I next want to thank the people who read portions of the manuscript and sent me comments or who responded to questions that came up as I wrote; at times I felt I was being helped by an army of experts. For this I am grateful to Yuichi Amitani, Eric Bapteste, Gillian Barker, David Baum, John Beatty, Ken Burnham, David Christensen, Eric Cyr Desjardins, Ford Doolittle, John Earman, Anthony Edwards, Branden Fitelson, Steven Frank, Richard Healey, Jonathan Hodge, Dan Hartl, Edward Holmes, John Huelsenbeck, James Justus, Bret Larget, Paul Lewis, William Mann, Sandra Mitchell, John Norton, Ronald Numbers, Samir Okasha, Roderick Page, Bret Payseur, Will Provine, Alirio Rosales, Bruce Russell, Larry Shapiro, Mike Steel, Christopher Stephens, Scott Thurow, and Carl Woese.

I am deeply indebted to the Vilas Trust at the University of Wisconsin; were it not for the research support provided by my William Vilas Professorship, I would not have been able to work so long and hard on this project. I also am grateful to the Rockefeller Foundation for the month's stay I had during May–June 2006 at their research center, the Villa Serbeloni in Bellagio, Italy. This is where I wrote a draft of Chapter 1 in delightful circumstances that still make me smile each time I think of them. Finally, I want to thank Sandra Mitchell and John Norton at the University of Pittsburgh's Center for Philosophy of Science for organizing a workshop on my book manuscript that took place in March 2007; I learned a lot during this event and the book is better because of it.

CHAPTER 1

Evidence

Scientists and philosophers of science often emphasize that science is a fallible enterprise. The evidence that scientists have for their theories does not render those theories certain. This point about *evidence* is often represented by citing a fact about *logic*: The evidence we have at hand does not deductively entail that our theories must be true. In a *deductively valid argument*, the conclusion must be true if the premises are. Consider the following old saw:

All human beings are mortal.
Socrates is a human being.
Socrates is mortal.

If the premises are true, you cannot go wrong in believing the conclusion. The standard point about science's fallibility is that the relationship of evidence to theory is *not* like this. The correctness of this point is most obvious when the theories in question are far more *general* than the evidence we can bring to bear on them. For example, theories in physics such as the general theory of relativity and quantum mechanics make claims about what is true at *all* places and *all* times in the entire universe. Our observations, however, are limited to a very small portion of that immense totality. What happens here and now (and in the vicinity thereof) does not deductively entail what happens in distant places and at times remote from our own.

If the evidence that science assembles does not provide certainty about which theories are true, what, then, does the evidence tell us? It seems entirely natural to say that science uses the evidence at hand to say which theories are *probably* true. This statement leaves room for science to be fallible and for the scientific picture of the world to change when new evidence rolls in. As sensible as this position sounds, it is deeply controversial. The controversy I have in mind is not between science and

nonscience; I do not mean that scientists view themselves as assessing how probable theories are while postmodernists and religious zealots debunk science and seek to undermine its authority. No, the controversy I have in mind is alive *within* science. For the past seventy years, there has been a dispute in the foundations of statistics between Bayesians and frequentists. They disagree about many issues, but perhaps their most basic disagreement concerns whether science is in a position to judge which theories are probably true. Bayesians think that the answer is *yes* while frequentists emphatically disagree. This controversy is not confined to a question that statisticians and philosophers of science address; scientists use the methods that statisticians make available, and so scientists in all fields must choose which model of scientific reasoning they will adopt.

The debate between Bayesians and frequentists has come to resemble the trench warfare of World War I. Both sides have dug in well; they have their standard arguments, which they lob like grenades across the no-man's-land that divides the two armies. The arguments have become familiar and so have the responses. Neither side views the situation as a stalemate, since each regards its own arguments as compelling. And yet the warfare continues. Fortunately, the debate has not brought science to a standstill, since scientists frequently find themselves in the convenient situation of not having to care which of the two approaches they should use. Often, when a Bayesian and a frequentist consider a biological theory in the light of a body of evidence, they both give the theory high marks. This allows biologists to walk away happy; they've got their answer to the biological question of interest and don't need to worry whether Bayesianism or frequentism is the better statistical philosophy. Biologists care about making discoveries about *organisms*; the *nature of reasoning* is not their subject, and they are usually content to leave such "philosophical" disputes for statisticians and philosophers to ponder. Scientists are *consumers* of statistical methods, and their attitude towards methodology often resembles the attitude that most of us have towards consumer products like cars and computers. We read *Consumer Reports* and other magazines to get expert advice on what to buy, but we rarely delve deeply into what makes cars and computers tick. Empirical scientists often use statisticians, and the "canned" statistical packages they provide, in the same way that consumers use *Consumer Reports*. This is why the trench warfare just described is not something in which most biologists feel themselves to be engulfed. They live, or try to live, in neutral Switzerland; the Battle of the Marne (they hope) involves others, far from home.

This book is about the concept of evidence as it applies in evolutionary biology; the present chapter concerns general issues about evidence that will be relevant in subsequent chapters. I do not aim here to provide anything like a complete treatment of the debate between Bayesianism and frequentism, nor is my aim to end the trench warfare that has persisted for so long. Rather, I hope to help the reader to understand what the shooting has been about. I intend to start at the beginning, to not use jargon, and to make the main points clear by way of simple examples. There are depths that I will not attempt to plumb. Even so, my treatment will not be neutral; in fact, it is apt to irritate both of the entrenched armies. I will argue that Bayesianism makes excellent sense for many scientific inferences. However, I do agree with frequentists that applying Bayesian methods in other contexts is highly problematic. But, unlike many frequentists, I do not want to throw out the Bayesian baby with the bathwater. I also will argue that some standard frequentist ideas are flawed but that others are more promising. With respect to frequentism as well, I feel the need to pick and choose. My approach will be "eclectic"; no single unified account of all scientific inference will be defended here, much as I would like there to be a grand unified theory.

One further comment before we begin: I have contrasted Bayesianism and frequentism and will return to this dichotomy in what follows. However, there are different varieties of Bayesianism, and the same is true of frequentism. In addition, there is a third alternative, likelihoodism (though frequentists often see Bayesianism and likelihoodism as two sides of the same deplorable coin). We will separate these inferential philosophies more carefully in what follows. But for now we begin with a stark contrast: Bayesians attempt to assess how probable different scientific theories are, or, more modestly, they try to say which theories are more probable and which are less. Frequentists hold that this is not what the game of science is about. But what do frequentists regard as an attainable goal? Hold that question in mind; we will return to it.

1.1 ROYALL'S THREE QUESTIONS

The statistician Richard Royall begins his excellent book on the concept of evidence (Royall 1997: 4) by distinguishing three questions:

(1) What does the present evidence say?
(2) What should you believe?
(3) What should you do?

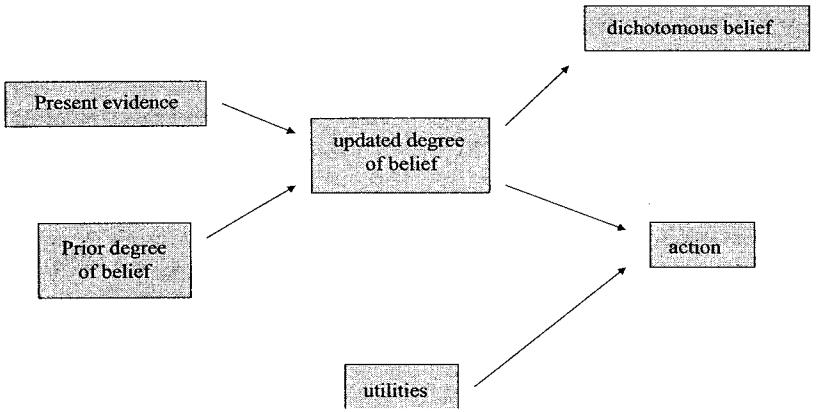

Figure 1.1 Present evidence and its downstream consequences.

If you are rational, you form your beliefs by consulting the evidence you have just gained, and when you decide what to do (which actions to perform), you should take account of what you believe. But answering question (2) requires more than an answer to (1), and answering question (3) requires more than an answer to (2). The extra elements needed are depicted in Figure 1.1.

Suppose you are a physician and you are talking to the patient in your office about the result of his tuberculosis test. The report from the lab says "positive." This is your present evidence. Should you conclude that the patient has tuberculosis? You want to take the lab report into account, but you have other information besides. For example, you previously had conducted a physical exam. Before you looked at the test report, you had some opinion about whether your patient has tuberculosis. The lab report may modify how certain you are about this. You update your degree of belief by integrating the new evidence with your prior information. This may lead you say to him "your probability of tuberculosis is 0.999."

If your patient is a philosopher who enjoys perverse conversation, he may reply, "but tell me, doctor, do I have tuberculosis, or not?" He doesn't want to know how *probable* it is that he has tuberculosis; he wants to know *whether* he has the disease – *yes or no*. This raises the question of whether a proposition's having a probability of 0.999 suffices for one to believe it, where belief is conceptualized as a dichotomous category: Either you believe the proposition or you do not. It may seem that a high degree of belief suffices for believing a proposition (even if it does not

suffice for being certain that the proposition is true), but there are complications. Consider Kyburg's (1970) lottery paradox. Suppose 1,000 lottery tickets are sold and the lottery is fair. *Fair* means that one ticket will win and each has the same chance of winning. If high probability suffices for belief, you are entitled to believe that ticket no. 1 will not win, since the probability of ticket 1's not winning is $\frac{999}{1000}$. The same is true of ticket no. 2; you should believe that it won't win. And so on, for each of the 1,000 tickets. But if you put these 1,000 beliefs (each of the form *ticket i will not win*) together with the rest of what you believe, your beliefs have become contradictory: You believe that some ticket will win (since you believe the lottery is fair), and you have just accepted the proposition that no ticket will win. Kyburg's solution to this puzzle is to say that acceptance does not obey a rule of conjunction; you can accept A and accept B without having to accept the conjunction $A\&B$.[1] This may be the best one can do for the concept of dichotomous belief, but it raises the question of whether we really need such a concept. After all, our everyday thought is littered with dichotomies that, upon reflection, seem to be crudely grafted to an underlying continuum. For example, we speak of people being *bald*, but we know that there is no threshold number of hairs that marks the boundary.[2] We are happy to abandon these crude categories when we need to, but we return to them when they are convenient and harmless.

If it makes sense to talk about rational acceptance and rational rejection, those concepts must bear the following relation to the concept of evidence:

If learning that E is true justifies you in *rejecting* (i.e., disbelieving) the proposition P, and you were not justified in rejecting P before you gained this information, then E must be evidence *against* P.

If learning that E is true justifies you in *accepting* (i.e., believing) the proposition P, and you were not justified in accepting P before you gained this information, then E must be evidence *for* P.

A theory of rational acceptance and rejection must provide more than this modest principle, which may seem like a mere crumb, hardly worth

[1] See Kaplan (1996) for a theory of rational acceptance that, unlike Kyburg's, obeys the conjunction principle.

[2] I say we "know" this, but Williamson (1994) and Sorenson (2001) have argued that in each use of a vague term, there is a cutoff, even if speakers are not aware of what it is. Their position is counterintuitive, but it cannot be dismissed without attending to their arguments (which we won't do here).

mentioning at all. But, in fact, it *is* worth stating, since later in this chapter it will do some important philosophical work.[3]

Even if this modest principle linking evidence and rational acceptance seems obvious, there is an old philosophical reason for pausing to ponder it. In the seventeenth century, Blaise Pascal sketched an argument that came to be called *Pascal's wager*. Earlier proofs of the existence of God had tried to demonstrate that there is evidence that God exists; Pascal endeavored to show that one ought to believe in God even if all the evidence one has is evidence *against*. The rough idea is this: If there is a God, you'll go to Heaven if you're a believer and go to Hell if you're not; on the other hand, if there is no God, it won't much affect your well-being whether or not you believe. Pascal wrote when probability theory was just starting to take its modern mathematical form, and his argument is a nice illustration of ideas that came to be assembled in *decision theory*. Though there is room to dispute the details of this argument (on which see Mougin and Sober 1994), the wager is of interest here because it appears to challenge the "modest" principle just enunciated. The wager purports to provide a reason for accepting the proposition that God exists even though it does not cite any evidence that there is a God. It is easy to think of nontheological arguments that pose the same challenge. Suppose I promise to give you $1,000,000 if you can get yourself to believe that the President is now juggling candy bars. If I am trustworthy, I have given you a reason to believe the proposition though I have not provided any evidence that it is true.

Commentators on Pascal's wager often distinguish two types of rational acceptance. The *act of accepting* a proposition can make good prudential sense, but that does not mean that *the proposition accepted* is well supported by evidence. When acceptance is driven by the costs and benefits that attach to the act of believing, I'll call this "prudential acceptance." When it is driven by the bearing of evidence on the proposition believed, I'll use the term "evidential acceptance." The modest principle linking evidence and "acceptance" really pertains to *evidential* acceptance. The principle, modified in this way, is true; in fact, it may even be true *by definition*. However, this does not settle whether it is ever permissible to

[3] It is interesting that the concept of evidence relates pairs of propositions to each other, while the concepts of acceptance and rejection relate propositions to persons. Smoke is evidence for fire, regardless of whether any agent takes this fact to heart. However, rational acceptance (or rejection) means that a person is justified in accepting (or rejecting) some proposition. The present disciplinary divide between philosophers of science and epistemologists coincides to a considerable degree with this distinction between questions concerning how propositions are related to each other and questions concerning how propositions are related to persons.

indulge in prudential acceptance. William James (1897) defends the right to believe when the evidence is silent in his essay "The Will to Believe." W. K. Clifford (1999) replies, in "The Ethics of Belief," that it is always wrong "to believe upon insufficient evidence." I will not try to adjudicate between these two positions. Suffice it to say that the modest principle stated earlier is binding on those who commit to having evidence control what they believe.

It may seem a long jump from Pascal's seventeenth-century theology to the hard edges of twentieth-century statistics, but Pascal's concept of prudential acceptance lives on in frequentism. The following remark by Neyman and Pearson (1933: 291) has often been quoted:

> No test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of [an] hypothesis [...] But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

Neyman and Pearson think of acceptance and rejection as *behaviors*, which should be regulated by prudential considerations, not by "evidence," which, for them, is a will o' the wisp. The prudential considerations they have in mind do not involve going to Heaven or Hell, but rather pertain to having true beliefs or false ones. There is no such thing as allowing "evidence" to regulate what we believe. Rather, we must embrace a policy and stick to it. If we do so, we can be certain (or, at least, it is overwhelmingly probable) that the percentage of false beliefs we accumulate over the long run will be held below some predesignated minimum. Not that present-day frequentists are all so dismissive of the concept of evidence (§1.4). But frequentists, early and late, have often embraced the idea of *prudential* belief.

Let us return to Figure 1.1. Suppose you, the physician, are 99.9 percent certain that your patient has tuberculosis, this degree of belief being based on the present tuberculosis test result and on other information you had from before. The thing to notice next is that your degree of belief does not, by itself, dictate what you should *say* or *do*. Should you tell your patient what you think? Should you remain silent? Should you lie? Should you hand him the pink pills you have in your desk? A rational decision about what to do requires more than the evidence you have and more than the degree of belief you have; a choice of action requires the input of values (which economists call *utilities*).

1.2 THE ABCS OF BAYESIANISM

Bayesianism is an answer to Royall's question (2): What should you believe? Bayesianism refines this question, substituting the concept of degree of belief for the dichotomous concept of believing or not believing a proposition. In our running example, Bayesianism addresses the question of how certain you should be that your patient has tuberculosis, given that his tuberculosis test came back positive.

Bayes' theorem

Bayesianism is based on Bayes' theorem, but the two are different. Bayes' theorem is a result in mathematics.[4] It is called a theorem because it is derivable from the axioms of probability theory (in fact, from a standard definition of conditional probability). As a piece of mathematics, the theorem is not controversial. Bayesianism, on the other hand, is a philosophical theory – it is an epistemology. It proposes that the mathematics of probability theory can be put to work in a certain way to explicate various concepts connected with issues about evidence, inference, and rationality.

Here is the rough idea of how Bayesianism uses Bayes' theorem: Before you make an observation, you assign a probability to the hypothesis H; this probability may be high, medium, or low (all probabilities by definition must be between 0 and 1, inclusive). After you make the observation, thereby learning that some observation statement O is true, you update the probability you assigned to H to take account of what you just learned. The probability that H has before the observation is called its *prior probability*; it is represented by $Pr(H)$. The word "prior" just means *before*; it doesn't mean that you know its value a priori (i.e., without any empirical input at all). The probability that H has in the light of the evidence O is called H's *posterior probability*; it is represented by the conditional probability $Pr(H \mid O)$; read this as "the probability of H, given O." Bayes' theorem shows how the prior and the posterior probability are related.

Now for the derivation of the theorem. Forget for just a moment that H means hypothesis and O means observation. Just regard them as any two

[4] A special case of the theorem was derived by Thomas Bayes and was published posthumously in the *Proceedings of the Royal Society* for 1764. Bayes' derivation was laborious and not fully general, very unlike the now-standard streamlined derivation I'll describe here.

propositions. Kolmogorov's (1950) definition of conditional probability is this:

$$Pr(H \mid O) = \frac{Pr(H \;\&\; O)}{Pr(O)}.$$

The definition is intuitive. For example, what is the probability that a card drawn at random from a standard deck is a heart, given that it is red? According to the Kolmogorov definition, this conditional probability has the same value as the ratio $Pr(\text{heart} \;\&\; \text{red})/Pr(\text{red})$. The denominator has a value of $\frac{1}{2}$. The proposition in the numerator, *heart & red*, is equivalent to *heart*, so the value for the numerator is $\frac{1}{4}$. Hence, the conditional probability has a value of $\frac{1}{2}$. By switching Hs and Os with each other in the Kolmogorov definition, you can see that it also is true that

$$Pr(O \mid H) = \frac{Pr(O \;\&\; H)}{Pr(H)}.$$

This means that the probability of the conjunction $H\&O$ can be expressed in two different ways:

$$Pr(H \;\&\; O) = Pr(H \mid O)\; Pr(O) = Pr(O \mid H)Pr(H).$$

From the second equality in the previous line, we obtain

$$\text{Bayes' theorem: } Pr(H \mid O) = \frac{Pr(O \mid H)Pr(H)}{Pr(O)}.$$

Here is some more terminology. I've already mentioned the *posterior probability* and the *prior probability* that appear in Bayes' theorem, but two other quantities are also mentioned. $Pr(O)$ is the *unconditional probability of the observations*. And R. A. Fisher dubbed $Pr(O \mid H)$ the *likelihood of H*. Because Fisher's terminology has become standard in statistics, I will use it here. However, this terminology is confusing, since in ordinary English, "likely" and "probably" are synonymous. So, beware! You need to remember that "likelihood" is a technical term. The likelihood of H, $Pr(O \mid H)$, and the posterior probability of H, $Pr(H \mid O)$, are different quantities and they can have different values. The likelihood

of H is the probability that H confers on O, not the probability that O confers on H. Suppose you hear a noise coming from the attic of your house. You consider the hypothesis that there are gremlins up there bowling. The likelihood of this hypothesis is very high, since if there are gremlins bowling in the attic, there probably will be noise. But surely you don't think that the noise makes it very probable that there are gremlins up there bowling. In this example, $Pr(O \mid H)$ is high and $Pr(H \mid O)$ is low. The gremlin hypothesis has a high likelihood (in the technical sense) but a low probability.

Let me add two more details that underscore the distinction between H's probability and its likelihood.

$$Pr(H) + Pr(notH) = 1$$

and

$$Pr(H \mid O) + Pr(notH \mid O) = 1$$

as well. The probability of a proposition and the probability of its negation sum to one; this is true for prior and also for posterior probabilities. But likelihoods need not sum to one; $Pr(O \mid H) + Pr(O \mid notH)$ can be less than 1, or more. Suppose you observe that Sue is a millionaire and wonder whether she won her wealth in last week's lottery. Your observation is very improbable under the hypothesis that she bought a ticket in the lottery and also under the hypothesis that she did not. To summarize this point: If you know the probability of H, you thereby know the probability of *notH*; but knowing the likelihood of H leaves the likelihood of *notH* completely open.

Another difference between likelihoods and probabilities concerns the difference between logically stronger and logically weaker hypotheses. Consider the following two hypotheses about the next card you'll be dealt from a standard deck:

H_1 = It's a heart.

H_2 = It's the Ace of Hearts.

The hypothesis H_2 is *logically stronger* than H_1; this means that H_2 entails H_1, but not conversely. Suppose the dealer is careless and you catch a glimpse of the card before it is dealt; you observe O = the card is red. Notice that H_1 has the higher posterior probability; $Pr(H_1 \mid O) = \frac{1}{2}$ while

$Pr(H_2 \mid O) = \frac{1}{26}$. But the two hypotheses have identical likelihoods, since $Pr(O \mid H_1) = Pr(O \mid H_2) = 1$. It is a theorem of probability theory that

If proposition X entails proposition Y, then $Pr(X) \leq Pr(Y)$, and $Pr(X \mid \text{data}) \leq Pr(Y \mid \text{data})$ no matter what the data are.

Logically stronger hypotheses can't have higher probabilities than logically weaker hypotheses, but they can have higher likelihoods. This point about likelihoods is illustrated by the relationship of H_1 and H_2 to the observation O' = the card is an ace.

A rule for updating

The different quantities used in Bayes' theorem are all available *before* you find out whether the statement O is true. You can know the value of $Pr(H \mid O)$ without knowing whether O is true, just as you can know that a conditional (an if/then statement) is true without knowing whether its antecedent (the if part) is true. All Bayes' theorem tells you is how the different probabilities it mentions, all assigned values at the same time, must be related. The theorem is, so to speak, a *synchronic* statement. But, as mentioned, Bayesianism provides advice about how you should change your degree of belief as you acquire new evidence. Bayes' theorem, therefore, must be supplemented by a rule for updating: This rule describes how probabilities should be related *diachronically*.

The rule of updating by strict conditionalization says that if O is the totality of the new information you have acquired, your *new* probability for H should be equal to your *old* value for $Pr(H \mid O)$. In other words: $Pr_{\text{now}}(H) = Pr_{\text{then}}(H \mid O)$, if O is all the evidence you acquired between then and now.

Before the result of the tuberculosis test is placed before you, you know the value of $Pr(S$ has tuberculosis $\mid$ the test is positive) and $Pr(S$ has tuberculosis $\mid$ the test is negative). These are your old posterior probabilities. When you learn that the test turned out positive, your new degree of belief for the proposition that S has tuberculosis is the one you assigned to the first of these conditional probabilities.

When I say that this rule for updating applies to "your" probability, does this mean that the Bayesian framework concerns only subjective degrees of belief? No – it is more general than this. You can think of this rule as giving normative advice to agents on how they should adjust the

amount of certainty they have. But a rule for updating also provides advice concerning what you should think the objective probability of a proposition is. If you think that the objective prior probability of drawing the Ace of Hearts from a normal deck is $\frac{1}{52}$, and you think that the objective posterior probability of the card's being the Ace of Hearts, given that it is red, is $\frac{1}{26}$, and you learn (just) that the next card drawn will be red, then your new objective probability for the card's being the Ace of Hearts should be $\frac{1}{26}$. It is useful to keep Bayesianism's *epistemological* advice about how probabilities should be assigned and manipulated separate from the *semantic* question of what probability statements mean. Not that interesting connections can't be drawn between the two issues. But first things first.

Strict conditionalization involves the idealization that an act of observation has the result that you find out that an observation statement is true or that it is false. What you learn isn't just that O is *probably true*; you learn that O is *true*. You then use this information to modify the degree of belief you have for some other proposition H. Bayesianism with strict conditionalization is a kind of hybrid philosophy, in which you accept or reject O but you do not apply the concept of dichotomous belief to H. Richard Jeffrey (1965) proposed a rule for updating in which you acquire only a degree of belief in O; the concept of dichotomous belief is thoroughly abandoned. Jeffrey's *probability kinematics* describes how your newly acquired degree of belief in O should affect your degree of belief in H.[5] For the purposes of this book, we can ignore Jeffrey's refinement and think of Bayesianism in terms of the idea of strict conditionalization. In what follows, I won't go to the trouble of distinguishing old probability assignments from new ones. Since I'll be focusing on the version of Bayesianism that uses the rule of strict conditionalization, I'll treat the posterior probability $Pr(H \mid O)$ as representing your updated degree belief once you learn that O is true (provided that O is *all* you learned).

Notice that the rule for updating by strict conditionalization addresses the case in which you *now* have a probability for proposition H, and you also had a (conditional) probability for that proposition *earlier*. It therefore fails to apply to cases of conceptual innovation in which H involves concepts that you just formulated. You didn't have a conditional

[5] Although Jeffrey's conditionalization is more realistic than strict conditionalization in terms of its characterization of the input, it has a logical oddity that strict conditionalization avoids. The *order* in which new evidence arrives can affect the final degree of belief in Jeffrey's conditionalization, but not in strict.

probability for H earlier because H uses concepts you didn't have available back then. This is an especially important feature of some scientific innovations; scientists often work within the confines of a fixed stock of concepts, but every so often they break out. Evolutionists sometimes draw a distinction between micro- and macroevolution (§2.19); the former describes changes that occur within an enduring species whereas the latter describes changes that result in the appearance of new species. Kuhn's (1962) distinction between normal science and revolutionary science is similar; there is science pursued within an existing "paradigm" and science that results in the formation of new paradigms. Bayesian updating by strict conditionalization makes more sense in connection with the micro-changes that occur within normal science; it is controversial whether it can represent the macro-changes that occur in scientific revolutions.[6]

Posterior probabilities, likelihoods, and priors

Let's apply Bayes' theorem to the running example that you are a doctor and your patient has a positive tuberculosis test result. You want to use this new information to figure out how certain you should be that he has tuberculosis. Bayes' theorem says that

$$(4)\quad Pr(\text{tuberculosis} \mid +\text{ result}) = \frac{Pr(+\text{ result} \mid \text{tuberculosis})Pr(\text{tuberculosis})}{Pr(+\text{ result})}.$$

Bayes' theorem also can be stated for the hypothesis that S does *not* have tuberculosis:

$$(5)\quad Pr(\text{no tuberculosis} \mid +\text{ result}) = \frac{Pr(+\text{ result} \mid \text{no tuberculosis})Pr(\text{no tuberculosis})}{Pr(+\text{ result})}.$$

Combining (4) and (5) yields the following equality of ratios:

$$(6)\quad \frac{Pr(\text{tuberculosis} \mid +\text{ result})}{Pr(\text{no tuberculosis} \mid +\text{ result})} = \frac{Pr(+\text{ result} \mid \text{tuberculosis})}{Pr(+\text{ result} \mid \text{no tuberculosis})} \times \frac{Pr(\text{tuberculosis})}{Pr(\text{no tuberculosis})}.$$

[6] See Eells (1985) and Earman (1992) for discussion of the closely related problem of old evidence. The problem described above is located in what Earman calls "the problem of new theories."

Notice that the quantity $Pr(+\text{ result})$, the unconditional probability of the observations, which is present in both (4) and (5), now has disappeared. Proposition (6) says that the ratio of posterior probabilities equals the ratio of likelihoods times the ratio of priors.

Before you observe the test result, you have your two prior probabilities; these must sum to one, but their ratio may of course be greater than unity, or less. Will your observation of the positive test result lead you to change your degrees of belief? They cannot if the two likelihoods are the same. If

$$Pr(+\text{ result} \mid \text{tuberculosis}) = Pr(+\text{ result} \mid \text{no tuberculosis}),$$

the ratio of the posterior probabilities will be the same as the ratio of priors. In this case, the observation is uninformative. In fact, you needn't even bother to check how the test came out. On the other hand, if

$$Pr(+\text{ result} \mid \text{tuberculosis}) > Pr(+\text{ result} \mid \text{no tuberculosis}),$$

your observation makes a difference. A positive test result will increase your confidence that S has tuberculosis (and reduce your confidence that he does not). In this case, the observation has the effect of making the ratio of posterior probabilities larger than the ratio of priors. The likelihood ratio, the first product term on the right-hand side of (6), is *the* pathway by which the test result can lead you to revise your degree of belief in whether S has tuberculosis. For Bayesianism, there is no other.

Another way to see this point is to delve more deeply into the instance of Bayes' theorem given in (4). What does "the unconditional probability of the observation" mean? A positive test result can occur when S has tuberculosis, but it also can occur when S does not (in which case the test result is mistaken). Both these possibilities are represented in the unconditional probability of the observations:

$$\begin{aligned}(7)\quad Pr(+\text{ result}) = {} & Pr(+\text{ result} \mid \text{tuberculosis})Pr(\text{tuberculosis}) \\ & + Pr(+\text{ result} \mid \text{no tuberculosis})Pr(\text{no tuberculosis}).\end{aligned}$$

The unconditional probability of the observation is the *average* probability that the observation has under the two alternative hypotheses, where the average is taken by using weighting terms supplied by the prior

probabilities; in other words, $Pr(+\text{ result})$ is a weighted average of the two likelihoods. If we use (7) to rewrite (4), we obtain:

$$(8)\ Pr(\text{tuberculosis} \mid +\text{ result}) = \frac{Pr(+\text{ result} \mid \text{tuberculosis})Pr(\text{tuberculosis})}{Pr(+\text{ result} \mid \text{tuberculosis})Pr(\text{tuberculosis}) + Pr(+\text{ result} \mid \text{no tuberculosis})Pr(\text{no tuberculosis})}.$$

If $Pr(+\text{ result} \mid \text{tuberculosis}) = Pr(+\text{ result} \mid \text{no tuberculosis})$, the denominator in (8) is equal to $Pr(+\text{ result} \mid \text{tuberculosis})$, in which case (8) simplifies to

$$Pr(\text{tuberculosis} \mid +\text{ result}) = Pr(\text{tuberculosis}).$$

Without a difference in likelihoods, the posterior probability must have the same value as the prior; the observation has not affected your degree of belief.

Confirmation

As mentioned earlier, Bayesianism is more than Bayes' theorem. The philosophy goes beyond the mathematics because the philosophy proposes definitions of key epistemological concepts. For example, Bayesianism defines confirmation as probability-raising and disconfirmation as probability-lowering:

(Qual) O confirms H if and only if $Pr(H \mid O) > Pr(H)$.
O disconfirms H if and only if $Pr(H \mid O) < Pr(H)$.
O is confirmationally irrelevant to H if and only if $Pr(H \mid O) = Pr(H)$.

Confirmation does not mean *proving true* and disconfirmation does not mean *proving false*; confirmation and disconfirmation mean only that an observation should increase or reduce your confidence that H is true. Thus, the observation that O is true can confirm H even though $Pr(H \mid O)$ is still low; the posterior probability just has to be higher than the prior. And O can disconfirm H even though $Pr(H \mid O)$ is still high; O just has to lower H's probability. Bayesian confirmation and disconfirmation involve *comparisons* of probabilities; they say nothing about the *absolute values* of any probability. Bayes' theorem allows an equivalent definition of Bayesian confirmation to be extracted from the one given above:

O confirms H if and only if $Pr(O \mid H) > Pr(O \mid notH)$.

To see whether O confirms H, don't ask whether H, if true, would lead you to expect that O is true. Rather, ask whether H makes O more probable than *not*H does.

The definitions stated in (Qual) characterize a *qualitative* concept of confirmation. They do not provide a measure of *degree* of confirmation; (Qual) doesn't say *how much* O confirms H. How might a *quantitative* concept be defined? Here are some candidates to consider, where $DoC(H,O)$ represents the degree to which O confirms H:

(Diff) $$DoC(H, O) = Pr(H \mid O) - Pr(H).$$

(Ratio) $$DoC(H, O) = \frac{Pr(H \mid O)}{Pr(H)}.$$

(L-Ratio) $$DoC(H, O) = \frac{Pr(O \mid H)}{Pr(O \mid notH)}.$$

All three of these definitions agree that (Qual) is true. However, they are not *ordinally equivalent*; they can disagree as to whether O_1 confirms H_1 more than O_2 confirms H_2. For example, suppose that

$$Pr(H_1 \mid O_1) = 0.9 \quad Pr(H_1) = 0.5$$
$$Pr(H_2 \mid O_2) = 0.09 \quad Pr(H_2) = 0.02.$$

According to (Diff), the difference measure, O_1 confirms H_1 more than O_2 confirms H_2, since $0.4 > 0.07$. But, according to the ratio measure, the reverse is true, since $\frac{9}{5} < \frac{9}{2}$. The fact that these and other measures sometimes disagree has given rise to a lively debate among Bayesians as to which measure is best (Fitelson 1999). Bayesians who despair of resolving this question try to restrict their discussion of confirmation to the qualitative definition (Qual).

Do we need to measure degree of confirmation? Perhaps the qualitative notion is enough. After all, there seems to be little reason to compare how much the fossil record confirms the Darwinian theory of evolution with how much Eddington's observation of light bending during an eclipse confirms the GTR. True, but there are other scientific contexts in which quantitative questions about confirmation matter. For example, in Chapter 4 we'll consider the hypothesis that two or more species share a common ancestor, and we'll investigate whether the *adaptive* similarities that the species share or the *neutral* similarities that they share provide stronger evidence in favor of that hypothesis. Even if

$Pr(X$ and Y have a common ancestor $\mid X$ and Y share adaptive trait $T_1) > Pr(X$ and Y have a common ancestor) and $Pr(X$ and Y have a common ancestor $\mid X$ and Y share neutral trait $T_2) > Pr(X$ and Y have a common ancestor).

there is another question that remains to be addressed. If it makes sense to ask which kind of similarity provides stronger evidence for common ancestry, (Qual) is not enough.

Reliability

What does it mean to say that a tuberculosis test is "reliable"? Does it mean that what the test says has a high probability of being true? That is, does it mean that

(9) $Pr(\text{tuberculosis} \mid +\ \text{result})$ and $Pr(\text{no tuberculosis} \mid -\ \text{result})$ are both large?

Or does it mean that when the person taking the test has tuberculosis (or not), the procedure can be relied upon to say what is true? That is, does it mean that

(10) $Pr(+\ \text{result} \mid \text{tuberculosis})$ and $Pr(-\ \text{result} \mid \text{no tuberculosis})$ are both large?

As emphasized earlier, it is important not to confuse $Pr(O \mid H)$ and $Pr(H \mid O)$. Recall the example about the gremlins. But what does the word "reliability" mean?

Here's how I think the term is used in ordinary English: When a witness is reliable, what he or she says is probably true. Witnesses who are apt to pick up on what is true might be said to be *sensitive*; if the proposition is true, they will probably notice that it is and tell you. In my view, ordinary usage pairs "reliable" with (9) and "sensitive" with (10). But whether or not this is how the terms are used in everyday discourse, *aficionados* of probability have come to use the term "reliability" to indicate that (10) is true, not that (9) is.[7] A reliable tuberculosis test procedure has a large likelihood ratio for each possible test outcome:

(R) $$\frac{Pr(+\ \text{result} \mid \text{tuberculosis})}{Pr(+\ \text{result} \mid \text{no tuberculosis})} \gg 1.0 \quad \frac{Pr(-\ \text{result} \mid \text{no tuberculosis})}{Pr(-\ \text{result} \mid \text{tuberculosis})} \gg 1.0.$$

[7] Actually, the terminology is more varied. For example, a "reliable" method for ranking options given a set of data is sometimes defined as one that usually returns the same ranking across different data sets; a method that ignores the data and always imposes the same ranking would be perfectly "reliable" in this sense.

Given this meaning, your patient S can obtain a positive test result on the reliable tuberculosis test you gave him and still it is highly improbable that he has tuberculosis. This will be true if the prior probability of S's having tuberculosis is sufficiently low (imagine that S is drawn at random from a population in which tuberculosis is very rare and then is given the test). To verify that this can happen, have another look at the relationship of the three ratios described in proposition (6).

Why is the term "reliability" often used by probabilists with the meaning described in (R)? Is this sheer perversity on their part? In fact, there is reason to focus on (R) even though people take tuberculosis tests to find out if they (probably) have the disease. Imagine using the same test procedure in two populations. In the first, people frequently have tuberculosis; in the second, they rarely do. There is a useful sense of "reliability" in which the test procedure is equally reliable in the two populations. Yet, if people are sampled at random in the two populations and then take the test, Pr(tuberculosis) is higher in the first population than in the second. If the test is equally reliable in the two cases, Pr(tuberculosis | + test outcome) will be higher in the first case than in the second. Tuberculosis tests are in this respect like a great many detectors and measurement procedures. Whether the test returns a positive or a negative verdict is determined just by facts specific to the person or thing taking the test; thermometers are related to ambient temperature in the same way, and pregnancy tests are related to pregnancy in that way as well. Whether the person has a common or a rare condition is irrelevant to what the test will say. To put the point abstractly, *likelihoods are often independent of priors.* But posterior probabilities depend on both likelihoods *and* priors. This feature that a test procedure has, which is stable across different applications in different populations, is worth noting; this is why the ratios described in (R) are important.

In saying that the posterior probability of tuberculosis "depends" on priors and likelihoods, but that the likelihoods are "independent" of priors and posteriors, I am describing the *physical* characteristics of test procedures, not the *mathematical* relationships characterized by Bayes' theorem. In Bayes' theorem, each of the quantities mentioned is a mathematical function of the other three; given any three values, you can calculate the fourth. However, this symmetry with respect to mathematical dependence is not present when we consider physical relationships. Whether a tuberculosis test is apt to yield a positive result depends on whether the person taking the test has tuberculosis, not on whether tuberculosis is common or rare.[8]

Expectation and expected value

It is often said that a baby born in the USA today can expect to live about seventy-eight years. What does this mean? The reality is that a baby not only might have a longer life than this, or a shorter one. Each possible lifespan has its own probability; p_1 is the probability of living exactly one year, p_2 is the probability of living exactly two, and so on. The figure of seventy-eight years is the mathematical expectation, a technical term:[9]

$$\begin{aligned}&\mathrm{E}(S\text{'s longetivity} \mid S \text{ is born in the USA in 2008})\\ &= 1(p_1) + 2(p_2) + \cdots + n(p_n) = \sum i(p_i) = 78 \text{ years.}\end{aligned}$$

$\mathrm{E}(x \mid y)$ represents the expected value of x given y; notice that x is a quantity and y is a proposition. Probabilities must fall between 0 and 1, but expected values need not. The expected value is an average; in fact, it is a *weighted* average, because the different possible longevities have different probabilities.

If seventy-eight years is the life expectancy, does that mean that you should expect a US newborn to live about seventy-eight years? That depends on how different possible longevities are distributed around this mean value. Figure 1.2 shows three hypothetical distributions. Each is symmetrical and is centered on seventy-eight years, so 78 is the average value according to each. It wouldn't make much sense to expect a baby to live about seventy-eight years if (a) were true. According to (a), a baby will probably live only a very short life or a very long one; it will be exceedingly rare for a baby to live about seventy-eight years. In (b), all lifespans from 0 to 156 years are equally probable, so here again it would not make sense to use the expected value as the value you should expect. In (c), not only is 78 the expected value, but it is highly probable that a US newborn will live about seventy-eight years. There is less variation around the mean value in (c) than there is in (a) and (b). In (c), it is sensible to use the expected value as the approximate value you'd expect.

[8] In §4.5, we'll examine a kind of evolutionary process, one that involves frequency dependent selection, in which priors and likelihoods do not exhibit this type of independence.

[9] To keep the example simple, I assume that lifespans come in whole numbers of years. This permits the expected value to be expressed as a summation over discrete quantities. If we take time to be a continuous quantity, the expectation will be an integral.

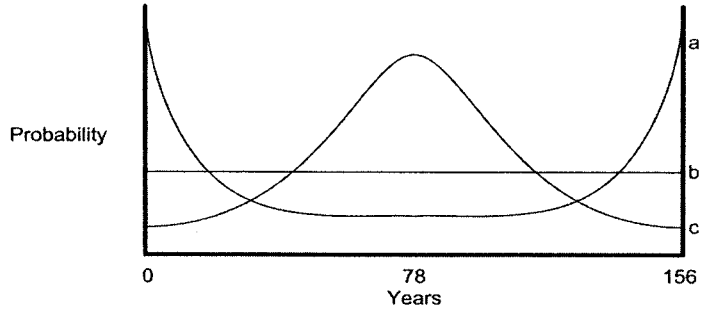

Figure 1.2 Three possible distributions of longevities. Each has the same expected value, seventy-eight years.

Induction

One of the important contributions that Bayesianism has made to understanding scientific reasoning is that it has thrown light on the traditional idea of learning by induction. Induction, as I use the term, means making an inference about a population based on a sample drawn from it. The inference may concern what *the next object* sampled will probably be like, or what *all the objects* in the population are probably like. There is a lot more to scientific reasoning than inductive sampling, but it is enlightening to see what induction looks like through Bayesian lenses.

Here is a seemingly plausible principle of inductive reasoning that Reichenbach (1938) called *the straight rule*:

If you toss a coin n times and h of those tosses come up heads, infer that Pr(the coin lands heads | the coin is tossed) $= h/n$.

This rule is not the only one to consider. For example, Laplace (1820) described a *rule of succession*:

If you toss a coin n times and h of those tosses come up heads, infer that Pr(the coin lands heads | the coin is tossed) $= (h+1)/(n+2)$.

The two rules disagree (though they disagree less the more you toss). Which is the right one to use? Reichenbach's rule looks simple and it seems to "go by the evidence," while Laplace's seems to introduce a funny correction to what the evidence is saying. Is this a good reason to prefer Reichenbach to Laplace? Bayesianism provides a framework for answering

this question. But, more importantly, Bayesianism exposes a deficiency present in both rules; there is a kind of assumption that neither rule makes explicit but that needs to be in place if any such rule is to make sense. Notice that both rules draw a conclusion about the value of a posterior probability, based on the evidence at hand, but neither rule states values for any prior probability. Bayesianism asserts that this is *magical thinking*. The observations alone cannot give you a posterior probability; you need to have a prior probability as well. A central thesis of Bayesianism is: *no probabilities out without some probabilities in.*

Laplace was well aware of this point. He identified an assignment of prior probabilities that allowed him to *prove* that the rule of succession is correct. Let p be the probability of heads on each toss. We assume that tosses are independent of each other; results on earlier tosses don't affect the probability of heads on later ones. Laplace's assumptions about prior probabilities include the postulate that p has the same chance of falling between 0.1 and 0.2 as it has of falling between 0.8 and 0.9 and that its chance of falling between 0.3 and 0.6 is the same as its chance of falling between 0.4 and 0.7. Perhaps it sounds strange to assign a probability to a probability; if so, think of p as a physical property of the coin, perhaps one that concerns how symmetrical it is. In any event, to fully describe how Laplace conceived of the prior probabilities associated with p, we need to think about the fact that there are infinitely many values that p might have. This means that Laplace can't express his postulate about prior probabilities by saying that all point values of p have the same probability. If they all have a probability of zero, they sum to zero; and if they all have a positive value, they sum to infinity. What is required is that they sum to unity. The solution is to shift from talk of *probability* to talk of *probability density*, an idea depicted in Figure 1.3. Densities take values from zero to infinity. The prior density represented in the figure always has a value of 1, so the area under this density curve has a value of unity. Probabilities are areas under density curves. Laplace's assumption was that the prior density curve is flat. Each point value for p has a probability of zero and a probability density of 1.[10]

According to this prior density curve, the expected value of p is $\frac{1}{2}$. Notice that the curve is symmetrical around $p = \frac{1}{2}$. Imagine a factory that manufactures coins according to this prior density function. A tenth of

[10] Laplace thought that this assumption is justified by the principle of indifference, which we'll examine in the next section. Here we'll simply examine the assumption's consequences.

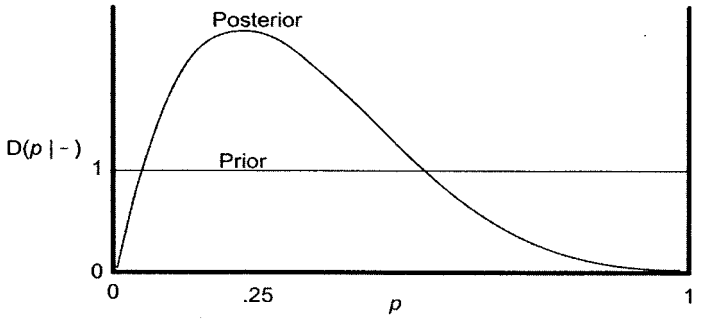

Figure 1.3 A flat prior density distribution for p and the non-flat posterior density occasioned by observing one head in four tosses. The prior expected value of p is 0.5; given this prior, the posterior expected value of p is 0.33.

the coins it produces have $0 < p < 0.1$, a tenth have $0.1 < p < 0.2$, and so on. So the average coin produced from this factory has a value of $p = \frac{1}{2}$. If you draw a coin at random from this prior distribution, and if you allow yourself to think of the expected value of p as the value you should expect p to have (thus setting aside the previous section's warning about how expected values should be interpreted), you can say that Laplace's assumption about priors entails that you should expect the coin to be fair before you have tossed it even once. This vindicates what the rule of succession says when $h = n = 0$; in this case, $\frac{(h+1)}{(n+2)} = \frac{1}{2}$. The next step is to understand what happens when you start tossing the coin. Does Laplace's rule give correct values for the expected value of p, conditional on the observations you have made? Surprisingly, the answer is *yes*.

We already know from the gremlins example that the hypothesis with the highest likelihood need not be the one with the highest posterior probability. The reason is that the prior probability is an "anchor"; the observations can lead the posterior probabilities to depart from the priors, but the priors still influence what values those posterior probabilities will have. If you obtain one head in four tosses, you have some evidence that the expected value of p is lower than $\frac{1}{2}$. But this does not permit you to ignore the prior expected value. This is why the posterior expectation moves away from the prior value of $\frac{1}{2}$ in the direction of $\frac{h}{n} = \frac{1}{4}$ and ends up somewhere in between, with a posterior expectation of $\frac{1}{3}$. How much of a shift the rule of succession tells you to make depends not just on the

frequency of heads in the observations, but on the absolute number of tosses. Observing one head in four tosses occasions a smaller shift away from $\frac{1}{2}$ than observing 100 heads in 400 tosses. The posterior expectation in the former case, as just noted, is $\frac{1}{3}$, while that in the latter case is $\frac{101}{402}$.

Laplace's rule is correct if you start with a flat prior density and you think that the proper target of this inductive rule is to infer the expected value of p. Where does that leave Reichenbach? Perhaps there is another assignment of prior probabilities that justifies the straight rule. Let's investigate this question by initially changing the subject. Instead of thinking about the *probabilities* of hypotheses, let's think about their *likelihoods*. Suppose we observe five heads in twenty tosses of the coin. What value of p = *Pr*(the coin lands heads | the coin is tossed) will maximize the probability of the observations, again assuming that tosses are independent of each other? The maximum likelihood estimate of this parameter is $p = \frac{5}{20} = 0.25$. The likelihood of this hypothesis is depicted in Figure 1.4, relative to the observations we actually made (five heads in twenty tosses) and also with respect to other observations that could have occurred but did not. The figure also represents the likelihood of the hypothesis that $p = \frac{3}{4}$ relative to different possible data sets. Note that the hypothesis $p = \frac{1}{4}$ says that the actual observations were more probable than the hypothesis $p = \frac{3}{4}$ says they were. In fact, the $p = \frac{1}{4}$ hypothesis makes the data more probable than *any* assignment of a point value to p does; it provides the estimate of *maximum* likelihood. The maximum likelihood estimate of p is just the sample frequency; it doesn't matter

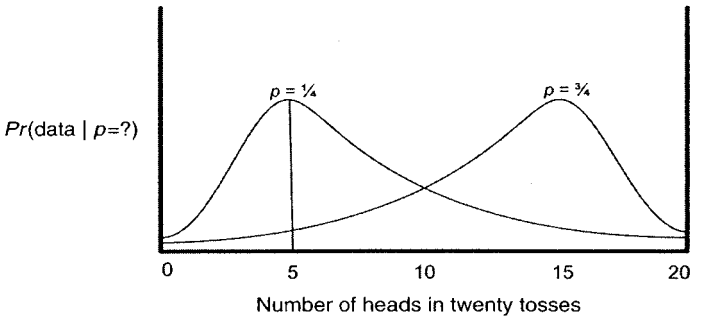

Figure 1.4 When the coin lands heads in five of twenty tosses, the maximum likelihood estimate of p = *Pr*(the coin lands heads | the coin is tossed) is $p = \frac{1}{4}$. The likelihood of the estimate $p = \frac{3}{4}$ is lower.

whether you observe one head in four tosses, or five in twenty, or 100 in 400 – the maximum likelihood estimate is the same.

The fact that the hypothesis $p = \frac{1}{4}$ has a higher *likelihood* than the hypothesis $p = \frac{3}{4}$ does not say anything about their *probabilities*. If those hypotheses are to have posterior probabilities, they must have priors. So what priors should we assign? More specifically, is there a prior density distribution of values for p that allows Reichenbach's rule to always generate the right value for the posterior expected value of p? Surprisingly, the answer is *no*. Notice that the straight rule pays no attention to the prior values; it simply goes by the maximum likelihood estimate. There is no prior distribution that legitimizes this policy.[11] The rule of succession is typical in this regard; it moves the estimate from the prior expected value of $\frac{1}{2}$ towards the maximum likelihood estimate of h/n, but does not go all the way there. The only case in which the rule of succession yields a value that is identical with the maximum likelihood estimate is when $h/n = 0.5$; in this case $(h + 1)/(n + 2)$ also equals 0.5. The general point is that *every* prior distribution will have a prior expected value, and this will always exert some influence on what the posterior expected value is. The straight rule cannot be given a Bayesian foundation.[12]

Trouble in Paradise

If all scientific inferences resembled the problem you face when your patient's tuberculosis test has a positive result, Bayesianism would be a thoroughly adequate philosophy of scientific inference. Before describing the fly in the ointment (in fact, there are two), let us examine some features of this example.

In the example of tuberculosis diagnosis, the two hypotheses are exclusive and exhaustive.[13] This is why $Pr(S$ has tuberculosis$) + Pr(S$ does not have tuberculosis$) = 1.0$. What is more, when you assign values to these prior probabilities, you are not merely reporting your subjective degree of certainty. You can point to frequency data concerning how

[11] Or, more precisely, no prior distribution that obeys the axioms of probability permits this. A flat *improper* prior (which goes outside the unit interval) can do so.

[12] Not that Reichenbach thought that the straight rule requires a Bayesian justification. Rather, he was impressed with the fact that the straight rule converges on the true value of p as the data set is made large without limit. This property, which statisticians call *statistical consistency*, will be discussed in §1.7 and §4.8.

[13] I assume here that your patient, S, exists and that this is not up for test.

often people have tuberculosis in the population to which S belongs. Of course, S belongs to many populations; for example, suppose that S lives in the USA, lives in Wisconsin, and lives in Madison, and that the frequencies of tuberculosis in these three populations differ. Philosophers often recommend considering the narrowest population on which you have frequency data, but I don't think that that is the only consideration. It matters whether you can regard S as being drawn at random from this or that population; if you can, the frequency data for that population provide a defensible prior. Although there are interesting issues here as to what the best assignment of value to the prior probability is, the point I want to emphasize is that frequency data are relevant and available.

The same virtue attaches to the values assigned to the likelihoods $Pr(+$ result $|$ tuberculosis$)$ and $Pr(+$ result $|$ no tuberculosis$)$. These are not numbers pulled from thin air, nor are they mere introspective reports about your attitudes. Rather, they too can be justified by pointing to frequency data. It is a familiar fact that scientific instruments, including the devices employed in medical diagnosis, are used to test hypotheses. The point of relevance here is that those devices are themselves tested. You can see how well a tuberculosis test performs by giving the test to a large number of people whom you know have tuberculosis and also to a large number whom you know do not. Frequencies within large samples provide a substantial justification for one assignment of values to the likelihoods rather than another.

In saying this, I am not denying the main lesson of the previous section. Frequency data do not by themselves *deductively entail* an assignment of value to a posterior probability. The fact that $p = h/n$ is the maximum likelihood estimate for a coin's probability of landing heads does not entail that this is the most probable value; still less does it entail that this is the true value. It is useful to think of the probability one is trying to estimate as a theoretical quantity; the evidence one uses to make this estimate is an observed frequency. The observations do not deductively entail the theory. However, with large samples, almost any prior probability will produce the same, or nearly the same, assignment of posterior probabilities. This is what Bayesians mean when they refer to the *swamping of priors*. Two agents can begin with different prior probabilities, but if they both update by using a sufficiently large data set, their posterior probabilities will be very close; the difference in priors has *washed out*. In this case, you will not go far wrong by ignoring whatever prior probabilities you start with and just using Reichenbach's straight

rule. The rule is invalid, as noted, but the values it delivers will usually be sensible for large random samples.

It is important to recognize how important it is for prior probabilities to be grounded in evidence. We often calculate probabilities to resolve our own uncertainty or to persuade others with whom we disagree. It is no good assigning prior probabilities simply by asking that they reflect how certain we feel that this or that proposition is true. Rather, we need to be able to cite reasons for our degrees of belief. Frequency data are not the only source of such reasons, but they are one very important source. The other source is an empirically well-grounded theory. When a geneticist says that Pr(the offspring has genotype Aa | mom and dad both have the genotype Aa) $= \frac{1}{2}$, this is not just an autobiographical comment. Rather, it is a consequence of Mendelism, and the probability assignment has whatever authority the Mendelian theory has. That authority comes from empirical data.

I don't want to overstate my praise for the objectivity of the quantities that figure in the Bayesian answer to the question of whether your patient has tuberculosis. Skeptical questions can always be pursued back to a point where you do not know how to answer, or you "answer" by stamping your foot and insisting on the legitimacy of assumptions that cannot be further justified. This is true for any claim about knowledge or justification; the present context is no exception. But to insist that the Bayesian solution to the diagnostic problem is "purely subjective" is to mistake the part for the whole. The objective component is substantial and compelling.

There is a world of difference between this quotidian case of medical diagnosis and the use of Bayes' theorem in testing a deep and general scientific theory, such as Darwin's theory of evolution or Einstein's general theory of relativity. The difference may be, at the end of the day, a matter of degree, but still the difference is profound. When we assign prior probabilities to these theories, what evidence can we appeal to in justification? We have no frequency data as we do with respect to the question of whether S has tuberculosis. If God chose which theories to make true by drawing balls from an urn (each ball having a different theory written on it), the composition of the urn would provide an objective basis for assigning prior probabilities, if only we knew how the urn was composed. But we do not, and, in any event, no one thinks that these theories are made true or false by a process of this kind. As I mentioned, frequency data are not the only convincing justification that an assignment of prior probabilities can have. An empirical theory, like Mendelism, that

is itself justified by observations can provide such probabilities. But this possibility does not bear fruit in the case of Darwin's theory or Einstein's; we have no empirically well-grounded theory of the processes by which theories like Darwin's or Einstein's are made true. In fact, maybe there is no such theory; perhaps Darwin's and Einstein's theories simply are true (or not), with no chance process leading to the one outcome or the other.

Although frequency data and a well-supported empirical theory can provide a basis for assigning prior probabilities, the principle of indifference cannot. This idea used to be a cornerstone of Bayesianism, but it is rare for contemporary Bayesians to have anything good to say about it. The principle says that if you are completely ignorant about which of a set of exclusive and exhaustive propositions is true, that you should assign them equal probabilities that sum to one. The problem with this principle is that there are multiple ways to slice the logical space into parts, which means that the same proposition can receive different prior probabilities depending on how the cake is sliced. It once was hoped that logic and language would somehow ground the principle of indifference, but this no longer seems even remotely plausible; logic and language do not furnish prior probabilities, at least not if prior probabilities are to have some authority in arguments in which people disagree. So do not fall into the trap of reasoning thus:

> Either God exists or he does not.
>
> Therefore, Pr(God exists) $=$ Pr(God does not exist) $= \frac{1}{2}$.

This is a trap because the pie can also be divided in three:

> Either God exists and Christianity is true, God exists and Christianity is false, or there is no God.
>
> Therefore, Pr(God exists and Christianity is true) $=$ Pr(God exists and Christianity is false) $=$ Pr(God does not exist) $= \frac{1}{3}$.

If the principle of indifference licenses the first inference, why does it not license the second? And if it licenses both, it has lapsed into contradiction.

Laplace appealed to the principle of indifference to justify the prior density distribution he used to derive the rule of succession, so the dilemma of embracing either arbitrariness or contradiction arises in this

context as well. Bertrand's paradox provides a nice illustration of how the principle of indifference goes wrong in the continuous case. Suppose I tell you that a cube manufactured by a certain factory has edges that are between 1 and 2 inches in length. If this is all you know about the cube, you might conclude that all possible lengths between 1 and 2 have the same prior density (= 1). This implies that

$$Pr(\text{the length of an edge is between 1 and 1.5 inches})$$
$$= Pr(\text{the length of an edge is between 1.5 and 2 inches}) = \frac{1}{2}.$$

However, the information I gave you also allows you to see that each side of the cube has an area that is somewhere between 1 and 4 square inches, and this might suggest that all possible areas between 1 and 4 have the same prior densities (= 1). This entails that

$$Pr(\text{the area of a side is between 1 and 2.5 square inches})$$
$$= Pr(\text{the area of a side is between 2.5 and 4 square inches}) = \frac{1}{2}.$$

The problem is that assigning equal priors to the lengths an edge might have contradicts assigning equal priors to the areas a side might have.

The questions just explored concerning the assignment of values to prior probabilities also attach to likelihoods, or rather they attach to some of them. In the case of S and whether he has tuberculosis, assignments of values to $Pr(+ \text{ test result} \mid S \text{ has tuberculosis})$ and to $Pr(+ \text{ test result} \mid S$ does not have tuberculosis) can be justified. The problem is that only half of this is true in many other testing situations. For example, when Arthur Stanley Eddington tested the general theory of relativity (GTR) by examining how much bend there was in starlight during a solar eclipse, he was able to ascertain a value for $Pr(\text{observation} \mid GTR)$. But what value could he assign to $Pr(\text{observation} \mid notGTR)$? The negation of the GTR is what philosophers call a *catchall hypothesis*. There are many specific theories (T_1, T_2, ..., T_n) that are incompatible with the GTR. The likelihood of *notGTR* is the average likelihood of these specific alternatives, weighted by the probability they have conditional on the GTR being false:

$$Pr(\text{observation} \mid notGTR) = \sum_i Pr(\text{observation} \mid T_i) Pr(T_i \mid notGTR).$$

Some alternatives to the GTR have not even been formulated yet, so it is hard to see how anyone can say what their likelihoods are. And what objective meaning could there be in saying that various alternatives have this or that probability of being true if the GTR is false? If the likelihood of the catchall hypothesis *notGTR* cannot be calculated, there is no saying whether Eddington's observation confirms the GTR, since

$$Pr(GTR \mid \text{observation}) > Pr(GTR) \text{ if and only if}$$
$$Pr(\text{observation} \mid GTR) > Pr(\text{observation} \mid notGTR).$$

As it happens, Eddington did not test the GTR against its negation; rather, he tested it against Newtonian theory, which made a concrete prediction about how much the light in the eclipse should bend. It turned out that

$$Pr(\text{observation} \mid \text{GTR}) \gg Pr(\text{observation} \mid \text{Newtonian theory}).$$

Unlike "S has tuberculosis" and "S does not have tuberculosis," the GTR and Newtonian theory are not exhaustive. Of course, if we think of the likelihoods as merely reflecting subjective degrees of confidence, someone might assert, as an autobiographical remark, that the GTR has a higher likelihood than its negation; but someone else, with equal autobiographical sincerity, could assert the opposite. And both would be right if the probabilities involved were merely subjective. In science, we want more than this.[14]

Let me comment, finally, on $Pr(\text{observation})$, the unconditional probability of the evidence. In the case of the tuberculosis test, the unconditional probability of a positive test result can be estimated empirically. You can estimate how often people have tuberculosis and how often not; and you can estimate how often people in each group who take the test have positive test results. This allows you to estimate the value of $Pr(+ \text{ test result})$, since this quantity is defined as $Pr(+ \text{ test result} \mid \text{tuberculosis}) Pr(\text{tuberculosis}) + Pr(+ \text{ test result} \mid \text{no tuberculosis}) Pr(\text{no tuberculosis})$. But what of the comparable quantity in Eddington's test? What is the unconditional probability that starlight bends a certain amount during an eclipse of the type that Eddington studied? It isn't true that the prior probabilities on *GTR* and *notGTR* are reflected in the fact that a given proportion of the physical systems that populate our universe

[14] Earman (1992: 117) uses the Eddington example to illustrate the problem of assigning likelihoods to catchalls.

are relativistic while the rest are not. We can't estimate Pr(observation) by seeing how often starlight bends during eclipses. This reveals, incidentally, why it can be misleading to say that Pr(observation) describes how "unsurprising" the observations are. Even if it is true that starlight *always* bends the same amount during eclipses of the type that Eddington observed, this does not mean that Pr(observations) ≈ 1. The relevant question is what the average probability is of this observation *under each hypothesis considered*, where the average is taken by using the *prior probabilities* of the hypotheses.

Philosophical Bayesianism, Bayesian statistics, and logic

Bayesian philosophers of science assign prior probabilities to scientific theories like the GTR and do not hesitate to assign likelihoods to catchall hypotheses – for example, to the GTR's negation. They concede that there is a subjective element in these assignments, though they hasten to note that there are numerous subjective elements in frequentism as well (we will examine these in due course). Bayesian philosophers think that it is a matter of intellectual honesty to acknowledge subjective elements when they intrude. They are inevitable. What could justify pretending that they are not there?

Bayesian statisticians in their professional work rarely assign prior probabilities to "big" theories like the GTR and they rarely assign likelihoods to catchalls like *notGTR*. But both these practices are standard in connection with hypotheses that are more modest. For example, when Bayesians consider the genealogical relationships that humans, chimps, and gorillas might bear to each other (§4.8), they often assign equal priors to the three competing hypotheses $(HC)G$, $H(CG)$, and $(HG)C$. Given the observed similarities and differences that those three species exhibit, it is possible to compute the likelihoods of the three hypotheses and then to compute their posterior probabilities. The effect of assigning equal priors is that all the real work is done by the likelihoods; if the priors are equal, the hypothesis of greatest likelihood must also be the hypothesis that has the greatest posterior probability. Bayesians might just as well say that what interests them here is the likelihoods and make no judgment at all about priors or posteriors. A similar comment applies when Bayesian statisticians perform *sensitivity analyses*; by examining various assignments of priors, they calculate how changing the priors affects the calculated posterior probabilities. Here again, what one is learning about are the likelihoods of the hypotheses under study; given the likelihood ratio of

Roberto Mungue Steyer

H_1 to H_2, changing the ratio of priors will bring with it changes in the ratio of posterior probabilities. Describing these changes is just a way of describing the likelihood ratio.

Even though Bayesian statisticians often soft-pedal their assignments of prior probabilities to hypotheses, there is a deeper commitment on the part of Bayesians that concerns how likelihoods are sometimes computed. If a coin is tossed twenty times and seven heads are obtained, it is perfectly clear what the probability of that outcome is according to the hypothesis that the coin is fair (i.e., that $p = \frac{1}{2}$). But consider the hypothesis that the coin is *not* fair: i.e., that $p \neq \frac{1}{2}$. What is the probability of seven heads in twenty tosses according to this catchall? There are many ways the coin might fail to be fair, which correspond to different values of p, and these different values of p confer different probabilities on the observations. The likelihood of the hypothesis that $p \neq \frac{1}{2}$ is an *average* over the likelihoods of all the point values that p might have if it differs from $\frac{1}{2}$. This average takes the form of the following summation:

$$\begin{aligned} &Pr\left(7 \text{ heads} \mid p \neq \frac{1}{2} \;\&\; 20 \text{ tosses}\right) \\ &= \sum\nolimits_i Pr(7 \text{ heads} \mid p = i \;\&\; 20 \text{ tosses}) \\ &\quad \times Pr\left(p = i \mid p \neq \frac{1}{2} \;\&\; 20 \text{ tosses}\right). \end{aligned}$$

The hypothesis that $p \neq \frac{1}{2}$ is, in this respect, just like the negation of the GTR. Notice that priors on different values of p do not occur in this expression, but something rather like them does. As we will see, frequentists also consider hypotheses like $p \neq \frac{1}{2}$, but they do not compute the *average* likelihoods of those hypotheses. The handling of such hypotheses (which statisticians call "composite") is a fundamental divide that separates Bayesians from frequentists.

For Bayesian philosophers, rationality does not require you to deny the subjective elements that inevitably intrude in inference; rather, the point is to regulate that subjectivity in the right way. For them, being rational has to do with how you *change* what you believe as new evidence arrives; your starting point is not something that Bayesian philosophers feel they need to address. Bayesian philosophers often see Bayesianism as analogous to deductive logic in this respect (Howson 2001). Deductive logic does not tell you what you should take your premises to be; logic is solely in the business of giving advice on what follows from them. So, the fact that

priors and likelihoods are sometimes subjective is just a fact of life with which we all have to deal. Subjective Bayesians see themselves as facing these facts squarely in the face; they think their critics are ostriches burying their heads in the sand.

If Bayesianism is simply the logic that each of us should use to regulate our degrees of belief, the criticisms I have described of that philosophy do not apply. But an epistemology should do more than this. We need to identify which of our probability assignments can be justified interpersonally. And we also need to see if there are objective considerations that Bayesians ignore. The first of these tasks leads to likelihoodism; the second will lead us to consider frequentist ideas.

1.3 LIKELIHOODISM

Strength in modesty

The problems with Bayesianism just described suggest a fallback position that preserves much of what Bayesianism has to offer while abandoning the elements of the philosophy that are too subjective. This is likelihoodism. When prior probabilities can be defended empirically, and the values assigned to a hypothesis' likelihood and to the likelihood of its negation are also empirically defensible, you should be a Bayesian.[15] When priors and likelihoods do not have this feature, you should change the subject. In terms of Royall's three questions (§1.1), you should shift from question (2), which concerns what your degree of belief should be, to question (1), which asks what the evidence says. The likelihoodist does not answer this question by using the Bayesian concept of confirmation; you don't ask if the evidence raises, lowers, or leaves unchanged the hypothesis' probability. Rather, you compare only those hypotheses to each other that have determinate likelihoods. For example, instead of trying to compare the GTR to its own negation, you do what Eddington did: You compare the GTR with a specific alternative theory, Newtonian theory, and you use the law of likelihood (so named by Hacking 1965) to interpret the data:

Law of likelihood: The observations O favor hypothesis H_1 over hypothesis H_2 if and only if $Pr(O \mid H_1) > Pr(O \mid H_2)$. And the degree to which O favors H_1 over H_2 is given by the likelihood ratio $Pr(O \mid H_1)/Pr(O \mid H_2)$.

[15] Sometimes we can say what the value is of $Pr(O \mid H)$ without needing empirical information. For example, we know a priori (if we know anything a priori) that Pr(the next ball drawn will be green | 20 percent of the balls in the urn are green and the draw will be random) = 0.2.

The concept of *favoring* used in the law of likelihood involves a three-place relation that connects two hypotheses and a body of evidence. One also might call it the relation of *differential support*, although this terminology is apt to mislead; it may encourage the impression that the law of likelihood says that O supports H_1 to one degree, that O supports H_2 to another, and that the question is whether the first is greater than the second. This is not what the law means. According to likelihoodism, there is no such thing as the degree to which O supports a single hypothesis. Support is essentially *contrastive*.

The law of likelihood contains two ideas: a *qualitative assessment* of the bearing of the observations on the two hypotheses (expressed by an inequality) and a *quantitative measure* of how strongly or weakly the observations favor one hypothesis over the other (expressed by the likelihood ratio). The quantitative component goes beyond what the qualitative component says, just as the choice of a measure of degree of confirmation goes beyond the Bayesian definition of qualitative confirmation. And a similar question applies: even assuming that the qualitative law of likelihood is true, why should you use the likelihood *ratio* as your measure? The likelihoodist wants a measure of favoring that does not require any assignment of values to prior or posterior probabilities, or any assignment of values to the likelihoods of catchalls (if those values can't be defended by evidence), so that precludes using the possible definitions of degree of confirmation mentioned in §1.2. But why not define favoring in terms of the likelihood *difference*, $Pr(O \mid H_1) - Pr(O \mid H_2)$? One reason is suggested by a pattern that arises when there are multiple pieces of evidence that are independent of each other, conditional on each of the two hypotheses considered. Suppose, for example, that

$Pr(O_i \mid H_1) = 0.99$, for each of the 1,000 observations $O_1, \ldots, O_{1,000}$.
$Pr(O_i \mid H_2) = 0.3$, for each of the 1,000 observations $O_1, \ldots, O_{1,000}$.

With conditional independence, we have

$$Pr(O_1 \,\&\, \ldots \,\&\, O_{1,000} \mid H_1) = (0.99)^{1,000}$$
$$\text{and } Pr(O_1 \,\&\, \ldots \,\&\, O_{1,000} \mid H_2) = (0.3)^{1,000}.$$

The likelihood of each of these hypotheses, relative to the 1,000 observations, is very close to zero, so their difference is tiny; however, the ratio of the two likelihoods is $(33)^{1,000}$, which is huge. Since each of these 1,000 observations favors H_1 over H_2, the 1,000 observations should do so

more powerfully than any of them does singly. This recommends the likelihood ratio over the likelihood difference as a measure of strength of evidence. Can it be shown that the likelihood ratio is the best of all possible measures? Perhaps a compelling argument for this stronger conclusion can be given, or perhaps this part of the law of likelihood should be regarded as a postulate to be judged by the intuitiveness and usefulness of its applications. In any event, there is a feature of this example that will come up later in this chapter and in subsequent chapters as well: A probabilistic hypothesis such as H_1 can do an excellent job predicting what happens in each of 1,000 experiments, in each case assigning a very high probability to the outcome that in fact takes place. Yet, the likelihood of the hypothesis goes down and down as one triumph is laid upon another. This underscores the fact that it is the *relationship* between the likelihoods of different hypotheses that matters, not the *absolute value* of any single hypothesis' likelihood.

Since likelihoodism agrees that Bayesianism makes sense in many cases, we can consider how the Bayesian concept of *confirmation* is related to the law of likelihood's qualitative notion of *favoring* when both uncontroversially apply (e.g., in the example of tuberculosis diagnosis discussed in §1.2). For O to confirm H_1, it must be true that $Pr(O \mid H_1) > Pr(O \mid notH_1)$. The observation provides Bayesian confirmation of H_1 precisely when H_1 has a higher likelihood than its negation. In contrast, the favoring relation posited by likelihoodism need not pit H_1 against its own negation; the question is whether H_1 has a higher likelihood than H_2, for some alternative hypothesis H_2 that is of interest. Here's a simple example that illustrates how O can provide Bayesian confirmation of H_1 without O's favoring H_1 over a hypothesis H_2 that is incompatible with H_1:

Example 1: Let O = the card is red, H_1 = the card is a heart, H_2 = the card is a diamond. Then $Pr(O \mid H_1) = 1$, $Pr(O \mid notH_1) = \frac{1}{3}$, and $Pr(O \mid H_2) = 1$.

And here's an example that exhibits the opposite pattern in which O does not provide Bayesianism confirmation of H_1 though it favors H_1 over a hypothesis H_2 that is incompatible with H_1:

Example 2: let O = the card is a 7, H_1 = the card is a heart, H_2 = the card is the Ace of Spades. Then $Pr(O \mid H_1) = \frac{1}{13} = Pr(O \mid notH_1)$, and $Pr(O \mid H_2) = 0$.

The likelihoodist concept of favoring describes what the evidence says about the competition between any two hypotheses that both probabilify the data at hand. The Bayesian concept of confirmation addresses a special case; it describes what the evidence says about the competition between a hypothesis and its own negation. Both questions are of interest from a Bayesian point of view. On the other hand, if Bayesianism has the problems described in §1.2, we need the concept of favoring for those problem cases, since Bayesian confirmation will not be able to do the needed work.[16]

Three objections to likelihoodism

The law of likelihood is a proposal; it is not a mathematical theorem (like Bayes' theorem). The law proposes that the informal concept of favoring (or differential support) be explicated in terms of the formal concept of likelihood comparison. To judge this proposal, we must determine how well it conforms to, and renders precise and systematic, our use of the informal concept. Our goal here, familiar from other projects of philosophical explication, is not to exactly mimic the everyday concept, which may contain various ambiguities, opacities, incoherences, indeterminacies, and even contradictions (Carnap 1947, 1950). The philosopher's job is not the same as the lexicographer's.

The previous paragraph conveys a formula that philosophers often offer that describes how the definitions they propose ought to be judged, and there is something to it. However, something more is needed with respect to the case at hand. Something important would be missed if the law of likelihood were judged solely on the basis of how it clarifies the meaning of the English word "likely." As already noted, Fisher's use of the term "likelihood" is radically at variance with ordinary usage. However, this is not an objection to Fisher's *idea*, just a comment on the infelicity of his choice of *label*. What matters about the law of likelihood is whether it isolates an epistemologically important concept. The same is true of the likelihoodist's use of terms like "favoring" and "support." A formal proposal that describes how an informal concept should be understood is to be judged by the light it throws on the informal concept, but it also should be judged by the light it throws, period.[17]

[16] There is more to likelihoodism than I have described here. For example, there is the likelihood principle. For discussion of what this principle means and how it is related to the law of likelihood, see Grossman (unpublished). One difference is that the law of likelihood describes the bearing of a single data set on two hypotheses while the likelihood principle says when two data sets are evidentially equivalent.

[17] A similar point was already visible in the discussion of what "reliability" means in §1.2.

The need to restrict the law of likelihood

Suppose you are Madison's top meteorologist. You gather data on the present weather configuration in the Midwest and (let us suppose) you have at hand a true theory of how weather systems change. Your job is to make a weather forecast. Based on the information you have, you infer that the probability of snow in Madison tomorrow is 0.9. It would be natural for you to express this by saying that your information *supports* the prediction that there will be snow; and it also would be natural to say that your information *favors* the hypothesis that it will snow over the hypothesis that it will not. But here the support and the favoring reflect facts about the *probabilities* of hypotheses not about their *likelihoods*. What your data and theory tell you is that

$$Pr(\text{snow tomorrow} \mid \text{present data \& theory}) = 0.9$$
$$> Pr(\text{no snow tomorrow} \mid \text{present data \& theory}) = 0.1.$$

You are not computing whether

$$Pr(\text{present data} \mid \text{snow tomorrow}) > Pr(\text{present data} \mid \text{no snow tomorrow}).$$

Your data and theory favor your weather prediction by making it probable, not by giving it a likelihood higher than that of some competing hypothesis.

An even starker example is provided by the following example. Suppose you want to predict whether the next card dealt to you will be a heart. The dealer looks at this card and, before he turns it over and places it in front of you, says, "This is the Ace of Hearts." You know that the dealer is truthful. What, then, is your epistemic situation? You're interested in ascertaining the truth value of the hypothesis H = the next card is a heart. From what the dealer says, you know that proposition O is true where O = the next card is the Ace of Hearts. Should you compute the likelihood of H or the probability of H? The likelihood of H is:

$$Pr(O \mid H) = \frac{1}{13}.$$

The probability of H is

$$Pr(H \mid O) = 1.0.$$

Surely you should focus on the probability. And it would not be an abuse of language to say that the dealer's comment *strongly supports* the hypothesis that the next card will be a heart; what the dealer says *favors* that hypothesis over the hypothesis, say, that the next card will be a spade.

These examples and others like them would be good objections to likelihoodism if likelihoodism were not a fallback position that applies only when Bayesianism does not.[18] The likelihoodist is happy to assign probabilities to hypotheses when the assignment of values to priors and likelihoods can be justified by appeal to empirical information. Likelihoodism emerges as a statistical philosophy distinct from Bayesianism only when this is not possible. The present examples therefore provide no objection to likelihoodism; we just need to recognize that the ordinary words "support" and "favoring" sometimes need to be understood within a Bayesian framework in which it is the probabilities of hypotheses that are under discussion; but sometimes this is not so. Eddington was not able to use his eclipse data to say how probable the GTR and Newtonian theory each are. Rather, he was able to ascertain how probable the data are, given each of these hypotheses. *That's* where likelihoodism finds its application.

How can a preposterous hypothesis be extremely likely?

The gremlin example invites the following objection to the law of likelihood: The hypothesis that there are gremlins bowling in the attic has a likelihood that is as high as a likelihood can be; it has a value of 1. So, the law of likelihood says that the gremlin hypothesis is very well supported. But this is silly. The noises we hear do not make it at all likely that there are gremlins up there bowling. This is not a well-supported hypothesis at all. Hence, the law of likelihood is false.

The complaint that the gremlin hypothesis can't be "likely" or "well supported" is easily explained by the fact that the speaker assigns the gremlin hypothesis a very low prior. Imagine that the objector has inspected thousands of attics and has never seen a gremlin and that reputable authorities have assured him that gremlins are a myth. When he arrives at your house, his prior that there are gremlins bowling in your attic is low; once he hears the noises, his probability that there are

[18] Fitelson (2007) uses this kind of problem to argue that the law of likelihood is false and should be modified to read as follows: O favors H_1 over H_2 if and only if $Pr(O \mid H_1) > Pr(O \mid H_2)$ and $Pr(O \mid notH_1) < Pr(O \mid notH_2)$. This principle does not follow from the Law (notice that both are biconditionals), though if the right-hand side of Fitelson's modified principle is true, so is the right-hand side of the law of likelihood. Notice also that using Fitelson's principle requires one to have likelihoods for catchall hypotheses, which likelihoodism maintains are often unavailable.

gremlins up there bowling remains low, though the Bayesian must concede that the observation increases the hypothesis' probability.[19] This is why the objector judges that the gremlin hypothesis is not "likely," by which he means that it is not very probable. Fair enough, but that is not an objection to the law of likelihood. As noted, we need to recognize that Fisher's terminology was not well chosen. The terms "likely" and "probably" are used interchangeably in ordinary English, but that is not an objection to the law of likelihood.

Although Bayesians sometimes make this objection to the law of likelihood, the fact of the matter is that Bayesianism is committed to the view that likelihoods are the one and only vehicle by which observations can change the probabilities we assign to hypotheses. This was the point I discussed in connection with proposition (6). Bayesians as well as likelihoodists need a word to use in describing the epistemological significance of the fact that $Pr(E \mid H) > Pr(E \mid notH)$. The law of likelihood uses the word "favoring," and "differential support" might be used here as well. Of course, the law of likelihood also applies this term in a wider context, namely when one is comparing H with an alternative hypothesis other than its own negation. But the point of this term is not to assess the overall plausibility of H but to describe what a particular observation says about the competition between H and some alternative hypothesis. The law of likelihood does not say that the gremlin hypothesis is rendered plausible by the noise you hear.

Edwards (1972) discusses the same sort of objection in connection with another example. You draw a card from a deck and it turns out to be the seven of spades. Now consider the hypothesis that each of the cards in the deck is a seven of spades; this hypothesis has a likelihood of 1.0. In contrast, the likelihood of the hypothesis that the deck is "normal" is only $\frac{1}{52}$. This leads the law of likelihood to conclude that the card you've observed favors the stacked hypothesis over the normal hypothesis. But surely, the objection concludes, the stacked hypothesis is not more plausible or better supported. I leave it to the reader to construct and evaluate the likelihoodist's reply.

Likelihoodism and the definition of conditional probability

Likelihoodists think they have a philosophy that comes into its own when no evidence is available to back up assignments of prior probabilities. But

[19] To see this, consider the following consequence of Bayes' theorem: If H entails E and $0 < Pr(E) < 1$ and $0 < Pr(H) < 1$, then $Pr(H \mid E) > Pr(H)$.

how can this be true, given the Kolmogorov definition of conditional probability (§1.2)? Recall that the definition says that

$$\text{(K)} \qquad Pr(O \mid H) = \frac{Pr(O \;\&\; H)}{Pr(H)}.$$

There, in the denominator on the right-hand side, a prior probability has popped up, just what likelihoodists say they can do without when they talk about likelihoods!

The answer to this challenge is that likelihoodists should think of the Kolmogorov definition as correct only when various unconditional probabilities are "well defined." When they are not, the concept of conditional probability can and should be taken to stand on its own; it does not need to be defined in terms of unconditional probabilities. There are good reasons for this approach that do not depend on any qualms one might have about Bayesianism. For example, consider the fact that Kolmogorov's (K) says that the conditional probability is undefined if $Pr(H) = 0$. But surely there are contexts in which a conditional probability has a value even though the conditioning proposition has a probability of zero. Suppose I make you the following promise: If the coin I am about to toss lands heads, I will buy you a ticket in a fair lottery in which 1,000 tickets are sold. If the coin fails to land heads, you will have no ticket, and so you can't win the lottery. You know that I am trustworthy, so you conclude that Pr(you win the lottery | the coin lands heads) $= \frac{1}{1{,}000}$. However, I then take measures to ensure that the coin *cannot* land heads. Maybe I bend the coin, or place it in a tossing device that ensures tails every time, or maybe I just lock it in a vault and thereby ensure that the coin can never be tossed. If you buy the Kolmogorov definition of conditional probability, the information that the coin can't land heads should lead you to say that the conditional probability just stated is not correct. The value is not $\frac{1}{1{,}000}$; rather, it is *not defined.* On the other hand, if conditional probability is a primitive concept, the conditional probability can have the value given even though the conditioning proposition has a probability of zero (Hajek 2003). This position has the additional virtue of allowing Pr(the coin lands heads | the coin lands heads) to have a value of *unity* instead of being *not defined.*

There is an epistemic point that is also worth considering. We often know the value of $Pr(O \mid H)$ even though we have no clue as to the value of $Pr(H)$. As mentioned in §1.2, we can estimate the value of Pr(+ test result | tuberculosis) by giving the test to thousands of people whom we know have tuberculosis. This procedure does not require that we know how

common or rare tuberculosis is, and so we may be entirely in the dark as to the value of Pr(tuberculosis). The defender of Kolmogorov's definition is right to reply that proposition (K) is not a claim about *knowledge*; it does not say that to *know* the value of a conditional probability you first must *find out* the values of the two unconditional probabilities cited. (K) asserts a symmetric *mathematical* (or *logical*) dependence, not an asymmetric *epistemic* dependence. The right question to ask about Kolmogorov's (K) is whether there must exist unconditional probabilities for $H\&O$ and for H if there is such a thing as the conditional probability $Pr(H \mid O)$.

The answer depends on what we mean by probability and on the example we consider. Bayesians usually adopt the idealization that rational agents have degrees of belief for all the sentences of their language. The Bayesian framework is one in which a *complete probability function* is deployed over all the sentences in some language. If $O_1, O_2, \ldots O_n$, and $H_1, H_2, \ldots H_m$ are all sentences in the language, then the probability function assigns a prior probability to each of those atomic sentences and to all Boolean combinations definable from them (e.g., to the negations of each and to all disjunctions and conjunctions constructed from this set). Posterior probabilities are definable from the relevant priors via proposition (K). This is not the best way to understand what likelihoodists are up to. According to likelihoodism, the language we speak is far more wide-ranging than the probability models we use. On a given occasion, we may specify a value for $Pr(O \mid H_1)$ and for $Pr(O \mid H_2)$, but none for $Pr(O \mid notH_1)$, and none for $Pr(H_1)$ or $Pr(H_2)$. We use this *partial* probability function to do the needed work. Not only don't we *know* the value of $Pr(O \mid notH_1)$, or of $Pr(H_1)$, or of $Pr(H_2)$; in addition, there may be no such values to know. The model we use does not include these even as unknown quantities.

What likelihoodists mean by probability is not simply that an agent has some degree of belief. For one thing, the concept of probability needs to be interpreted more normatively. $Pr(O \mid H)$ is the degree of belief you *ought* to have in O given that H is true. But likelihoodists also like to think of these conditional probabilities as reflecting objective matters of fact. If Pr(the card is the Ace of Hearts | the card is dealt from this deck) $= \frac{1}{52}$, this is because of the physical composition of the deck and the physical properties of the process of dealing. When likelihoodists insist that probabilities must be "objective," they mean that probabilities must be grounded in such physical details.[20] When the physical processes at

[20] The word "objective" used by likelihoodists does not mean what so-called objective Bayesians have meant by the term: that probabilities must be derivable from logical features of the language we speak.

work generate frequency data, these data provide evidence we can use to infer the values of the underlying probabilities.[21]

Is Kolmogorov's (K) the right way to think about conditional probability when probability is understood in the way that likelihoodists propose? If there exists a physical process that leads people with tuberculosis who are tested to have a positive test result with a certain frequency, is there also a physical process that leads some people, but not others, to have tuberculosis? Arguably so, in which case $Pr(+$ test result | tuberculosis) and Pr(tuberculosis) will both figure in a useful model. But now consider Eddington. There was a physical process that led the light to bend during the eclipse; this is the process that the GTR purports to describe. But is there, in addition, a physical process whose result was that the GTR, or some competing theory, became true? Arguably not. If not, likelihoodists will not include $Pr(GTR)$ in their probability model. This is why your interpretation of probability should influence whether you regard Kolmogorov's (K) as a proper definition or just as a postulate that is true in favorable circumstances.

Kolmogorov's proposition (K), like Bayes' theorem, should be understood as having a certain rider attached. They do not assert that all the quantities they describe make sense. Rather, each of them should be understood in terms of the following preface: *in any model that uses the following quantities, here is how those quantities must be related.* When (K) is understood in this way, you can see that the following criticism is misguided: "If you assign a value to a hypothesis' likelihood, you are committed to saying that the hypothesis has a prior, whether you know its value or not."

The principle of total evidence

Bayesians and likelihoodists have their disagreements, but they agree on the principle of total evidence. This principle says that you should take account of everything you know. As stated, this idea is vague, but it gains precision when it is applied to concrete problems, as we shall see. It is a "pragmatic" principle in the philosophical sense of that term. This doesn't mean that it is something that cynics rather than idealists

[21] Although observed frequencies provide *evidence* concerning the values of probabilities, there are lots of contexts in which probabilities can't be *defined* in terms of (actual or hypothetical) frequencies; see Sober (1994, 2008b). For this reason, I prefer a "no-theory theory of probability," according to which probabilities are theoretical terms that cannot be defined in terms of observables.

embrace; rather, the point is that it gives advice about how probabilities should be *used* to solve problems. As there are many probability problems, the principle has many applications, and so the principle may be more plausible in some contexts than in others. I'll begin by describing a few settings in which the principle seems to make excellent sense. It will emerge in the next section that the principle of total evidence is controversial; it constitutes one of the fault lines that separate some central ideas in frequentism from both Bayesianism and likelihoodism.

Suppose two witnesses provide independent reports about what they saw at the scene of a crime. And suppose that each is at least minimally reliable in the sense described in §1.2, meaning that, for some relevant range of propositions:

$$Pr[W_i(P) \mid P] > Pr[W_i(P) \mid notP], \text{ for } i = 1, 2.$$

Here $W_i(P)$ means that witness i asserts that proposition P is true. The principle of total evidence says that you should take account of the testimony of *both* witnesses if that is the total evidence you possess. However, the principle is usually interpreted as saying that *more is better than less*; you should take account of both testimonies, rather than just one of them, even if there is more information available than what the two witnesses say.

Why are two witnesses better than one? If the witnesses agree that P is true, and the two witnesses go about their business independently,[22] the two pieces of testimony discriminate more powerfully between P and $notP$ than either of them does by itself, in the sense that

$$\frac{Pr[W_1(P) \;\&\; W_2(P) \mid P]}{Pr[W_1(P) \;\&\; W_2(P) \mid notP]} > \frac{Pr[W_i(P) \mid P]}{Pr[W_i(P) \mid notP]} > 1, \text{ for each } i = 1, 2.$$

This is because

$$\frac{Pr[W_1(P) \;\&\; W_2(P) \mid P]}{Pr[W_1(P) \;\&\; W_2(P) \mid notP]} = \frac{Pr[W_1(P) \mid P]}{Pr[W_1(P) \mid notP]} \times \frac{Pr[W_2(P) \mid P]}{Pr[W_2(P) \mid notP]}$$

and each of the ratios on the right is greater than one. This just reflects the common sense fact that two independent and (at least minimally) reliable witnesses who agree that P is true provide stronger evidence in favor of P than either witness does alone.[23]

[22] There can be (and will be!) a relation of *unconditional dependency* between what independent reliable witnesses say, in that $Pr[W_1(P) \mid W_2(P)] > Pr[W_1(P)]$. The relevant notion of independent witnesses is independence *conditional on the proposition reported*: $Pr[W_1(P) \;\&\; W_2(P) \mid P] = Pr[W_1(P) \mid P] \times Pr[W_2(P) \mid P]$.

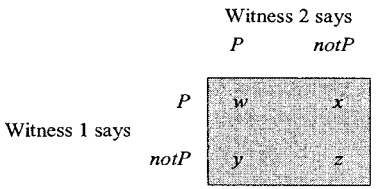

Figure 1.5 When two independent and reliable witnesses each report on whether proposition P is true, two yeses provide stronger evidence for P than one, and one yes provides stronger evidence than zero. Each cell represents the likelihood ratio $Pr(\text{testimony} \mid P)/Pr(\text{testimony} \mid notP)$ that goes with each of the four possible testimonies; $w > x$, $y > z$.

This example makes it look as if the principle of total evidence is justified by our hunger for strong evidence. But this can't be right. For suppose the two witnesses *disagree*. If you take both pieces of testimony into account, you may have no basis at all for discriminating between P and $notP$, whereas if you selectively focus on just one witness's testimony, you will. The principle of total evidence in this case tells you to resist the desire for telling evidence; if the total evidence says that you have little or no basis for discriminating between the two propositions, so be it.

When reliable witnesses reach their judgments independently of each other (conditional on P's being true and conditional on P's being false), this induces a kind of evidential *monotonicity*; if there are two witnesses, two votes for P provide stronger evidence that P is true than one vote would provide, and one vote provides stronger evidence for P than if neither witness had asserted that P is true. These comparisons are represented by the likelihood ratios depicted in Figure 1.5. As simple and familiar as this fact about multiple independent testimonies is, it is important to bear in mind that there is no rule written in Heaven that separate pieces of evidence must be independent. Suppose you are a cook in a restaurant. The waiter brings an order into the kitchen – someone in the dining room has ordered toast and eggs for breakfast. You wonder if this evidence discriminates between two hypotheses – that your friend Smith placed the order or that your friend Jones did so. You know the

[23] This point about multiple witnesses bears on Hume's analysis of the epistemology of reports about the alleged occurrence of miracles, on which see Earman's (2000) book and my review of it (Sober 2004d).

eating habits of each; the probabilities of different breakfast orders, conditional on Smith's placing the order, and conditional on Jones's placing the order, are shown in Figure 1.6. These probabilities give rise to the following curious fact: The order's being for *toast and eggs* favors Smith over Jones (since $0.4 > 0.1$); but the fact that the customer asked for *toast* provides no evidence on this question (since $0.5 = 0.5$); and the fact that the customer asked for *eggs* doesn't either (since, again, $0.5 = 0.5$). Here the whole of the evidence is more than the sum of its parts.

Figure 1.7 depicts the opposite pattern in which a new set of inclinations is attributed to your two friends. If Smith and Jones are disposed to behave as described, an order of *toast and eggs* fails to discriminate between the two hypotheses (since $0.4 = 0.4$). But the fact that the order included *toast* favors Smith over Jones (since $0.7 > 0.6$), and the same is true of the fact that the order included *eggs* (since $0.6 > 0.4$). Here the whole of the evidence is less than the sum of its parts.

Although the principle of total evidence says that you must use all the relevant evidence you have, it does not require the spilling of needless ink.

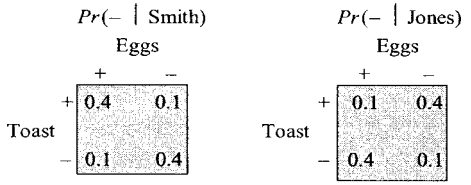

Figure 1.6 Smith and Jones differ in their inclinations to place different orders for breakfast. The breakfast order of toast and eggs provides evidence about which of them placed the order, although the fact that the order included toast does not, and neither does the fact that the order included eggs.

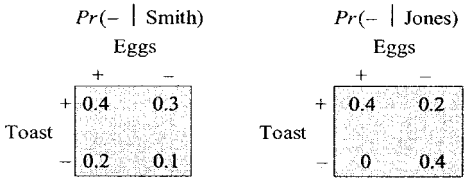

Figure 1.7 A new set of breakfast inclinations for Smith and Jones. Now the breakfast order of toast and eggs provides no evidence about which of them placed the order, though each part of the order favors Smith over Jones.

It does not require you to record irrelevant information. Consider the two hypotheses about coin tossing depicted in Figure 1.4. One of them says that $p = \frac{1}{4}$ while the other says that $p = \frac{3}{4}$, where p is the coin's probability of landing heads. I earlier described the data by saying that there were five heads in the twenty tosses of the coin. But why am I not obliged to describe the exact sequence of heads and tails that formed the data? There are many ways to get five heads in twenty tosses. A proposition that states just the sample frequency is *logically weaker* than a description of the exact sequence (in that the latter implies the former, but not conversely). Isn't it a violation of the principle of total evidence to use the sample frequency as a description of the data?

If we represent strength of evidence by the likelihood ratio, the answer is *no*. Consider each of the specific sequences in which there are five heads in twenty tosses. The two hypotheses we are considering ($p = \frac{1}{4}$ and $p = \frac{3}{4}$) agree that each of these exact sequences has a probability of $p^5(1-p)^{15}$ though they disagree about what the true value of p is. The likelihood ratio of $p = \frac{1}{4}$ to $p = \frac{3}{4}$, relative to a description of the exact sequence of heads and tails we observe, has the value:

$$\frac{Pr(\text{exact sequence} \mid p = \frac{1}{4})}{Pr(\text{exact sequence} \mid p = \frac{3}{4})} = \frac{(\frac{1}{4})^5(\frac{3}{4})^{15}}{(\frac{3}{4})^5(\frac{1}{4})^{15}} = 3^{10}.$$

If there are N exact sequences that can produce five heads in twenty tosses[24] the probability of obtaining *some sequence or other* in which there are five heads in twenty tosses has a value of $Np^5(1-p)^{15}$. Using this logically weaker description of the data, we obtain the following likelihood ratio:

$$\frac{Pr(5 \text{ heads} \mid p = \frac{1}{4})}{Pr(5 \text{ heads} \mid p = \frac{3}{4})} = \frac{N(\frac{1}{4})^5(\frac{3}{4})^{15}}{N(\frac{3}{4})^5(\frac{1}{4})^{15}} = \frac{(\frac{1}{4})^5(\frac{3}{4})^{15}}{(\frac{3}{4})^5(\frac{1}{4})^{15}} = 3^{10}.$$

Notice that the Ns have cancelled. There is no need to use the logically stronger description of the data that states the exact sequence of heads and tails, since it makes no difference to the likelihood ratio (Fisher 1922b; Hacking 1965: 80–1). In this sense, the sample frequency is a *sufficient* statistic. Notice the role played by the likelihood *ratio* in this argument; if you represented weight of evidence in some other way (e.g., via the

[24] N, the number of specific sequences in which there are m successes in n trials, is calculated by the formula for $\binom{n}{m}$, meaning *from n objects choose m*; $N = n!/m!(n-m)!$.

likelihood *difference*), maybe N would not disappear. Notice also how powerfully the data favor one hypothesis over the other, even though both say that the total data set was very improbable.

Whether the sample frequency is a sufficient statistic depends on the hypotheses being evaluated. In the example just described, the two hypotheses agree that tosses are independent of each other. But suppose this is something you want to test. And suppose further that the exact sequence of heads and tails is observed to be

H T H T H T H T H T H T H T H T H T

This sequence contains 50 percent heads, but it would be a mistake to think that this logically weakened description captures all the information in the data that is evidentially relevant. The *order* of heads and tails is evidentially relevant as well.

The logically weaker description of the data, the sample frequency, is a disjunction. One of the disjuncts describes the exact sequence that *did* occur; the other disjuncts describe exact sequences that *did not*. When $p = \frac{1}{4}$ and $p = \frac{3}{4}$ are the two hypotheses under test, there is nothing wrong with describing the data in this disjunctive form, saying that this sequence *or* that sequence *or* that other sequence was the one that occurred without saying which. The principle of total evidence is not a rule against disjunctions. Rather, the rule says that logically weakening your description of the data is not permitted when this changes your assessment of what the evidence indicates. Applying the principle requires a rule for interpreting what the evidence says about the hypotheses under test. At this point, likelihoodists appeal to the law of likelihood and use the likelihood ratio. Bayesians can agree with the above argument, since for them the likelihood ratio is *the* vehicle by which ratios of priors are transformed into ratios of posterior probabilities, as proposition (6) attests. Likelihoodists and Bayesians are on the same page when it comes to the principle of total evidence.[25]

The limits of likelihoodism

Likelihoodism addresses the first of Royall's three questions (§1.1) while remaining silent on the other two; it confines itself to the task of interpreting what the evidence says while giving no advice on what you should

[25] I will not try to address the deeper question of what the ultimate justification is of the principle of total evidence. I. J. Good (1967) provides a decision-theoretic justification.

believe or do. Even so, the question remains of whether likelihoodism accomplishes the relatively modest goal it sets for itself. The problem is that there are many scientific hypotheses of interest that are *composite*, rather than *simple*. These are technical terms. The two hypotheses about the coin (that $p = \frac{1}{4}$ and that $p = \frac{3}{4}$) depicted in Figure 1.4 are both simple in the sense that each says exactly how probable each possible outcome of the experiment is. Composite hypotheses are more ambiguous; they circumscribe a *family* of probabilities that an observation might have without singling out just one. An example would be the hypothesis that $p > \frac{1}{4}$; this hypothesis does not say what the probability is of observing exactly five heads in twenty tosses. There are many values that p might have if it exceeds $\frac{1}{4}$, and each specific value has its own likelihood relative to a given observation; composite hypotheses are disjunctions (sometimes infinite disjunctions) of simple hypotheses.

Hypotheses that look as if they are composite can in reality turn out to be statistically simple, if background information of a certain sort is available. Imagine that there are three kinds of coins that a factory manufactures – a third have $p = \frac{1}{4}$, a third have $p = \frac{1}{2}$, and a third have $p = 1.0$. If you chose a coin made at this factory at random, then if the coin before you has $p > \frac{1}{4}$, there are just two possibilities – that $p = \frac{1}{2}$ and $p = 1.0$ – and these are equiprobable. The average of these is $p = \frac{3}{4}$. Likelihoodists have no problem with assessing the hypothesis that $p > \frac{1}{4}$ in this kind of context. True to their antisubjectivist inclinations, they are happy to consider this hypothesis because there is an objective answer to the question of what observations we should expect to make if the hypothesis that $p > \frac{1}{4}$ is true. Absent this kind of information, they decline to assess the hypothesis at all. Rather, they relegate $p > \frac{1}{4}$ to the same epistemic limbo to which they consign *notGTR*, the catchall hypothesis that the GTR is false.

It is arguable that science often does not need to assess how the evidence bears on such catchall hypotheses. Eddington was able to compare the GTR with Newtonian theory, and maybe that is enough. However, other composite hypotheses seem to play a central role in the activity of science, so the likelihoodist denial that they can be handled should raise more eyebrows. For example, population geneticists often want to say whether the gene-sequence data gathered from a number of species favor the hypothesis of random genetic drift or the hypothesis of selection. The drift hypothesis is often statistically simple: For example, with respect to the two alleles A and a that might exist at a given genetic locus, the drift hypothesis says that they are identical in fitness. It says that $w_A = w_a$,

which means that $w_A - w_a = 0$. In contrast, the hypothesis of selection is composite; it says that $w_A \neq w_a$; in other words, it says that $w_A - w_a = \theta$, where θ is a parameter whose value is not equal to zero. Notice that there are many different values that θ might have if it isn't equal to zero. Each of these specific values for θ entails its own probability for the data at hand. But what does the bare hypothesis of selection itself predict? As the previous example about the coin factory suggests, this question would be answerable if we had an objective basis for assigning probabilities to the different values θ might take if it were nonzero. But, alas, we often lack this type of information. For this reason, it is often impossible to compare drift with selection within the framework of likelihoodism. Although physicists may be content to compare the GTR with Newtonian theory and to feel no need to ponder the catchall hypothesis that the GTR is false, population geneticists have wanted to test drift against selection and have even claimed to have done so. We will examine the question of whether and how this is possible in Chapter 3. For now, the point is that we have isolated an issue that unites Bayesians and frequentists; these two old enemies maintain that likelihoodism is too austere. Frequentists think they have good methods for testing composite hypotheses and Bayesians deny that the hypotheses in question are really composite. Both rush in where likelihoodists fear to tread.

1.4. FREQUENTISM I: SIGNIFICANCE TESTS AND PROBABILISTIC *MODUS TOLLENS*

I began this chapter by painting with a broad brush. I said that Bayesians hold that science is in the business of determining which theories are probably true while frequentists hold that this is not at all what science is about. I then complicated the story by adding likelihoodists to the cast of characters. They often eschew the goal of assigning probabilities, but in many respects they are more like Bayesians than frequentists, as we now will see. The fact that there are three positions here, not two, complicates the problem of saying what frequentism amounts to. It is not enough to say that frequentists reject the goal of assigning probabilities to hypotheses, since that point, though correct, does not separate them from likelihoodists. What can be said that is distinctive of what frequentism is *for*? We will uncover some of its differences with the other two philosophies in due course. But we must bear in mind that frequentism is not a single unified theory. Rather, it is a motley of different techniques that are often only loosely connected with each other; sometimes they are even in

conflict. In §1.2, I mentioned that Bayesianism gives *epistemological* advice about probability assignments; what probability statements *mean* (which "interpretation of probability" is correct) is a separate, semantic, question. A similar point applies to frequentism. Frequentism is not the thesis that probability statements are claims about actual or hypothetical frequencies, though this semantic thesis is something that many frequentists endorse. Rather, frequentism is a thesis about epistemology. Frequentists assess a rule of inference by examining the (expected) frequencies of good and bad outcomes when the rule is applied repeatedly.

The first frequentist method that I want to consider is R. A. Fisher's idea of *significance tests*. Fisher conceived of this procedure as a corrective to what he thought was wrong with the Neyman–Pearson theory of hypothesis testing, which I'll discuss in the next section. I take these two approaches in reverse chronological order because Fisher's theory is in some ways easier to grasp than the Neyman–Pearson approach and because its contrast with likelihoodism is more obvious.

To get started, let's consider a simple rule of deductive reasoning, *modus tollens*. This is a form of argument familiar to philosophers and scientists; it is the centerpiece of Karl Popper's views on falsifiability (which I'll discuss in §2.8):

(MT)
$$\frac{\text{If } H, \text{ then } O \quad \textit{not}O}{\textit{not}H}$$

Modus tollens, like other rules of deductive logic, says what follows from what. It does not, in the first instance, give advice. Still, it is natural to interpret *modus tollens* as saying that if the hypothesis H entails the observation statement O, and O turns out to be false, then H should be *rejected*. I use a single line to separate premises from conclusion to indicate that *modus tollens* is deductively valid (meaning, recall, that if the premises are true, the conclusion must be). Since (MT) is valid, perhaps the following "probabilistic extension" of the rule constitutes a sensible principle of nondeductive reasoning:

(Prob-MT)
$$\frac{Pr(O \mid H) \text{ is very high} \quad \textit{not}O}{\textit{not}H}$$

According to *probabilistic modus tollens*, if the hypothesis H says that O will *very probably* be true, and O turns out to be false, then H should be

rejected. Equivalently, the suggestion is that if H says that some observational outcome (*not*O) has a very low probability, and that outcome nonetheless occurs, then we should regard H as false. I draw a double line between premises and conclusion in (Prob-MT) to indicate that the argument form is not supposed to be deductively valid. But maybe it is a sensible form of inference nonetheless.

Before addressing whether probabilistic *modus tollens* is correct and how it is related to deductive *modus tollens*, I want to discuss a parallel question. Consider *modus ponens*:

(MP)
$$\frac{\text{If } O, \text{ then } H \qquad O}{H}$$

Modus ponens is deductively valid, and this may suggest that the following probabilistic extension of the principle is also correct:

(Prob-MP)
$$\frac{Pr(H \mid O) \text{ is very high} \qquad O}{\overline{\overline{H}}}$$

(Prob-MP) says that if O renders H very probable, and O is true, then we should accept H. My brief comments in §1.2 on the lottery paradox suggest that we should be wary of this rule of acceptance. But (Prob-MP) has a close cousin, which we have already examined:

(Update)
$$\frac{Pr_{\text{then}}(H \mid O) \text{ is very high} \qquad O \qquad O \text{ is all the evidence we have gathered between then and now.}}{Pr_{\text{now}}(H) \text{ is very high}}$$

This is nothing other than the rule of updating by strict conditionalization. (Update) is a sensible rule, and it also has the property of being a generalization of deductive *modus ponens*. By parity of reasoning, should we conclude that probabilistic *modus tollens* is a good rule because it generalizes deductive *modus tollens*?

Friends of (Prob-MT) need to say where the probability cutoff for rejection is located. How low must $Pr(O \mid H)$ be for O to justify rejecting H? Richard Dawkins (1986: 144–6) addresses this question in the context of discussing how theories of the origin of life should be evaluated. He says that an acceptable theory can say that the origin of life on Earth was somewhat improbable, but it cannot go too far. If there are n planets in the universe that are "suitable" locales for life to originate, then an acceptable theory of the origin of life on Earth must say that that event had a probability of at least $\frac{1}{n}$. Theories that say that terrestrial life was less probable than this should be rejected. Creationists also have set cutoffs. For example, Henry Morris (1980) says that theories that assign to an event a probability less than $\frac{1}{10^{110}}$ should be rejected, and William Dembski (2004) says that a theory that assigns to a "specified event" (a technical term in Dembski's framework) a probability less than $\frac{1}{10^{150}}$ should be rejected.[26] Morris and Dembski obtain these numbers by attempting to calculate how many times elementary particles could have changed state since the universe began.

Dawkins, Dembski, and Morris have all made the same mistake. It isn't that they have glommed on to the wrong cutoff. The problem is deeper: *There is no such cutoff.* Probabilistic *modus tollens* is an incorrect form of inference (Hacking 1965; Edwards 1972; Royall 1997). Lots of perfectly reasonable hypotheses say that the observations are very improbable. As noted earlier, if H confers a very high probability on each of the observations $O_1, O_2, \ldots, O_{1,000}$ (but a probability that is short of unity), it will confer a very low probability on their conjunction, if the observations are independent of each other, conditional on H. A probability that is very large but less than one, when multiplied by itself a large number of times, will yield a very small probability. Adopting probabilistic *modus tollens* would have the effect of eliminating all probabilistic theories from science once they are repeatedly tested.

It may seem that the kernel of truth in (Prob-MT) can be rescued by modifying the argument's conclusion. If it is too much to conclude that H is false, perhaps we should conclude just that the observations constitute evidence against H:

(Evidential Prob-MT)
$$\frac{Pr(O \mid H) \text{ is very high.} \qquad \textit{not}O}{\overline{\overline{\textit{not}O \text{ is evidence against } H.}}}$$

This principle is also unsatisfactory, as an example from Royall (1997: 67) nicely illustrates. Suppose I send my valet to bring me one of my urns.

[26] For discussion of Dembski's (1998) framework for inferring the existence of intelligent designers, see Fitelson et al. (1999).

I want to test the hypothesis (H) that the urn he returns with contains 0.2 percent white balls. I draw a ball from the urn and find that it is white. Is this evidence against H? It may not be. Suppose I have only two urns; one of them contains 0.2 percent white balls, while the other contains 0.01 percent white balls. In this instance, drawing a white ball is evidence *in favor* of H, not evidence *against* it.[27]

The use of genetic data in forensic identity tests provides a further illustration of Royall's point. Suppose that two individuals match at twenty independent loci; they are heterozygotes at each. At each locus, each individual has one copy of a rare allele (frequency = 0.001) and one copy of the alternative, common, allele (frequency = 0.999). The probability of this twenty-fold matching, if the two individuals are full sibs, is about $[(0.001)(0.5)]^{20}$. This is a very small number, but that hardly shows that the sib hypothesis should be rejected. In fact, the data *favor* the sib hypothesis over the hypothesis that the two individuals are unrelated. If they are unrelated, the probability of the observations is about $[(0.001)(0.001)]^{20}$. The two likelihoods are both very small, but the first is 500^{20} times larger than the second (Crow et al. 2000: 65–7).[28]

These examples reflect a central idea in the likelihoodist theory of evidence: judgments about evidential meaning are essentially *contrastive*. To decide whether an observation is evidence against H, you need to know what the alternative hypotheses are; *to test a hypothesis requires testing it against alternatives*.[29] In the story about the valet, observing a white ball is very improbable according to H, but in fact that outcome is evidence *in favor of* H, not evidence against it. This is because O is even more improbable according to the alternative hypothesis. Probabilistic *modus tollens*, in both its vanilla and evidential versions, needs to be replaced by the *law of likelihood*. The relevance of this point is not confined to urn problems and forensic DNA. It will play an important role in Chapter 4

[27] A third formulation of probabilistic *modus tollens* is no better than the other two. Can one conclude that H is *probably* false, given that H says that O is highly probable, and O fails to be true? The answer is *no*; inspection of Bayes' theorem shows that $Pr(notO \mid H)$ can be low without $Pr(H \mid notO)$ being low.

[28] Notice how the likelihood *ratio*, not the likelihood *difference*, figures in this argument.

[29] There are two exceptions to the thesis that testing is always contrastive. If a true observation statement entails H, there is no need to consider alternatives to H; you can conclude without further ado that H is true; this is just *modus ponens*. And if H entails O and O turns out to be false, you can conclude that H is false, again without needing to contemplate alternatives; this is just *modus tollens*. It is a separate question how often these forms of argument apply to testing in science. They rarely do. Observations almost never entail theories, and theories almost never entail observations. More on this later.

when we consider the question of why the similarities observed in two or more species is evidence for those species' having a common ancestor. Within the framework developed there, an observed similarity O provides *stronger* evidence in favor of the common ancestry (CA) hypothesis the *lower* the value is of $Pr(O \mid CA)$. The reason the evidence for CA is strengthened by lowering the value of this conditional probability is that lowering the value of $Pr(O \mid CA)$ leads the value of $Pr(O \mid SA)$ to plunge even more; here SA is the hypothesis of separate ancestry.

There is a reformulation of probabilistic *modus tollens* that makes sense, but it is Bayesian:

(Bayesian Prob-MT)

$Pr_{\text{then}}(O \mid H)$ is very high.
$Pr_{\text{then}}(O \mid notH)$ is very low.
$Pr_{\text{then}}(H) \approx Pr(notH)$
not-O

$Pr_{\text{now}}(H)$ is very low.

Although the conclusion of this argument follows *deductively* from the premises (given the rule of updating by strict conditionalization and that $notO$ is all you learned between then and now), this is a form of argument that frequentists will not touch with a stick. The reason is not that it is invalid (it is not) but that it requires premises that frequentists regard as too subjective.[30]

Fisher's (1959) test of significance is a version of probabilistic *modus tollens* and that is bad enough. But it has the additional defect that it violates the principle of total evidence. In a significance test, the hypothesis you are testing is called the "null" hypothesis, and your question is whether the observations you have are sufficiently improbable according to the null hypothesis. However, you don't consider the observations in all their detail but rather the fact that they fall in a certain region. You use a logically weaker rather than a logically stronger description of the data. Here's an example (from Howson and Urbach 1993: 176) that illustrates the point. You want to test the hypothesis that a coin is fair (i.e., the hypothesis that the probability of heads is 0.5) by tossing the coin twenty times. Assume that the tosses are independent of each other. Suppose you obtain four heads. You then compute the

[30] Wagner (2004) shows that a bound on the value of $Pr(notH)$ can be derived from the values of $Pr(O \mid H)$ and $Pr(notO)$; he calls his result a probabilistic version of *modus tollens*. This is not the probabilistic *modus tollens* whose nonexistence I argue for above.

probability of a disjunction in which "four heads" is one of the disjuncts. You need to look at all the outcomes that the null hypothesis says are *at least as improbable* as the one you actually obtained:

$$Pr(\text{0 or 1 or 2 or 3 or 4 or 16 or 17 or 18 or 19 or 20 heads} \mid \text{the coin is fair and the coin is tossed 20 times}) = p.$$

The probability of this disjunction, conditional on the null hypothesis, is called the *p*-value for the test outcome.

This *p*-value has two interpretations, corresponding to two different conceptions of what a significance test is supposed to accomplish. Sometimes significance testers draw a conclusion as to whether the null hypothesis should be rejected. To do this, they specify a value for α, the "level of significance"; the null hypothesis is rejected if the *p*-value is less than this cutoff. If $\alpha = 0.05$ is your level of significance, then four heads in twenty tosses will suffice to reject the null hypothesis, since the *p*-value of this outcome is 0.012; had you obtained six heads in twenty tosses, this outcome would not suffice to reject the null, since the *p*-value in this instance is 0.115. It is generally conceded that choosing a value for α is an arbitrary matter of convention. The other interpretation of significance tests is that they measure the strength of the evidence against the null hypothesis; the lower the *p*-value of the outcome, the stronger the evidence against. This comparative idea, by itself, does not say whether six heads in twenty tosses is (in an absolute sense) evidence against the hypothesis that the coin is fair, but it does say that four heads in twenty tosses would be stronger evidence against it. If we stipulate that a *p*-value of 0.05 is the cutoff between "strong evidence against the null hypothesis" and not, then we know how to interpret six heads in twenty tosses, and also how to interpret four in twenty and two in twenty. The first of these is not strong evidence against the null while the second and third are. There is arbitrariness here as well.

Both interpretations of significance tests are vulnerable to the fact that there are many descriptions of the data that might be used, and changing these can lead to different conclusions about the null hypothesis. I mentioned that obtaining six heads in twenty tosses does not allow you to reject the null hypothesis (if you set $\alpha = 0.05$), since the probability of obtaining between zero and six or between fourteen and twenty heads is greater than 0.05. In this example, we thought of each possible number of heads that might occur in twenty tosses (0, 1, 2, ... 18, 19, 20) as an element in the outcome space and then gathered

together the fourteen elements there that each has a probability of occurring under the null hypothesis that is less than or equal to the probability of obtaining exactly six heads. But the outcome space can be sliced up differently.[31] For example, instead of having twenty-one categories, you might decide to collapse some of these together. If you combine five heads and ten heads into one category, and fourteen heads and fifteen heads into another, you now have an outcome space with nineteen categories, not twenty-one. If you then construct a disjunction of the categories from this list that each has a probability that is less than or equal to the probability of getting exactly six heads, you'll discover that the probability of the relevant disjunction under the null hypothesis is 0.49, which will lead you to reject the null hypothesis (Howson and Urbach 1993: 182–3). Whether you reject the null depends on how you slice the cake.

It might be objected that collapsing the twenty-one categories into these nineteen is "unnatural," or that finer-grained taxonomies are preferable to ones that are coarser-grained. Defenders of significance tests have not attempted to develop an account of naturalness, and it is unclear how much help significance tests could extract from such an account. However, it is abundantly clear that insisting on logically stronger descriptions of the data does not help the significance tester. Instead of having twenty-one categories in the outcome space, why not treat each specific sequence of heads and tails as a separate element, with the result that our outcome space now has 2^{20} members, each with the same probability under the null hypothesis of $(\frac{1}{2})^{20}$? When we obtain a specific sequence of heads and tails (say, one containing two heads) and then collect the other elements in the outcome space that are no more probable according to the null hypothesis, the result is that we construct a disjunction that contains *all* 2^{20} elements; the probability of this disjunction, under the null hypothesis, is unity. With this fine-grained outcome space, we'll never reject the null, no matter what the outcome is.

Turning now to the evidential interpretation of significance tests, it is important to see how it conflicts with likelihoodism. According to the law of likelihood, whether the observations are evidence against the hypothesis that the coin is fair depends on which alternative hypothesis you consider. If the alternative to the null hypothesis says that the probability of heads is 0.8, then observing four heads in twenty tosses will

[31] Compare this point with considerations about cake slicing that arise in connection with the principle of indifference (§1.2).

be evidence *in favor* of the null hypothesis, not evidence *against* it. If the modest principle stated in §1.1 is correct, this point also bears on the idea that significance testing provides a rule of rejection. If an observation justifies you in rejecting H, and you were not justified in rejecting H before you obtained the observation, then the observation must be evidence against H. The fact that significance tests don't contrast the null hypothesis with alternatives suffices to show that they do not provide a good rule for rejection.

Another odd property of significance tests concerns the way in which they are sensitive to sample size. Howson and Urbach (1993: 208–9) explain this point by describing a nice example due to Lindlay (1957). Suppose you wish to test the hypothesis (H_1) that 40 percent of the marbles in an urn are red. If you examine ten balls and choose $\alpha = 0.05$, you will reject H_1 if you see seven or more red balls. If you examine 100 balls and choose the same value for α, you will reject H_1 if you observed more than forty-eight red balls. And if you examine 1,000 balls, again with $\alpha = 0.05$, you will reject H_1 if you observe more than 403 red balls. As sample size increases, the observed frequency must be closer and closer to 40 percent for you to not reject H_1. With ten balls, you need to observe less than 70 percent; with 100 you need to observe less than 48 percent; and with 1,000, you need to observe less than 41 percent. This may not seem strange until you add the following detail. Suppose the alternative to H_1 is the hypothesis (H_2) that there are 60 percent red balls in the urn. The law of likelihood now entails that observing fewer than 50 percent red favors H_1 over H_2, that observing more than 50 percent red has the opposite evidential significance, and that *these interpretations of the observations are correct at all sample sizes.* If the law of likelihood is right, and if the modest principle stated in §1.2 correctly describes the connection between evidence and rejection, then we have here an objection to significance tests.

Although I have criticized the rejection and the evidential interpretations of significance tests, there is a more modest interpretation that is beyond reproach. Fisher (1956: 39, 43) put the point like this: If H says that O is very improbable, and O occurs, then we know that a disjunction is true – either H is false or something very improbable has occurred. This disjunction *does* follow. However, what does *not* follow is the first of Fisher's disjuncts; nor does it follow that we have obtained evidence against H. Another modest interpretation of significance tests is also appropriate: An observational outcome that a hypothesis says is very improbable may prompt you to search for a different hypothesis that says that the outcome

was less surprising. This is how I understand the following remark that Gossett made in the 1930s:

> [a significance test] doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the [p-value] is very small, say .00001; what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say 0.05 [...] you will be very much more inclined to consider that the original hypothesis is not true. (quoted in Hacking 1965: 83)

This gentle suggestion has good likelihoodist credentials.

If probabilistic *modus tollens* and significance tests have the flaws just described, can we abandon the *probabilistic* and simply rely on the *deductive* form? If H_1 entails O and O turns out to be false, it follows that H_1 is false. If H_2 is the only alternative to H_1, it further follows that H_2 is true. This is the pattern of reasoning that Sherlock Holmes endorses in *The Sign of Four* where Sir Arthur Conan Doyle has his hero say that "when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth." The *correctness* of this pronouncement is not in dispute; rather, it is the *applicability* of Holmes's dictum that I contest. In science, it is rarely the case that the hypotheses under test deductively entail observational claims. This is obvious in the case of hypotheses that use the concept of probability (as in my running example of the hypothesis that a coin is fair). But the point often holds when hypotheses make no mention of probability. For example, when Eddington tested Newtonian theory against relativity theory, the competing hypotheses did not provide point predictions about what he should observe when he measured the bend in starlight during a solar eclipse. Because his measurements were imprecise, he could say only that the observations would *probably* fall in one value range if Newtonian theory were true and that they would *probably* fall in a second interval if relativity theory were true. The pervasive pattern in science is that hypotheses confer (nonextreme) probabilities on observations.[32]

It may seem not to matter much whether a hypothesis says that O *cannot* occur or says only that O *very probably* will not occur. In fact, the difference is profound. If you observe that O is true, the former allows you to reject H without your needing to consider an alternative hypothesis. In contrast, the latter does not license rejection, and there is

[32] The fact that scientific theories typically confer probabilities on observations only when auxiliary information is added will be explored in the next chapter in connection with Duhem's thesis.

no saying whether the observation is evidence against H unless an alternative hypothesis is specified.

1.5 FREQUENTISM II: NEYMAN–PEARSON HYPOTHESIS TESTING

The theory of hypothesis testing set forth by Neyman and Pearson (1933), and subsequently developed in detail by Neyman, gives advice about rejection, not, in the first instance, advice about the interpretation of evidence. As noted in §1.1, Neyman and Pearson state that they are not interested in interpreting evidence but only in stating general rules for guiding "behavior." This claim notwithstanding, the interpretation of evidence and the rational acceptance and rejection of hypotheses *are* related if the modest principle enunciated earlier is correct; if learning that O is true justifies rejecting H, where the rejection of H was not justified before that knowledge was gained, then O must be evidence *against* H. The Neyman–Pearson theory, as we will see, violates this principle.

If you are going to decide whether to accept or reject a hypothesis in the light of a set of observations, there are two kinds of error to which you are vulnerable. Consider the tuberculosis test discussed earlier, but this time let's frame the problem in terms of the task of acceptance and rejection, not as a question concerning the interpretation of evidence. You, the physician, receive the report of your patient's tuberculosis test result. The report is either positive or negative, and the patient either has tuberculosis or does not. You have two options: You can accept the hypothesis that your patient has tuberculosis or you can reject it. There are two kinds of error you might commit: You might reject the hypothesis that he has tuberculosis when it is true, or you might accept the hypothesis when it is false. These options are depicted in Figure 1.8, as are

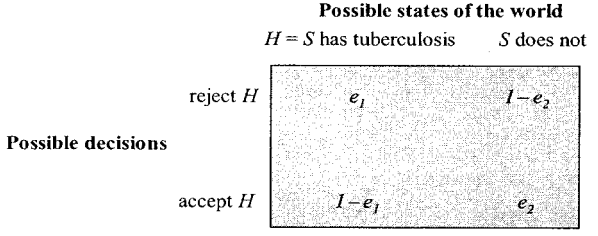

Figure 1.8 S either has tuberculosis or does not, and you, the physician, must decide whether to accept or reject the hypothesis H that S has tuberculosis. The four cells represent four possibilities; cell entries represent probabilities of the form Pr(decision | state of the world).

the probabilities of mistaken rejection (e_1) and of mistaken acceptance (e_2). If you were to ignore the report and merely toss a coin, your two error probabilities would then each have a value of 0.5. But you can do better if the test procedure you use is *reliable* in the sense described in §1.2; the more reliable the procedure, the smaller the error probabilities are. However, this does not mean that you will *probably* get the right answer if you use a reliable procedure. The error probabilities are of the form Pr(you accept hypothesis H | H is false) and Pr(you reject hypothesis H | H is true); they do not represent $Pr(H$ is false | you accept $H)$ and $Pr(H$ is true | you reject $H)$. Neyman–Pearson hypothesis testing is frequentist, not Bayesian.

Neyman–Pearson theory begins with the truism that it is better to have smaller error probabilities than larger ones. If you are going to base your decision about your patient's condition on what a test result says, you'll do better by using a more reliable testing procedure than one that is less. For example, suppose you can use a test kit that is made in Madison or one that is made in Middleton, where the two pairs of error probabilities are:

The Madison test kit: $Pr(-$ test result | S has tuberculosis$) = 0.02$.
$Pr(+$ test result | S does not have tuberculosis$) = 0.01$.

The Middleton test kit: $Pr(-$ test result | S has tuberculosis$) = 0.04$.
$Pr(+$ test result | S does not have tuberculosis$) = 0.03$.

Surely you'd want to use the Madison test kit, since both its error probabilities are lower. But how should you choose between the Madison kit and one made in Prairie du Chien? The error probabilities of this third test kit are:

The Prairie du Chien test kit: $Pr(-$ test result | S has tuberculosis$) = 0.01$.
$Pr(+$ test result | S does not have tuberculosis$) = 0.02$.

To choose between Madison and Prairie du Chien, you must decide which kind of error is worse to commit. Is it more important to avoid accepting that S has tuberculosis when he does not, or to avoid rejecting the hypothesis that S has tuberculosis when he does? One obvious way to decide this is to think about how your actions will be influenced by what you believe. Is it worse to treat someone for tuberculosis when he doesn't have the disease, or to fail to treat someone for tuberculosis when he does?

Notice how ethical considerations figure in this question. The issue is not strictly epistemological. In terms of Royall's three questions (§1.1), we are edging towards question (3) and away from questions (1) and (2).

The Neyman–Pearson theory recognizes that there are two types of error, but it does not treat them the same. First, you choose which of the two hypotheses under test you'll regard as the "null hypothesis." You then decide how large an error probability you will tolerate in connection with mistakenly rejecting the null:

$$Pr(\text{reject the null hypothesis} \mid \text{the null hypothesis is true}) < \alpha.$$

Scientists usually choose a value of $\alpha = 0.05$ while recognizing that this choice is arbitrary. The probability of rejecting the null hypothesis when it is true is called a Type-1 error. After putting an upper limit on how much Type-1 error you are prepared to tolerate, you then try to minimize the probability of the other kind of error:

$$Pr(\text{accept the null hypothesis} \mid \text{the null hypothesis is false}) = \beta.$$

The mistake of accepting the null hypothesis when it is false is a Type-2 error. So there are three steps in the Neyman–Pearson process: Decide which hypothesis is the null; set an upper limit on the probability of Type-1 error; and then minimize the probability of Type-2 error.

Suppose you decide that "*S* has tuberculosis" is your null hypothesis and you chose a value for α of 0.05; given these choices, all three test kits are acceptable so far. But now you want to minimize β. Madison does better on this score than either Middleton or Prairie du Chien. On the other hand, if you decide that "*S* does not have tuberculosis" is the null hypothesis while still hewing to the convention that $\alpha < 0.05$, you'll end up opting for the test kit from Prairie du Chien. Different decisions about what the null hypothesis is lead to different test procedures. Here is some more terminology: α (the probability of Type-1 error) is called the "size" of your test and $(1-\beta)$ is called its "power." Neyman–Pearson testing treats these asymmetrically: First get the size below some threshold, then maximize power.

To apply this framework in a way that brings out how it is related to likelihoodism, let's return to the coin-tossing problem discussed earlier. Suppose your plan is to toss the coin thirty times and that there are two hypotheses you want to consider. The first says that the probability of heads is $\frac{1}{4}$ on each toss while the second says that this probability is $\frac{3}{4}$. The probability that each hypothesis assigns to each possible outcome of your

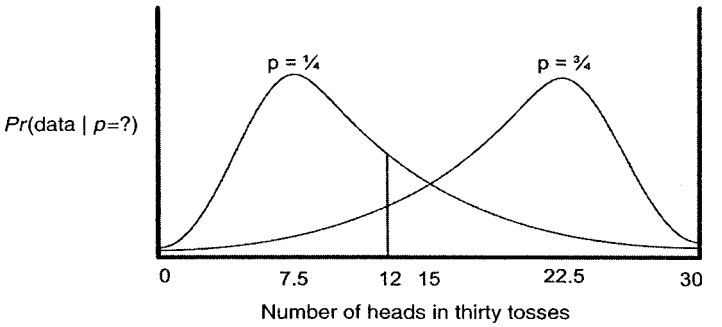

Figure 1.9 If $p = \frac{1}{4}$ is the null hypothesis and $p = \frac{3}{4}$ is the alternative to the null, and $\alpha = 0.05$ is chosen, the Neyman–Pearson theory says that the null hypothesis should be rejected if and only if twelve or more heads occur in thirty tosses of the coin.

experiment is depicted in Figure 1.9. Suppose you decide to regard the hypothesis that $p = \frac{1}{4}$ as your null hypothesis and you choose 0.05 as your value for α. You thereby require that the chance of rejecting this hypothesis, if it is true, must be less than or equal to 0.05. You now must use this stipulation to identify a "critical region." That is, you need to say what possible outcomes will suffice for rejecting the null hypothesis, given that you want to make sure that the probability of mistakenly rejecting the null hypothesis is no greater than 1 in 20. Many choices satisfy this requirement. For example, if you reject the null hypothesis precisely when there are *zero* heads in thirty tosses, the probability of rejecting the hypothesis that $p = \frac{1}{4}$ when the hypothesis is true is only $(0.75)^{30}$, which is tiny. The same can be said of the policy of rejecting the null hypothesis precisely when *all* thirty tosses land heads. With this policy, the chance of rejecting the null when it is true is only $(0.25)^{30}$, again a tiny number. Notice that neither of these judgments depends in any way on what the *alternative* to the null hypothesis happens to be. The fundamental difference between Neyman–Pearson testing and Fisher's test of significance is that the former is contrastive (pitting the null hypothesis *against a specified alternative*), while the latter is not. We now need to see what role the alternative to the null hypothesis plays in determining what the critical region will be. The critical region is determined by the joint fact that we want the chance of rejecting the hypothesis that $p = \frac{1}{4}$ if it is true to be no greater than 0.05 *and* we also want the chance of accepting this hypothesis if it is false (in which case $p = \frac{3}{4}$ is true) to be as small as possible. These

two requirements result in a unique policy. We should reject the hypothesis that $p = \frac{1}{4}$ precisely when there are twelve or more heads in the thirty tosses. This cutoff is depicted in Figure 1.9 (the example is from Royall 1997: 16–17).

Notice that this cutoff differs from the one drawn by the law of likelihood, which says that a data set with fourteen or fewer heads favors the first hypothesis while a data set with sixteen or more heads has the opposite evidential significance. If there are exactly fifteen heads in thirty tosses, the two hypotheses have the same likelihood. As noted before, the law of likelihood answers Royall's first question (what does the evidence say?) while the Neyman–Pearson theory provides a policy for acceptance and rejection. However, the two come into contact (and are incompatible) if it is a mistake to reject a hypothesis because one has obtained a set of observations that, in fact, are evidence *for* the hypothesis, not evidence *against* it. This is precisely what happens if you observe twelve, thirteen, or fourteen heads in thirty tosses. If you obtain any of these outcomes in your experiment, the Neyman–Pearson theory says to *reject* $p = \frac{1}{4}$, while the law of likelihood interprets each of these outcomes as evidence *in favor of* $p = \frac{1}{4}$ (given that the alternative hypothesis is $p = \frac{3}{4}$). If the law of likelihood is right, the Neyman–Pearson theory is wrong.

What procedure would the Neyman–Pearson theory recommend if you were to decide that $p = \frac{3}{4}$ is your null hypothesis? You then would draw a different cutoff, but it, too, would fail to coincide with the boundary drawn by the law of likelihood. With the hypothesis that $p = \frac{3}{4}$ as your null, you will reject this hypothesis precisely when eighteen or fewer tosses land heads. This means that if you observe between twelve and eighteen heads, your decision about which of the two hypotheses you'll reject depends on which is the null and which is the alternative. If the hypothesis that $p = \frac{1}{4}$ is your null hypothesis, you reject it when any of these outcomes occurs; but if $p = \frac{1}{4}$ is the alternative to the null, you do not. Life is harder on a hypothesis if it is treated as the null. Notice that the law of likelihood does not depend on how you label the various hypotheses you wish to evaluate, and there is no need to choose a value for α, either. This is a good thing, since both choices are arbitrary.

As noted, Neyman–Pearson theory first fixes a value for α and then seeks to minimize the value of β. This is why the cutoff it draws differs from the one dictated by the law of likelihood. The history of statistics might have been different. If the two types of error had been treated as equally serious, the goal would have been to minimize the sum $(\alpha + \beta)$ of the two error probabilities. This would have provided no guidance in the choice between Madison and Prairie du Chien, but it would have resulted in a crossover point of fifteen heads in thirty tosses (Royall 1997: 17), thus bringing the Neyman–Pearson philosophy into accord with the law of likelihood. In fact, there are many policies that correspond to different ways of handling the disutilities that attach to Type-1 and Type-2 errors. Even if avoiding Type-1 error is more important than avoiding Type-2, why should this mean that we need to stipulate a value for α? For example, setting $\alpha = 0.05$ means that it doesn't matter to you whether the chance of Type-1 error is 0.04 or 0.004. If making α small matters more than making β small, why not require that the sum $(10\alpha + \beta)$ be minimized? This is why the behaviorist justification of the Neyman–Pearson philosophy does not work on its own terms. Even if "acceptance" and "rejection" are taken to be behaviors that need have no connection to an assessment of evidence, the desire to reduce the frequencies of errors in one's lifetime (or in the lifetime of the enterprise of science) does not automatically entail the policy of first choosing a value for α and then minimizing β.

In discussing the principle of total evidence (§1.3), I described a few examples in which logically strengthening or logically weakening one's description of the data affects which of two hypotheses has the higher likelihood. This principle is also relevant to thinking about how the Neyman–Pearson theory bears on the question of how evidence should be assessed. We have already seen, in connection with the coin-tossing example depicted in Figure 1.8, that observing twelve heads in thirty tosses leads the Neyman–Pearson theory to reject the null hypothesis that $p = \frac{1}{4}$ and to accept the hypothesis that $p = \frac{3}{4}$ (or to not reject it) even though the former has the higher likelihood. But now let us logically weaken the description of the observations. Instead of saying "we observed *exactly twelve*," let us say "we observed *twelve or more* green balls." The law of likelihood judges that this logically weakened description of the data has a different evidential significance. Since α and β are both small, this weakened description of the data favors $p = \frac{3}{4}$ over $p = \frac{1}{4}$, and the likelihood ratio is $(1 - \beta)/\alpha$, a quantity substantially greater than unity. It *is* more probable that you'll get *twelve or more* heads in thirty tosses if $p = \frac{3}{4}$ than if $p = \frac{1}{4}$. Look at the areas under the two curves in Figure 1.8. The Neyman–Pearson theory and the law of likelihood are in accord with respect to how evidence should be interpreted *when information in the data set is thrown away*. However, this reconciliation has a price: We

have violated the principle of total evidence. From the point of view of likelihoodism and Bayesianism as well, this is a serious defect in the Neyman–Pearson theory.

In addition to the difficulties already noted, which strike both likelihoodists and Bayesians as fatal, there is a further fact about the Neyman–Pearson theory that especially irks Bayesians. How can "acceptance" and "rejection" be based just on the evidence at hand? True, if your test procedure is very reliable, a positive test result provides evidence that strongly favors the hypothesis that S has tuberculosis over the hypothesis that he does not. However, this is consistent with its being very improbable, given the positive test result, that S has tuberculosis. The Neyman–Pearson policy sometimes recommends accepting a hypothesis in the light of evidence that renders the hypothesis very improbable. This is what can happen when acceptance and rejection are controlled by likelihoods and priors are ignored. This criticism of the Neyman–Pearson theory does not require that prior probabilities *always* make sense. All that is needed is that they *sometimes* do, and this is something that non-Bayesians should concede.

In order to bring out one last feature of the Neyman–Pearson approach, let us consider a fourth tuberculosis test kit; it is made in Mazomanie:

The Mazomanie test kit: $Pr(-\text{ test result} \mid S \text{ has tuberculosis}) = 0.902$

$Pr(+\text{ test result} \mid S \text{ does not have tuberculosis}) = 0.001.$

If you decide that "S has tuberculosis" is the null hypothesis and set $\alpha = 0.05$, you will decline to use this test kit. But suppose you did so anyway, perhaps by mistake, and you obtained a positive test result. How should you interpret this evidence? A likelihoodist will say that you have just obtained strong evidence favoring the hypothesis that S has tuberculosis since the relevant likelihood ratio is large:

$$\frac{Pr_{\text{Mazomanie}}(+\text{ test result} \mid S \text{ has tuberculosis})}{Pr_{\text{Mazomanie}}(+\text{ test result} \mid S \text{ does not have tuberculosis})} = \frac{0.098}{0.001} = 98.$$

In fact, this evidence is precisely as strong as the evidence that attaches to a positive result produced by the Madison test kit. Even though the two test kits have different values for α and β, a positive test result produced by using the Madison test kit also produces a likelihood ratio of $0.98/0.01 = 98$. Yet, the Neyman–Pearson methodology instructs you not to use the Mazomanie test kit and embraces the one from Madison. How can it do so, if the two are *evidentially equivalent* when a positive test result is produced? The answer is that the Neyman–Pearson theory addresses the question of how one should choose a *general policy*. If you, the doctor, have to choose between using the Madison test kit on all your patients and using the Mazomanie test kit on all of them, the plausible choice is to opt for the one from Madison. Notice that the previous sentence answers a question that falls under Royall's question (3): What should you do? That is, which test kit should you use in your medical practice? It is not an answer to question (1): What is the evidential meaning of S's positive test result? Nor does it address question (2): Should you believe that S has tuberculosis? Hacking (1965) makes this point by distinguishing the task of *before-trial betting* and *after-trial evaluation*. The first involves designing an experiment, the second the interpretation of the results you obtained on the experiment you actually ran. Likelihoodists and Bayesians hold that both tasks are important but maintain that they are distinct. The Neyman–Pearson philosophy does not distinguish these tasks; once a general procedure has been chosen, there is no additional question as to how the result obtained by applying the procedure on a single occasion should be interpreted. This difference between the philosophies becomes vivid when a less than optimal test procedure is used and one wishes to interpret the result. This was my point in introducing the Mazomanie test kit. If you use this procedure and obtain a positive result, Neyman–Pearson frequentists will say that you shouldn't have used that test kit and will refuse to interpret the outcome; Bayesians and likelihoodists will say that using that test kit rather than the one from Madison turned out not to matter and will be happy to interpret the test outcome. Philosophers will recognize that this difference between the two statistical frameworks parallels the distinction in ethics between rule and act utilitarianism.

I have described the rudiments of the Neyman–Pearson theory in the context of the simple example of coin tossing, and this has allowed me also to describe some standard criticisms of that approach. However, frequentists may want to object that it is silly to test the hypothesis that $p = \frac{1}{4}$ against the hypothesis that $p = \frac{3}{4}$. Instead, why not just *estimate* the value of p (and draw a confidence interval around that estimate)? For example, if there are twelve heads in the thirty tosses, you can simply say that the maximum likelihood (ML) estimate of p is 0.4; as already noted, this doesn't mean that p *probably* has a value of 0.4 or even that the true value is probably *close* to 0.4. However, in saying that this is the ML estimate, you can sweep aside the problem of deciding which of $p = \frac{1}{4}$ and

$p = \frac{3}{4}$ is the null hypothesis and what your value for α ought to be. ML estimation may sound like likelihoodism or even Bayesianism, but frequentists have their own special rationale for this procedure. Frequentists do not accept the law of likelihood. Rather, they see the method of ML estimation as justified, when it is, because it has certain virtues as a *general policy*; for frequentists, there is no additional question about the evaluation of an individual ML *estimate*. It is *estimators*, not *estimates*, that is their focus. A central concept in the frequentist theory of estimation is that an estimator (i.e., a procedure for making estimates) must be *admissible*. A method of estimation is *in*admissible if there is another estimator that has a smaller expected error for all possible values that the parameter being estimated might take. Whereas inadmissibility arguably suffices to not use an estimator, admissibility is not sufficient for a method to be used. The reason is that there can be multiple admissible estimators that give contradictory advice. In any event, it turns out that ML estimation is an admissible procedure when one or two parameters are being estimated but not when the estimation problem involves three or more. With more than two parameters, there is another procedure, involving shrinkage in accordance with a formula derived by James and Stein (1961) that has a lower expected error no matter what the true values are of the parameters being estimated (Efron and Morris 1977). This is not the place to pursue questions about estimation any further; suffice it to say that frequentists can decline to use the Neyman–Pearson theory to test the hypothesis that $p = \frac{1}{4}$ against the hypothesis that $p = \frac{3}{4}$ and insist that maximum likelihood estimation of the value of p is the way to go.[33]

Although estimation may make more sense than Neyman–Pearson hypothesis testing when the two hypotheses are statistically *simple*, this option is not available to the frequentist when both the hypotheses being tested are *composite*. In this case, the standard Neyman–Pearson approach is the *likelihood ratio test*. Don't let this terminology mislead you; this test is a frequentist construct even though the likelihood ratio also appears in the law of likelihood, which is the central concept of likelihoodism. Here's an example that illustrates what the likelihood ratio test involves. You conduct the following experiment in your kitchen: You heat a pressure cooker to a given temperature and then observe how much pressure there is in the container. You don't observe the temperature and

[33] It is worth emphasizing that this change in strategy does nothing to vindicate the Neyman–Pearson theory as it applies to simple hypotheses. The objections have not been *met*; rather an altogether different frequentist approach has been suggested.

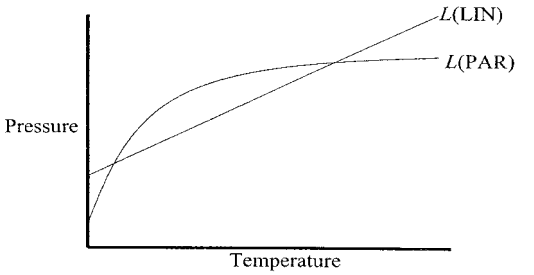

Figure 1.10 Each of the observations can be represented by a data point. L(LIN) is the straight line that fits the data best; L(PAR) is the parabola that fits best. The likelihood ratio test compares the models LIN and PAR by computing the likelihood ratio of L(LIN) and L(PAR).

pressure directly; rather, you observe the readings that a thermometer and a pressure meter provide. You know that these devices are reliable, but not perfectly reliable. You do this experiment multiple times, representing each observation by a point in the coordinate system depicted in Figure 1.10.

Suppose there are two models you want to test that both attempt to describe how temperature and pressure are related in this system. With the variables X and Y representing temperature and pressure, respectively, the two models are:

(LIN) $\quad y = a + bx + e$

(PAR) $\quad y = a + bx + cx^2 + e.$

LIN says that temperature and pressure are related linearly; PAR says that they are related parabolically. In these models, x and y are variables, while a, b, c, and e are parameters. Each model is an infinite disjunction; LIN is a disjunction of all straight lines in the X-Y plane; and PAR is a disjunction of all the parabolas. In other words, these models have existential quantifiers attached to their adjustable parameters; LIN, for example, says that *there exist* values for a, b, and e such that $y = a + bx + e$. The "e" in each model represents the fact that your observations are subject to error. Even if the true relationship between temperature and pressure is linear, you can't assume that the data you gather will fall exactly on a straight line. LIN postulates an error distribution around each of the straight lines it includes. Although a straight line is sometimes said to provide the "predicted" y-value for a given x-value, this is a bit misleading. What each straight line in LIN represents is the *average* (the

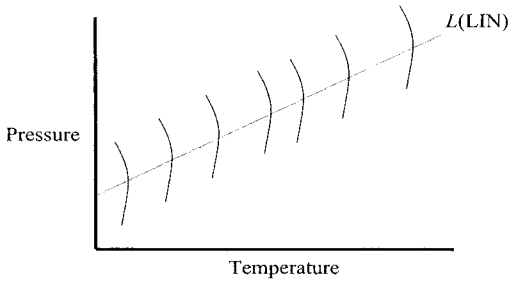

Figure 1.11 L(LIN) is the straight line that is closest to the data; the LIN model postulates an error distribution around this line. The observed pressure value for a given temperature need not coincide exactly with the average ("predicted") pressure value.

expected value; see §1.2) of the observed pressure-values that should be associated with a given value of temperature. This error distribution is depicted in Figure 1.11.

Here's how the likelihood ratio test applies to the comparison of LIN and PAR. First, you find the straight line that maximizes the probability of the data. This will be the straight line that is "closest" to the data; that is, the line that "fits" them best. Call this maximum likelihood straight line L(LIN). Then you do the same thing with PAR. There are many parabolas, some close to the data, others far away. You need to find the member of PAR that maximizes the probability of the data; this is L(PAR). These two "best cases" of LIN and PAR are depicted in Figure 1.10. In discussing how the Neyman–Pearson theory evaluates the two simple statistical hypotheses ($p = \frac{1}{4}$ and $p = \frac{3}{4}$) about coin tossing shown in Figure 1.9, I was able to discuss what each predicts about the data. But LIN and PAR are composite. Neither says how probable the data are that you generated in your kitchen (i.e., how probable the y values you observed are, given the x values you used). The Neyman–Pearson theory solves this problem by shifting from LIN to L(LIN) and from PAR to L(PAR). We test the two models by comparing the maximum likelihood members of each. It's as if LIN and PAR are two armies that compete by each sending forth its fittest champion. The armies stand idle and are evaluated by seeing which champion wins the *mano a mano*. The likelihood ratio test of LIN against PAR focuses on the likelihood ratio

$$\frac{Pr[\text{data} \mid L(\text{LIN})]}{Pr[\text{data} \mid L(\text{PAR})]}.$$

The question is whether this ratio is smaller than some arbitrarily chosen level of significance; if it is, you should reject LIN.

One interesting feature of the likelihood ratio test is that it avoids an arbitrariness that afflicts the Neyman–Pearson test of two simple hypotheses. In the coin-tossing example of testing $p = \frac{1}{4}$ against $p = \frac{3}{4}$, you need to decide which of these hypotheses is the null. As also was true in the example of the tuberculosis test, there is nothing inherent in these simple hypotheses that settles which is "really" the null. Considerations concerning which type of error you are more concerned to avoid are typically brought to bear, but this is a fact about *us*, not about the hypotheses themselves. Testing LIN against PAR is a different matter. Each of these models contains adjustable parameters, but it is LIN that says that $c = 0$ while PAR leaves open what value that parameter has. It is in this objective sense that LIN can be said to be the null hypothesis in this two-way competition. Frequentists sometimes describe the choice of null by talking about which of the hypotheses we want to nullify (i.e., reject), but there is no need for us and our desires to intrude into the story.

When I discussed the two simple hypotheses $p = \frac{1}{4}$ and $p = \frac{3}{4}$ about the coin and the problem of deciding which of them is the null hypothesis and what level of α to use, I considered the possibility that frequentists might decline to apply the Neyman–Pearson theory to this problem and instead would insist that the problem to address is how best to estimate the value of the parameter p, where $p = Pr$(the coin lands heads | the coin is tossed). Estimation is an issue that arises *after* you have settled on a given model of the experiment. You have already decided that each toss of the coin has the same probability of landing heads as every other and you have decided that the tosses are independent of each other. Given this framework, you can estimate p. Testing the composite hypotheses LIN and PAR is different. The problem of choosing a level of significance can't be set aside and an estimation problem considered in its stead. The reason is that the competition between LIN and PAR is a competition *between* models, while estimation is a task that is carried out *within* the confines of a single model. True, if you assume that LIN is true, you can estimate the values of the parameters in it; the same goes for PAR. But that hardly suffices to test LIN against PAR. In fact, you know in advance that L(LIN) can't have a higher likelihood than L(PAR). This is because LIN is *nested* in PAR. LIN is a special case of PAR; the equation for LIN can be obtained from the equation for PAR by setting the parameter $c = 0$. Given that the ratio on which the likelihood ratio test focuses can't have a value that is greater than unity, the frequentist's question is whether the

ratio is *significantly* less than unity; you have to look at the data to see whether this is so.

It is interesting to reflect on what the frequentist advice to "accept" or "reject" means in the context of these two composite models. LIN is *nested* in PAR, meaning that LIN logically entails PAR. If so, what would it mean to accept LIN and reject PAR? You can't regard LIN as true and PAR as false if the former entails the latter. It also makes no sense to decline to reject LIN and to reject PAR; if PAR is false, so is LIN. It might be replied that the frequentist can eliminate this problem by stipulating that the models worth talking about are *not* nested; this can be achieved by requiring that all the parameters in the two models have nonzero values. Now the models are incompatible. The problem with this reply is that the mathematics that underlies the likelihood ratio test requires that models be nested (Burnham and Anderson 2002).

Bayesians have an additional criticism of the Neyman–Pearson treatment of composite hypotheses, one that does not apply when only simple hypotheses are considered. The Neyman–Pearson theory compares LIN and PAR by comparing the members of each that have maximum likelihood, namely L(LIN) and L(PAR). But the likelihoods of LIN and PAR, the Bayesian will observe, are not these *maxima* but rather are their *average* likelihoods. Since LIN is a disjunction of straight lines (L_1, L_2, ...), it has a likelihood of the following form:

$$Pr(\text{data} \mid \text{LIN}) = \sum_i Pr(\text{data} \mid L_i) Pr(L_i \mid \text{LIN}).^{34}$$

Frequentists don't want to discuss these average likelihoods because it often is impossible to empirically justify an assignment of values to the weighting terms that have the form $Pr(L_i \mid \text{LIN})$. If the temperature and pressure in your pressure cooker are linearly related, what is the probability of the different specific straight-line relations that might obtain (and please answer this question without looking at the data you drew from your pressure cooker)? This is one motive that frequentists have for shifting from the *average* likelihood of the infinitely many straight lines that belong to LIN to the *unique* likelihood value that attaches to just one of them, namely to L(LIN). This *is* a motive for shifting, but not a justification for the likelihood ratio test. The justification offered is that if you follow the Neyman–Pearson procedure again and again, the expected value of your

[34] This should be an integral, not a discrete summation, but I prefer to use the latter to make this material accessible to a wider readership. Aficionados know how to correct this crudity.

Type-1 errors will be no more than α, and the expected value of your Type-2 errors will be β. It's the *general policy* that has this property, but the question may be asked of why this property of the general policy shows, in the concrete situation of evaluating LIN and PAR with the data you have from your kitchen experiment, that you should evaluate the two models by examining the maximum likelihood special cases of each. Frequentists regard this question as irrelevant, while Bayesians regard it as central.

Even if there is nothing arbitrary about saying that LIN is the null hypothesis when LIN is compared with PAR in a likelihood ratio test, there is another detail of this procedure that introduces a kind of arbitrariness that did not appear in the example of testing the two simple hypotheses $p = \frac{1}{4}$ and $p = \frac{3}{4}$ about coin tossing. To see what this new arbitrary element is, we need to consider a hierarchy of nested models, not just two. LIN and PAR are both polynomials; each has the form:

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_{n-1} x^{n-1} + b_n x^n.$$

LIN is a first-degree polynomial and PAR is second-degree. Let's consider a hierarchy of five polynomials by adding to our list three more – a third, fourth, and fifth degree. For simplicity, let's call these five models A, B, C, D, and E. We need to fit each of these five models to the data from our stovetop experiment and then figure out the likelihood ratios for adjacent pairs of fitted models. Suppose we obtain the following left-to-right likelihood ratios:

$$L(\text{A}) \leftarrow (0.1) \rightarrow L(\text{B}) \leftarrow (0.3) \rightarrow L(\text{C}) \leftarrow (0.05) \rightarrow L(\text{D}) \leftarrow (0.5) \rightarrow L(\text{E}).$$

There are two ways to apply the likelihood ratio test to this hierarchy: *step-up* and *step-down*. In each case, a level of significance needs to be chosen; suppose you select $\alpha = 0.15$. In step-up testing, you begin with the simplest model A and ask whether the likelihood ratio of L(A) to L(B) is less than 0.15. If it is, you reject A and then compare B and C and ask the same question. You continue to step-up until you can't anymore. The result of step-up testing on this sequence of models is to reject A in favor of B but then to fail to reject B in favor of C. The process terminates with B. In step-down testing, you begin with the most complex model, E, and compare it with the model that is one step down, namely D. The question is whether the likelihood ratio of L(D) to L(E) is less than 0.15. If it is, you stay with E. If it is not, you move from E to D. Given the numbers shown above, this step-down process terminates with D. The choice between step-up and step-down testing is arbitrary

and yet it can influence which models you accept and reject (Burnham and Anderson 1998).

1.6 A TEST CASE: STOPPING RULES

There is a classic puzzle that illustrates the clash between Bayesianism and likelihoodism on the one hand and significance tests and the Neyman–Pearson theory on the other. It concerns the "stopping rule" used when observations are gathered. This rule determines when the inquiry is over. In the example about coin tossing that I used to explain significance tests in §1.4, the stopping rule was to stop after the coin is tossed twenty times; it then turned out that six heads had occurred. The same outcome can occur if a different stopping rule is used. For example, you might decide to toss the coin until you obtain six heads, and it then turns out that the sixth head occurs on the twentieth toss. Here's the question: if you obtain the sixth head on your twentieth toss, should your interpretation of this result depend on which of the two stopping rules you used? Likelihoodism and Bayesianism say *no*, whereas the two versions of frequentism examined so far say *yes*.[35]

Let's begin with the likelihood analysis, which the Bayesian accepts; the issue about prior probabilities plays no role here. Although this problem is sometimes described as if it is supposed to be obvious that Bayesianism entails that the choice of stopping rule is irrelevant, the reason for this is worth tracing carefully. For the sake of a simpler example, let's shift for a moment to comparing a fixed-length experiment that involves tossing a coin three times with a flexible-length experiment in which you toss the coin until it lands heads. The possible outcomes of each of these experiments are depicted in Figure 1.12. If the coin is fair ($p = 0.5$), each specific sequence that can occur in the fixed length experiment has a probability of $\frac{1}{8}$; in the flexible length experiment, the probabilities of the different outcomes (reading from left to right) are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and so on. Suppose you obtain tails on the first two tosses and heads on the third but don't know what the value of p is. The probability of obtaining the sequence TTH is the same regardless of which experiment was performed:

$$Pr(\text{TTH} \mid \text{there will be 3 tosses}) = Pr(\text{TTH} \mid \text{there will be exactly one H})$$
$$= p(1-p)^2.$$

[35] This example is from Howson and Urbach (1993: 210–12); it is similar to an example given by Lindley and Phillips (1976).

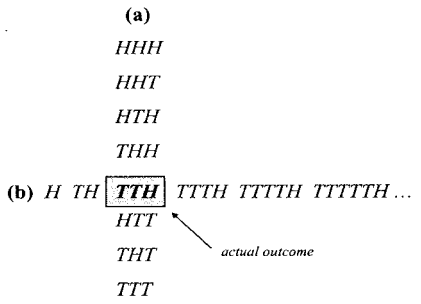

Figure 1.12 If a coin lands tails on the first two tosses and heads on the third, this outcome might be the result of two different experiments: (a) toss three times; (b)toss until heads occurs once (from Goodman 1999: 1000). The possible outcomes of both experiments are shown.

This means that if you are testing the hypothesis that $p = 0.5$ against the hypothesis that $p = 0.9$, the following equality obtains

$$\frac{Pr(\text{TTH} \mid p = 0.5 \;\&\; \text{there will be 3 tosses})}{Pr(\text{TTH} \mid p = 0.9 \;\&\; \text{there will be 3 tosses})} = \frac{Pr(\text{TTH} \mid p = 0.5 \;\&\; \text{there will be one heads})}{Pr(\text{TTH} \mid p = 0.9 \;\&\; \text{there will be one heads})}.$$

This equality indicates why Bayesians say that the choice of stopping rule is not relevant to the interpretation of the observations; the weight of evidence (as measured by the likelihood ratio) is the same, regardless of which experiment you performed. Returning to our initial example, I hope it is clear why it doesn't matter to the likelihoodist whether you obtained six heads in a fixed length experiment of twenty tosses or if it took you twenty tosses to obtain six heads in a flexible length experiment – the meaning of the evidence is the same.

Why does the difference between the two experimental designs matter to the significance tester? The answer begins with the fact that significance tests require you to consider the probability under the null hypothesis of a logically weaker description of the data – that you obtained the test result *or ones that are at least as improbable*. If the null hypothesis says that $p = 0.5$, the probabilities you need to think about to perform a test of significance for the fixed and the flexible length experiments are, respectively,

(Fixed) $Pr(0-6 \text{ or } 14-20 \text{ heads} \mid p = 0.5 \;\&\; \text{there will be 20 tosses})$

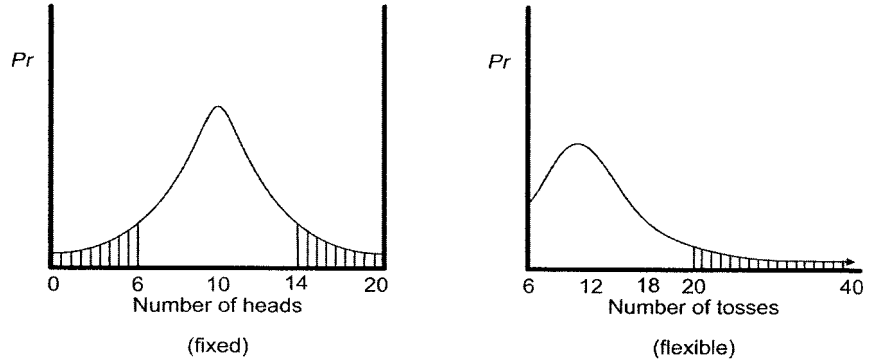

Figure 1.13 Suppose that a fixed-length experiment in which a coin is tossed twenty times and a flexible-length experiment in which a coin is tossed until six heads occur both result in six heads in twenty tosses. In each case, a significance test of the null hypothesis that the coin is fair focuses on the probability of obtaining that result or ones that are at least as improbable.

and

(Flexible) $Pr(20 \text{ or more tosses} \mid p = 0.5 \ \& \text{ there will be 6 heads})$.

The relevant regions of the two outcome spaces that the two significance tests consider are shown in Figure 1.13. It turns out that (Fixed) has a value of 0.115 and (Flexible) has a value of 0.0319. If you set your level of significance at $\alpha = 0.05$, you should not reject the null hypothesis if you performed the fixed experiment, but you should reject the null if you performed the flexible. Which experiment you performed to obtain your six heads in twenty tosses makes all the difference.[36]

It is not a unique feature of significance tests that the probability the null hypothesis confers on a logically weakened description of the data depends on which experiment was performed. Consider the simpler example depicted in Figure 1.12. You tossed the coin three times and obtained the exact sequence TTH. As already noted, this description of the data has a probability of $\frac{1}{8}$ under the null hypothesis that $p = 0.5$ regardless of which experiment was performed. However, with a logically

[36] The same result can arise in Neyman–Pearson hypothesis testing, for example, if the null hypothesis is tested against the composite alternative that $p \neq 0.5$ (Howson and Urbach 1993: 211).

weaker description of the data (in which you describe the mix of heads and tails but omit to mention their order), this agreement dissolves:

$$Pr(2\text{T and } 1\text{H} \mid \text{Null \& there will be 3 tosses}) = \frac{3}{8}.$$

$$Pr(2\text{T and } 1\text{H} \mid \text{Null \& there will be just 1 H}) = \frac{1}{8}.$$

As shown in Figure 1.12, there is just one way to get two tails and one heads in the flexible experiment, but there are three in the fixed-length experiment. The point is that likelihoodists don't care about the values of these single conditional probabilities but only about the values of various *ratios*, whereas significance testers think that what matters is the value of a single conditional probability – (Fixed) or (Flexible) as the case may be.

Given the importance that significance testers assign to the choice of stopping rule, what should they say about experiments in which it is unclear which stopping rule was actually used? Howson and Urbach (1993: 212) describe the following example. Suppose two scientists collaborate to perform a coin-tossing experiment; they obtain six heads in twenty tosses (with the sixth head occurring on the last toss) and then sit down to write an article in which they report their results, thinking that nothing is amiss. It then emerges that they had different plans in mind; the first scientist thought the plan was to toss twenty times; the second thought the plan was to toss until six heads occur. Of course, they should have talked things through beforehand, but what are they now to do? According to the logic of significance tests, they need to figure out what they would have done if other results had emerged. If they had obtained the sixth head on the nineteenth toss, would they have continued the experiment? If they had obtained only five heads by the twentieth toss, would they have persevered? Answering these questions requires information about the power relations between the two experimenters. Perhaps you are inclined to say that it doesn't matter what they would have done if the results had been otherwise; what matters is the results they in fact obtained and this result can be interpreted without psychoanalyzing the two scientists. If so, you are thinking like a likelihoodist.

Defenders of significance tests often suggest that Bayesians are hopelessly uncritical of how experiments are designed but that frequentists, in this respect, have their heads screwed on right. Suppose I decide to continue tossing a coin until I obtain results that go against the null hypothesis. If so, I apparently know in advance what conclusion I'll draw.

But if I cannot fail to reject the null hypothesis, regardless of whether that hypothesis is true, how can the experiment be said to test that hypothesis? And if the experiment doesn't test the null, why bother to run it in the first place? Frequentism explains why it is pointless to do this experiment, but frequentists often claim that Bayesians have a blind spot here; Bayesianism, they say, holds that there is nothing wrong with running this type of "try-and-try-again" experiment (Mayo 1996). What is even more galling to frequentists is that Bayesians have the temerity to proclaim this a *virtue* of their position, rather than acknowledging it to be the embarrassment to Bayesianism that it truly is.

This criticism of Bayesianism is sometimes stated as a very general claim: That Bayesianism *never* accords any epistemic import to the design of experiments and can offer no rationale for declining to perform experiments whose outcomes are known in advance. This criticism is vastly overstated, as a simple example from Eddington (1939) illustrates. You throw a net in a lake and wait until fifty fish have been caught. You pull the net out and see that all fifty fish are more than 10 inches long. How does this observation bear on the following two hypotheses? H_1 says that all the fish in the lake are more than 10 inches long; H_2 says that 50 percent of the fish are more than 10 inches long. Your first impulse is to think that the observations favor H_1 over H_2, but then you realize that this interpretation depends on what the net was like. If the net has 1 inch holes, the interpretation makes sense, but if the holes are 10 inches across, the observation fails to discriminate between the two hypotheses. The general point is that the bearing of observations on hypotheses often depends on the methods used to obtain the observations. When the outcome of an experiment is knowable beforehand and does not depend on which hypothesis is true, there is no point in performing this experiment; the law of likelihood provides a perfectly straightforward explanation of why this is so.[37]

Setting this hyperbolic criticism of Bayesianism to one side, let us look in more detail at fixed- and flexible-length experiments of the kind described in Figures 1.12 and 1.13. Let's begin by getting the facts straight in connection with frequentism. Consider an experiment that ends precisely when a significance test takes the data to indicate that the null hypothesis should be rejected. It is a certainty that this experiment will end if one uses the "nominal" value for the level of significance (Anscombe 1954). Using the nominal value means that at each stage one pretends that

[37] I discuss Eddington's example of an *observation selection effect* in connection with the fine-tuning version of the design argument in Sober (2004b).

the data were the result of an experiment designed to have that number of observations. Since this experiment's outcome is known in advance and does not depend on whether the null hypothesis is true, frequentists think there is an excellent reason not to run it. However, they are not opposed to "sequential trials." Armitage (1975) has described a protocol for such experiments in which one uses the "overall," rather than the "nominal," value for the level of significance. This new concept has the consequence that it is no longer a certainty that the experiment will end, and so it is no longer crazy, from a frequentist point of view, to run it. Armitage also describes how sequential trials can be structured so that accepting the null hypothesis as well as rejecting it is a possible outcome.

To understand what Bayesianism and likelihoodism say about this problem, we must be careful not to saddle these frameworks with ideas that are alien to them. Neither uses significance tests, and their experiments don't end with the "acceptance" or "rejection" of the null hypothesis. Both interpret experimental results by using the law of likelihood, so we need to be explicit about the alternative to the null hypothesis that is in contention. To this end, let's suppose that the null hypothesis (H_0) says that $p = 0.5$, that the alternative hypothesis (H_1) says that $p = 0.9$, and that the experiment you undertake will stop precisely when the frequency of heads engenders a likelihood ratio of H_0 to H_1 that is less than or equal to $1/k$ (where $k \geq 1$). If H_0 is true, is this experiment bound to end, thus resulting in misleading evidence that favors H_1? Robbins (1970) has shown that the probability of this experiment's ending when H_0 is true is less than or equal to $1/k$. If you define "strong evidence against the null" to mean a ratio that is less than $\frac{1}{8}$, then the probability of this misleading result is less than $\frac{1}{8}$. Commenting on this point, Royall (1997: 7) says that "if an unscrupulous researcher sets out deliberately to find evidence supporting his favorite but erroneous hypothesis [...] over his rival's [...] which happens to be correct, by a factor of at least k, then the chances are good that he will be eternally frustrated." Notice that this point has nothing to do with the prior or posterior probabilities of the hypotheses; it falls strictly within the likelihood framework.[38]

[38] Kadane et al. (1996) obtain similar results but within a fuller Bayesian framework and using the strong assumption of countable additivity. Suppose you decide to end the experiment precisely when the posterior probability assigned to H_1 exceeds some value v. If your prior for H_1 is r, the probability that the experiment will end, if H_0 is true, is no more than $r(1-v)/(1-r)v$. So if H_0 and H_1 each have priors of $\frac{1}{2}$, and you don't stop the experiment until H_1 has a posterior probability of at least 0.9, the probability of the experiment's ending is no more than 0.11. Notice

Thus, the try-and-try-again design in which you end the experiment only when you've obtained strong evidence against H_0 is *not* bound to end, if the criterion for its ending is formulated in terms of the likelihood ratio. If there is something wrong with this experimental design, it is not that you know in advance what will happen. One defect, noted by Teddy Seidenfeld, is that if the null hypothesis is true, this experiment has a serious chance of going on forever; if experiments cost money to run, Bayesians with finite funds have a good reason not to use this experimental design (Backe 1999: S360).[39] Fortunately, there are other designs that are far more sensible; for example, you could continue drawing evidence until strong evidence favoring H_0 over H_1, or strong evidence favoring H_1 to H_0, is obtained. The probability that this even-handed experiment will end, sooner or later, is unity (Wald 1947: 37–40; Backe 1999: S359); of course, it is not a foregone conclusion which result you'll obtain.

Where do these points leave the optional stopping problem? Significance testers abhor the try-and-try-again experimental arrangement when carried out with "nominal" p-values. However, with "overall" p-values, sequential experiments are not beyond the frequentist pale. And if you organize your test along Bayesian or likelihoodist lines, it is not true that try-and-try-again must result in the experiment's ending (where ending means attaining a likelihood ratio that represents strong evidence against the null). This shows that if the experiment *does* end, you really do have evidence (as defined by the likelihood ratio). Bayesians think that both the design of experiments and the interpretation of the results obtained are important topics; this is Hacking's (1965) distinction between before-trial betting and after-trial evaluation (§1.5). It is frequentists who often do not see the second as a problem separate from the first.

1.7 FREQUENTISM III: MODEL-SELECTION THEORY

The keys and the lamppost

When I raised the objection that Bayesianism often has no objective basis for assigning values to prior probabilities or to the likelihoods of catchall hypotheses, I did not describe a *different* theory for assigning such values

the relationship to the likelihood ratio in this example; given these values for r and v, the experiment ends precisely when the likelihood ratio of H_0 to H_1 is $\frac{1}{9}$ or less.

[39] Compare Jeffrey's (1983: 154) response to the St. Petersburg paradox: The bargain you are offered must be fraudulent, since no one has an infinite amount of money.

and then argue that this different theory is *better* than Bayesianism. No, what I did was change the subject. I retreated to likelihoodism, which addresses a different question – the question of how evidence ought to be interpreted. This pattern of shifting questions is not unique to the foundations of statistics nor is it unique to philosophy. Though politics is often called "the art of the possible," science deserves to be described in this way as well. If one problem cannot be solved, there is no reason why another should not be taken up that can. The only sin is to give the false impression that a new theory solves the same problem that an old one was unable to address. Science is sometimes like the man searching under the lamppost for the keys that he misplaced. When asked why he is searching there, he replies that that is where the light is. He does not reply that that is where his keys probably are.

In the previous two sections, I explained the rudiments of significance tests and of the Neyman–Pearson theory of hypothesis testing. I described some serious (and standard) objections to each. However, as mentioned at the start of the discussion of frequentism, this statistical philosophy is not a unified theory; rather, it is a loose confederation of ideas. The criticisms I've made of significance tests and of hypothesis testing don't necessarily attach to other frequentist ideas. The part of statistics called *model-selection theory* may have its problems, but it avoids the problems we so far have identified. There is no need to decide which hypothesis to call the null, and there is no need to choose a value for α. Indeed, there is no such thing as acceptance and rejection in model-selection theory. The name of this part of statistics is misleading; the problem addressed is one of model *comparison*, not model *selection*. Before we consider some of the solutions that have been proposed to the problem of model comparison, we need to understand what the problem is. An important element in this field has been the articulation of a new question: How should we estimate how accurate a theory's predictions will be?

Model building in science: Two pervasive patterns

In many areas of scientific research, a great deal of effort goes into the construction and evaluation of "models." This term has a technical meaning in statistics and a somewhat different nonmathematical meaning in the sciences themselves. As noted in the discussion of LIN and PAR in the previous section, models in the statistical sense of that term contain adjustable parameters; the statement that X and Y are related linearly is a model, while the statement that $y = 3 + 4x$ is not. This specific straight-line

equation has been obtained from the linear model by substituting point values for adjustable parameters. When scientists use the term "model," they often have a different idea in mind. For them, a model is a simplified hypothesis; it purports to explain or predict a set of observations without trying to represent all the factors that are relevant. Models are not fully *realistic*; rather, they contain *idealizations*. Physicists work with models that assume that planets are spherically symmetrical, that particles collide with perfect elasticity, and that balls roll down inclined planes that are perfectly frictionless; evolutionary biologists consider models that assume that populations are infinitely large, that mating is perfectly random, and that a trait has a single unchanging fitness value in each of the many generations of the population in which the trait evolves. These models are known to be false, but they are not dismissed out of hand. The hope is that there may be truth in these falsehoods. If the idealizations are *harmless*, their departures from the truth won't matter much; these idealized models will yield accurate predictions even though they are false (McMullin 1985; Hausman 1992). If your goal is to predict how much time a ball will take to roll down a ramp, assuming that the ramp is *perfectly* frictionless may be fine if the ramp is *nearly* frictionless and your measurements are somewhat imprecise.[40]

There are two pervasive facts about the use of models in science that are of considerable philosophical significance. The first is that scientists often test models that they know are false. This is especially clear for many of the hypotheses that are labeled *null models*. This term is often applied to hypotheses that say that there is no difference between two quantities. The hypothesis that two fields of corn plants have the same mean height is a null hypothesis in this sense; the same is true of the hypothesis that a coin is fair (since it says that there is no difference between the chance of heads and the chance of tails). It is interesting that we often know, with as much certainty as we can ever have in science, that these so-called null hypotheses are false. Consider the coin. Do you really think that the coin is *exactly* symmetrical, that the chance of heads (p) is *exactly* equal to the chance of tails ($1 - p$) on each toss and that this precise symmetry remains in place each and every time the coin is tossed? I, personally, do not. My expectation is that there are modest asymmetries in the shape and balance

[40] There is a third use of the term "model" found in the part of logic called model theory. Here a model is a set of objects, properties, and relations that make a set of sentences true. In this usage, models are not propositions. For historical and philosophical reflections on the use of models in science, see Hesse (1966), Morgan and Morrison (1999), Da Costa and French (2003), and Frigg and Hartmann (2006).

of the coin; I am virtually certain that $p \neq \frac{1}{2}$. I also feel pretty sure that the coin changes its shape, if only slightly, during its lifetime. So why do scientists bother to test the simple hypothesis that $p = \frac{1}{2}$ against the composite hypothesis that $p \neq \frac{1}{2}$? Or consider the two fields of corn. The null hypothesis says that there is no difference in their average heights. Again, I find myself as certain as I am about almost anything that this null hypothesis is not true. The falsity of the null hypothesis, of course, is not an *a priori* matter; however, I suggest that our empirical experience of the world assures us that the two means are not *exactly* the same (to 1 million decimal places and more). Yet, scientists test the null hypothesis that the difference is zero against one or another alternative hypothesis.

Given that null hypotheses are often known to be false before any statistical test is run, it is not surprising that statisticians sometimes argue that these null hypotheses are not worth testing (see, for example, Yoccoz 1991 and Johnson 1995). I do not draw this conclusion. If the goal of scientific inference were just to find out which theories are true, dismissing such null hypotheses without testing them would make sense. But if the goal is to discover which theories will make accurate predictions, there may be a point in testing null hypotheses. Maybe hypotheses known to be false will make accurate predictions. And if *all* the hypotheses under test are known to be false (since all contain idealizations), it may still be worthwhile to determine which of them can be expected to make the most accurate predictions. If idealized (and therefore false) models are proper objects of scientific testing, we need to change our conception of what the goal of scientific reasoning is. Bayesianism is usually understood as a theory for deciding which hypotheses are probably *true*; the Neyman–Pearson theory concerns which hypotheses we should accept as *true* and reject as *false*; and likelihoodism tells us whether our evidence favors the hypothesis that H_1 is *true* over the hypothesis that H_2 is *true*. Truth enters into each of these theories of inference. This obsession needs to be overcome.

A second fact about model building in science also is pregnant with philosophical meaning. It concerns an experience that scientists often have when they use models that are very complex. When scientists consider a body of data that they suspect was produced by multiple causes that interacted in complex ways, they may be tempted to invent a complex model as an explanation. Doesn't a complex reality need a complex theory to do it justice? However, when such models are fitted to the data by finding the maximum likelihood estimates of their adjustable parameters (as we did in the example in §1.5 about the pressure cooker), those fitted models often do a terrible job of predicting new data drawn from the

same system. Here's an example that illustrates the kind of pattern I have in mind. Suppose you made n observations of $<xy>$ pairs during your experiment with the pressure cooker. It is a mathematical fact that a polynomial of degree $n - 1$ can be found that fits those n data points *perfectly*. If you made two observations, there is a straight line (a first-degree polynomial) that passes exactly through them; if you made three observations, there is a parabola (a second-degree polynomial) that does the same thing. And so on. Sadly, the mathematical assurance that a sufficiently complex polynomial will fit the *old* data perfectly is no guarantee that the fitted polynomial will do a good job predicting *new* data. In fact, scientists often find that complex models do very poorly in predicting new data when fitted to old. Simpler models often do better. Here the complexity of a model corresponds to the number of adjustable parameters it contains.

Given this common experience that model-builders have, it may seem that the only lesson is the following vague rule of thumb: Don't make your models too complicated or too simple, either. This advice is sensible, but it isn't very helpful. How complicated is too complicated? What is remarkable is that this advice can be made more precise. Work in model-selection theory has shown that, in a variety of circumstances, it is possible to *estimate* how accurately a model will predict new data when it is fitted to old. There is much that remains to be learned about the mathematical underpinnings of this area, but what is striking is that there are mathematical structures here to be investigated. The fact that models that are very complex are often not good at predicting new data when fitted to old is not a brute fact. Rather, there is a body of mathematics that *explains* why complex models are often poor predictors and allows scientists to take measures to avoid using models that are too complex.

Akaike's framework, theorem, and criterion

Model-selection theory began as a subject in statistics with Hirotugu Akaike's 1973 paper. Akaike identified a problem, and he proposed a solution to it. It is important to keep separate these two parts of what he accomplished, since the problem he singled out for study has an importance that transcends the solution to the problem that he proposed. This is because the subject he founded led to the discovery of different solutions that are appropriate in different settings. There now are multiple model-selection criteria on the market, and it is widely recognized that different criteria should be used for different model-selection tasks.

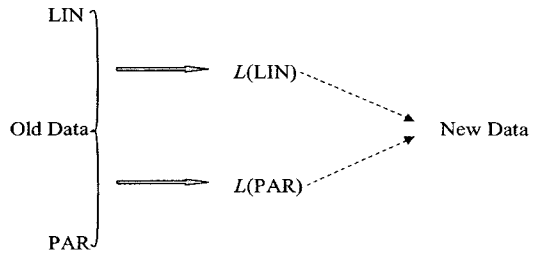

Figure 1.14 The prediction problem that Akaike considered. From the old data and a model you can deduce what the likeliest member of that model is. That likeliest model then makes probabilistic predictions about new data. Which model, LIN or PAR, will do better in predicting the new data when fitted to the old? Deductive relations are indicated by solid arrows, probabilistic by broken.

A simple version of the kind of problem that Akaike discussed is depicted in Figure 1.14. LIN and PAR are the two models we examined in §1.5 of how temperature and pressure are related in the pressure cooker. You use the available data to find the maximum likelihood estimates of the parameters that each model contains; that is, you use the data to find L(LIN) and L(PAR); L(LIN) is the member of LIN that has the highest likelihood, and L(PAR) is likewise the maximum likelihood member of PAR. You then ask the following question: If you were to draw new data from the pressure cooker, would L(LIN) or L(PAR) do a better job predicting this new data? I hope the reader finds it puzzling how this question could be answered. The only data you can consult is the old data you already drew from the pressure cooker. Since LIN is nested inside PAR, you know in advance that Pr[old data | L(PAR)] $\geq$ Pr[old data | L(LIN)], no matter what the old data are. The only way the two fitted models can have exactly the same likelihoods is if the data fall exactly on a straight line; otherwise L(PAR) will have the higher likelihood. As already noted, more complex models inevitably fit the data at hand better than simpler models. But we know from experience that more complex models often do worse at predicting, not better. What else is there to consider here besides the likelihoods of L(LIN) and L(PAR)?

Bayesians may feel inclined at this point to appeal to the prior probabilities of LIN and PAR. But here we run into a wall. Since LIN entails PAR, Pr(LIN) $\leq$ Pr(PAR). The simpler model cannot have the higher prior probability – a point that Popper (1959) emphasized. This problem can be circumvented if we stipulate that the only models we are willing to

consider must be incompatible with each other. For example, if we require that LIN and PAR be set aside, and LIN* and PAR* considered instead (where all the parameters in this latter pair of models have non-zero values), the axioms of probability theory do not settle in advance which of the two has the higher prior probability. But even though it is logically consistent to say that $Pr(\text{LIN}^*) > Pr(\text{PAR}^*)$, it is hard to see how this can be anything more than a stipulation. Consider the parameter c that attaches to the squared term in PAR*. The claim that $Pr(\text{LIN}^*) > Pr(\text{PAR}^*)$ is equivalent to the claim that $Pr(c = 0) > Pr(c \neq 0)$. What objective reason could there be for thinking that this inequality is true?

Let us make the prediction problem more precise. After you draw the old data and use them to identify L(LIN) and L(PAR), you want to know how well these two curves will predict new data drawn from the same pressure cooker. But there is no one form that the new data set must take. Different data sets will differ from each other, though all are produced by the same underlying mechanism. The reason for expecting variability among data sets is that observations are subject to error. This means that when we ask how well a given fitted model will do in predicting new data, what we want to ascertain is how well it will do *on average* in this prediction task. L(LIN) may accurately predict one new data set but do less well in predicting another. By the *predictive accuracy* of a model M we mean how well *on average* M will do when it is fitted to old data and the fitted model is then used to predict new (Forster and Sober 1994). Imagine carrying out the task described in Figure 1.14 again and again. The *expected* performance of a model is what we want to know about.

There is a second refinement that needs to be added to this definition of predictive accuracy. What does it mean to talk about how accurately L(LIN) or L(PAR) predicts a single new observation in the pressure-cooker experiment? This new observation takes the form of a pair of temperature and pressure values $<x, y>$. When the temperature value x is fed into L(LIN) or into L(PAR), the output is a predicted value for the pressure y. We then can determine how close or far away the predicted value for y is from the observed value. We might do this by taking the difference between the two values and squaring it. Greater accuracy then means a smaller squared distance. Or we might compute the value of Pr(observed pressure value y | fitted model & temperature value x), with a larger likelihood indicating a higher degree of accuracy. These two approaches are related, in that squared distance is inversely related to likelihood, given some standard assumptions. The next question is how we should measure accuracy of prediction when there is more than one data point in the new

data set. We could sum the squared distances or we could compute the likelihoods relative to all the new data. But notice that as the data set we are trying to predict gets larger, the sum of squares increases and the likelihood must decline as more and more terms are multiplied together. The problem is that we want to define the notion of predictive accuracy so that it does not turn out that a model is automatically less predictively accurate with respect to a larger data set than it is with respect to a smaller one. This is a general point about how we want to conceptualize measurement devices. When we ask about the accuracy of a bathroom scale or a thermometer or a tuberculosis test, the answer should not depend on how many times the device is used. A natural solution is to think of the predictive accuracy of a model as its average accuracy *per datum* (Forster and Sober 1994; Forster 2001). This point about the concept of predictive accuracy is not in Akaike (1973); he was thinking about model comparison where all the models are asked to predict the same new data set, which contains the same number of observations that the old data set contained. In this context, the difference between *per datum* predictive accuracy and total accuracy over the entire data set does not matter. For the moment, I'll follow Akaike's lead and omit mention of the fact that predictive accuracy is a *per datum* quantity. Later on, I'll return to the question of how larger and larger data sets should be brought into the picture.

After isolating the prediction problem of estimating how accurately a model will predict new data when fitted to old, Akaike (1973) derived a result that bears on it:

Akaike's Theorem: An unbiased estimate of the predictive accuracy of model $M = \log\{Pr[\text{data} \mid \text{L}(M)]\} - k$.

We use the old data to find the likeliest member of model M and then take the natural logarithm (= base e) of its likelihood. We then subtract k, which is the number of adjustable parameters that the model contains. Notice that Akaike's estimate pays attention to both the model's fit to data and its simplicity. Akaike's theorem led to the formulation of the following model-selection criterion:

The Akaike information criterion (AIC): The AIC score of a model M, $\text{AIC}(M) =_{\text{def}} \log\{Pr[\text{data} \mid \text{L}(M)]\} - k$.

The absolute value of a model's AIC score is not what is interesting about this criterion. What matters is how the scores of different models *compare* when the models are fitted to the same data set. AIC is a proposal that addresses the task of model *comparison*, not the task of model *acceptance* and *rejection* (Sakamoto et al. 1986: 84).[41]

How will the AIC scores for LIN and PAR compare? PAR will have a higher value for the first addend than LIN; L(PAR) will have the higher likelihood, and therefore the higher log-likelihood. But PAR contains one more adjustable parameter than LIN; in this respect, PAR is worse than LIN. Each model has one piece of good news and one piece of bad. The AIC scores of the two models depend on the character of the data. With some possible data sets, LIN will score better; with others, PAR will. The question is whether the data at hand depart sufficiently from linearity to justify the loss in simplicity that comes from shifting from LIN to PAR. AIC provides a principled basis for deciding how fit-to-data should be *traded off* against simplicity.

Three questions need to be answered about Akaike's theorem. What does "unbiased" mean? How is Akaike's theorem related to AIC? And what are the assumptions from which the theorem was derived?

A bathroom scale provides an unbiased estimate of your weight if the average of its values over many weighings is your true weight. In this hypothetical run of tests, we assume that your true weight remains the same. An unbiased estimator is *centered* on the true value; if your true weight is x, the *expected value* of the scale's readings is x. However, an unbiased scale may on any given occasion provide an estimate that is way too high or an estimate that is way too low. How much (squared) variation there will be, on average, among different estimates is called the *variance* of the estimator. A reading produced by a scale that has a high variance may have a very small probability of being close to the true value.[42] The best bathroom scale would be unbiased and have very small variance, but suppose you had to choose which of these virtues you prize more. Suppose one scale is perfectly unbiased and has large variance while a second has a small bias and a small variance. It is easy to imagine preferring the second to the first; this scale tends to read too high or tends

[41] I have presented AIC as a measure of predictive *accuracy*, so models with bigger scores are better than models with smaller ones. The reader should realize that AIC is usually presented as an expected *distance* from new data, in which case models are better the smaller their scores.

[42] Consider Figure 1.2 and suppose that the curves shown there represent the readings that three different unbiased scales might produce when an object that weighs 78 pounds is placed upon them.

to read too low (you don't know which) but it rarely is off by more than an ounce. This scale might be better than one that is centered on your true weight but tends to swing from 10 pounds too light to 10 pounds too heavy. These considerations indicate that it would be desirable to show, not just that AIC is an unbiased estimator but also that it is an estimator of minimum variance, or that it has a lower variance than other estimators that one might use. Akaike's theorem does not address this question; Sakamoto et al. (1986: 76–80) describe the variance of AIC estimates, but there is more to be learned about this subject. In any event, recall the frequentist setting of these questions about unbiasedness and variance. We are discussing the "operating characteristics" of a general policy, that of using AIC to estimate the predictive accuracy of models. Even if AIC is unbiased and has a low variance, that does not entail that when LIN scores better than PAR with respect to the data drawn from your pressure cooker, that LIN will *probably* be more predictively accurate. Posterior probabilities require priors, and this is something that frequentists disdain.

I turn next to the assumptions that Akaike used to prove his theorem. Akaike's proof uses the "normality assumptions" that are frequently exploited in mathematical statistics. This means, roughly, that each of the parameters in the model whose value you need to estimate will be such that repeated estimates form a normal distribution. There is, second, the assumption that old and new data sets are drawn from the same underlying reality. When you accumulate a data set on your pressure cooker and decide which of LIN and PAR will be more predictively accurate with respect to a new data set you have yet to see, you need to assume that the true but unknown law governing the pressure cooker won't change between your old observations and your new ones. Temperature and pressure must be related to each other in the same way across all data sets. There is a third assumption that goes into the proof. Your old data were accumulated by looking at various temperature values. These were chosen in accordance with some sampling procedure; perhaps these temperature values were drawn at random from a range of values. The theorem assumes that your new data will be drawn from the same distribution of temperature values. So, the relationship of X and Y is the same across data sets and so is the distribution of X values. Together these constitute a Humean *uniformity of nature assumption* (Forster and Sober 1994).

This last assumption means that Akaike's theorem and his criterion do not apply to inference problems in which you are trying to

extrapolate – situations in which your sample is constrained to come from one range of temperature values and you want to make a prediction concerning what is true outside that range (Forster 2000b, 2001). It is useful to think about this point in relation to the fact that AIC is asymptotically equivalent with a different model-selection method that is called take-one-out cross-validation (Stone 1977). The cross-validation criterion makes no mention of simplicity. Rather, to test a model like LIN, you set aside one of the n data points in the sample, fit LIN to the $n - 1$ data points that remain, and then see how well L(LIN) predicts the data point that was set aside. The procedure is repeated for each of the other data points; then you compute the average performance of the model across these n trials. That's the cross-validation score of LIN. The same procedure is carried out for other models and then the scores of different models are compared. Cross-validation is a general kind of procedure in which one gauges a model's performance by dividing one's data into a training (or calibration) set and a prediction (or test) set. In the application just described, the n data points are divided into $n - 1$ for training and 1 for prediction. This is called take-one-out cross-validation. The cross-validation framework also allows you to consider take-two-out, take-three-out, and so on. It is possible that one model scores better than another in terms of take-one-out, while the reverse is true for take-ten-out.The fact that AIC is equivalent with the former rather than the latter is telling. AIC is a solution to *one* prediction problem, but there are others.

It is interesting how AIC's comparison of LIN and PAR changes as the size of the data set increases. Consider the fact that a model's AIC score is influenced by two quantities, but only one of them changes as more data accumulate. The log-likelihoods of L(LIN) and L(PAR) both decline as more data roll in, but the number of adjustable parameters in each model of course stays the same. We know from the definition of AIC that AIC(LIN) > AIC(PAR) precisely when

$$\log\{Pr[\text{data} \mid L(\text{LIN})]\} - \log\{Pr[\text{data} \mid L(\text{PAR})]\} > -1.$$

The reason "−1" is on the right-hand side of this inequality is that PAR has one more adjustable parameter than LIN. Thus, what we want to know is whether

$$\log\left\{\frac{Pr[\text{data} \mid L(\text{LIN})]}{Pr[\text{data} \mid L(\text{PAR})]}\right\} > -1$$

and (since the logarithm is base $e \approx 2.72$) this is true precisely when

$$\text{(A)} \qquad \frac{Pr[\text{data} \mid L(\text{LIN})]}{Pr[\text{data} \mid L(\text{PAR})]} > \frac{1}{2.72} \approx 0.37.$$

This inequality describes what it takes for LIN to have the higher of the two AIC scores. It may be true for small and middling sized data sets, but, with a sufficiently large data set, the inequality must be false; PAR must score better.[43] This is a sensible feature of AIC; the greater simplicity of LIN over PAR can compensate for L(LIN)'s lower likelihood for some sample sizes, but eventually it cannot. If there is a slight parabolic bend in the data, you might want to ignore this when sample sizes are small, but if the bend is still there when you have lots of data, you'd be foolish to ignore it. The impact of simplicity on model evaluation *should* depend on sample size. The prediction problem that AIC is meant to address involves using an old data set to predict a new one *of approximately the same size.* If LIN scores better than PAR, given the data you have at hand, it does not follow that LIN would score better on a data set that is vastly larger.

The likelihood ratio test (§1.6) also applies to models like LIN and PAR, so it is useful to review some differences between it and AIC. First, there is the fact that the likelihood ratio test gives advice about whether the null hypothesis should be rejected; it therefore requires an arbitrary decision about how small the likelihood ratio of the two fitted models must be for one to reject the null. In contrast, AIC gives advice about model comparison, not model acceptance and rejection, and what it compares are estimates of predictive accuracy, not truth. Second, the mathematical underpinnings of the likelihood ratio test sanction its use only on nested models, but what could it mean to accept LIN and reject PAR, given that LIN entails PAR? The mathematics behind AIC justify its use on nested and non-nested models alike. Third, the likelihood ratio test violates the principle of total evidence; one doesn't look at the point

[43] Recall the point in §1.2 about the two witnesses whose testimonies agree. The fact that the testimonies are *independent* of each other, conditional on the proposition reported, was important in that discussion; similarly, in the present context, each datum is independent of every other, conditional on L(LIN) and conditional on L(PAR). When a small number of independent and reliable witnesses all say that proposition P is true, it is an open question whether the likelihood ratio of P to *not*P will exceed some threshold; but for any threshold you specify, the likelihood ratio must exceed that threshold if there are sufficiently many unanimous witnesses. Similarly, if the data set is large enough, the log-likelihood of L(PAR) will exceed that of L(LIN) by any threshold you name, including the one described in proposition (A).

values in the data but only at a logically weaker description that says whether or not those data points fall in a given region. AIC abides by the principle of total evidence.

Identifiability

AIC penalizes models for being complex, but there are some models that are so complex that AIC does not even apply. It isn't that models that have more than 23,453,450,965 parameters are in principle beyond the pale. Rather, the limitation I have in mind comes from the number of data points in the observations one has at hand. A model with more parameters than there are data points will (typically) not be *identifiable*. A failure of identifiability means that there is no such thing as *the* maximum likelihood estimate of the parameters that the model contains. This failure of uniqueness can occur even in simple models, provided that the data set is sufficiently small. Consider our old friend LIN, a simple model if ever there was one. Suppose your data set consists of a single data point. There are infinitely many straight lines that pass exactly through this point; each has a likelihood that cannot be bettered. What would it mean to talk about LIN's predictive accuracy in this case? One would have to envision fitting the model to this single datum and then using "the" fitted model to predict other data sets that contain a single new data point. However, there is no such thing as "the" fitted model in this case. AIC does not even apply.

A data set that contains a single observation may seem like a joke, but the point about identifiability applies to the larger data sets that scientists actually use. AIC cannot be applied to models that are not identifiable. This means that our data limit the kinds of theories we can evaluate. In contrast, Bayesianism does not prohibit the assignment of prior probabilities and likelihoods to such models; for subjective Bayesians, such quantities are always well defined. Wittgenstein says in the last line of the *Tractatus* that *whereof one cannot speak, one must remain silent*. AIC embodies a kind of Wittgensteinian circumspection; Bayesianism is bolder.

Is AIC statistically inconsistent?

I mentioned earlier that estimators can be assessed for their unbiasedness and for their variance. I now want to consider a third property of estimators that one might value or even demand. This is the property of

statistical consistency. Don't confuse this with the property of *logical* consistency. An estimator is statistically consistent when it converges on the true value of the parameter being estimated as more and more data are added. For example, suppose you want to infer the probability a coin has of landing heads when it is tossed. The policy of using the frequency of heads in a sample of tosses as your estimate is statistically consistent (a point that arose in connection with Reichenbach's straight rule in §1.2). This is the method of maximum likelihood estimation; by using this procedure, the estimate will converge on the true probability of heads as the number of tosses is increased. Does AIC converge on the true value of a model's predictive accuracy when the size of the data set is increased? That is, if one model is in fact more predictively accurate than another, can AIC be relied upon to award the first model the higher score as the size of the data set is increased without limit?

The question of AIC's consistency has often been misunderstood. The question is not whether AIC converges on the *true* model. AIC is not a device for assessing which model is true but provides an estimate of a model's predictive accuracy (Forster 2001); as already noted, it is perfectly legitimate to use AIC to evaluate a set of models all of which are known to contain idealizations and so all are known at the outset to be false. Also, when models are nested, you know in advance that the most complex model is true if any of them are. There is no need to use data or a model-selection criterion to ascertain this fact. Sometimes the question of consistency has been taken to be whether AIC converges on the true model that has the smallest number of adjustable parameters. So, if LIN and PAR are both true, the task assigned to AIC is to converge on LIN when the data are made large without limit. I pointed out before that this is not something that AIC will do. As a data set is made larger and larger, eventually the most complex model will have the best AIC score if the models considered are nested. This is not a defect in AIC. This most complex model *is* the model of greatest predictive accuracy for data sets that are large enough; AIC has succeeded in converging on the best model in that sense. However, the point of AIC is not to ascertain which models will be most predictively accurate for enormous or infinite data sets; the problem is to cope with the finite data sets one has at hand. If you make 200 observations of pressure and temperature on your pressure cooker, the problem is to figure out which model will do best in predicting what you'll observe if you draw another 200; it is a different problem to figure out which model will do best if old and new data sets contain 2,000,000,000,000,000 observations (Burnham and Anderson 2002: 298).

Demanding that AIC converge on the most predictively accurate of the models considered as data sets are made larger and larger is a bit like demanding that a bathroom scale converge on your true weight as you get heavier and heavier. The scale will fail to converge on a single value because the target is moving, not stationary. It makes more sense to demand that the scale's readings be centered on your true weight. If you weigh a single object of fixed weight again and again, will the average of these weighings converge on the object's true weight as the number of weighings increases? This is what the scale will do if it is unbiased. Repeatedly "weighing" a set of models using AIC will do the same thing, since AIC is an unbiased estimator.

Bayesian model selection

The criticism that AIC is statistically inconsistent is often voiced in the context of claiming that the Bayesian information criterion (BIC) derived by Schwarz (1978) is better. BIC will converge on the smallest true model, if the set of models you are considering includes one that is true. However, it is questionable why consistency in this sense should be thought a virtue if the competing models considered are not exhaustive; in this case, there is no guarantee that any of them is true. Also, if the models are nested, you know in advance that the largest model is true if any of them are. Why is it important to converge on the *smallest* true model, rather than on *a* true model? The latter task is easily achieved (if one of the models is true) and no model-selection criterion is needed to do this; the fact that the former task is harder does not explain why it is worthwhile.

Logically prior to this question about consistency is a more fundamental point of difference that separates BIC and AIC. As noted, the goal of AIC is to compare different models for their expected predictive accuracies. The goal of BIC, however, has nothing to do with predictive accuracy. This model-selection criterion has a Bayesian goal: to estimate the average likelihoods of composite models. LIN, for example, is an infinite disjunction of different straight lines, each of which confers its own probability on the data at hand. We saw earlier that the likelihood of LIN must be a weighted average over the likelihoods of these different straight lines, where the weighting terms have the form $Pr(L_i \mid LIN)$. Since BIC aims to estimate $Pr(\text{data} \mid LIN)$, the method must make assumptions as to what values these weighting terms have. Those not sold on Bayesianism despair of grounding these weighting terms in anything objective, and for that reason will be skeptical of BIC. Although a commitment to

the values of these weighting terms must figure in any valid derivation of BIC, the weighting terms do not appear in the final product, which is the criterion that Schwarz (1978) derived for the average likelihood:

The Bayesian information criterion: The likelihood of model $M \approx \log\{Pr[\text{data} \mid L(M)]\} - k[\log(n)]/2$.

Here, k is the number of adjustable parameters in the model, and n is the number of data. BIC imposes a bigger penalty for complexity than AIC does; notice also that the second addend in BIC increases as the sample size increases, which is not true of the second addend in AIC. Schwarz (1978) derives BIC by assuming that the models under consideration have the same priors. Given this assumption, the criterion not only estimates average likelihoods; it also estimates posterior probabilities.

BIC is often applied to nested models, the idea being that BIC identifies the model in the set of competitors that has the highest posterior probability. But, as already noted, no matter what the data say, LIN cannot be more probable than PAR if LIN entails PAR. When models are nested, one can tell *a priori* which model has the highest prior and the highest posterior probability; there is no need to consult the data to figure this out and no need to consult a model-selection criterion. If the data lead BIC to say that LIN has a higher posterior probability than PAR, the Bayesian criterion has simply made a mistake and its testimony should be set aside. This problem can be avoided by restricting the application of BIC to non-nested models.

Although BIC was derived as a device for estimating average likelihoods and posterior probabilities, we still may ask how well it performs as an estimator of predictive accuracy. We know from Akaike's theorem that AIC is unbiased; since BIC differs from AIC by a constant, BIC must therefore be a biased estimator of predictive accuracy. A further defect in BIC also follows: BIC's estimates of predictive accuracy have a larger expected squared error than the ones generated by AIC (Forster and Sober, in preparation).

The debate over AIC and BIC needs to be understood, in the first instance, as a debate over choice of goals – estimating predictive accuracy versus estimating average likelihood. Only after a goal has been chosen can the question be raised as to which criterion does better in achieving that goal.

The subfamily problem

A curve, since it contains no adjustable parameters, is a member of many models. For example, "$y = 3 + 4x$" is a member of LIN, but it also is a

member of PAR and of lots of other models besides. Given this, how is a curve's AIC score to be computed? Its log-likelihood is univocal, but what penalty should we impose on it for its degree of complexity? If we view the curve as a member of one model, we'll apply one penalty term, but if we view it as a member of a different model, we'll apply another. This is the subfamily problem (so called by Forster and Sober 1994).

One step towards solving this problem is to recognize that AIC applies to *models* and that there is no need for AIC to say which model is the one to which a curve "really" belongs. The predictive accuracy of a model is its average performance as it is fitted to old data sets and then makes predictions about new ones. There is no paradox in saying that LIN and PAR may differ in their predictive accuracies even if L(LIN) and L(PAR) happen to be identical curves in virtue of the (collinear) data set one has at hand. AIC also applies to *curves*, but this is because curves are a limit case; they are models that contain zero adjustable parameters. A curve's AIC score is just its log-likelihood (since its complexity penalty is zero). Thus, it can turn out that "$y = 3 + 4x$" has a lower likelihood than "$y = 3 + 4x + 0.001\ x^2$," and so the former has the lower AIC score, and yet LIN has a higher AIC score than PAR, where the two curves happen to be the best-fitting members of the two models, respectively. The two curves have their own AIC scores, LIN has a third, and PAR has a fourth.

Although this point shows that AIC is not guilty of contradicting itself (or of arbitrarily deciding which model a curve "really" belongs to), it does leave another question unanswered. How should we *use* AIC to make predictions? This is a *pragmatic* question in the sense of that term discussed earlier in connection with the principle of total evidence (§1.3). Should we apply AIC to the two curves L(LIN) and L(PAR) and therefore use the latter to make our predictions? Or should we apply AIC to LIN and PAR and allow the data to help us decide which model is better? Focusing exclusively on curves has the result that we always choose the curve that comes from the largest model. The motivation for using AIC is to find models that make accurate predictions; applying AIC only to fitted models prevents the criterion from helping us to achieve that end. But there is another reason to decline to use AIC in this way. AIC provides unbiased estimates of predictive accuracy, regardless of whether it is applied to LIN and PAR, or to L(LIN) and L(PAR), or to all four. One reason to score LIN and PAR, rather than L(LIN) and L(PAR), is that AIC has greater variance when it is applied to smaller models (Escoto 2004); applying AIC to fitted models is more apt to produce inaccurate estimates of predictive accuracy.

There is another dimension to this pragmatic problem. The fact that AIC is a comparative principle, not a criterion for acceptance, shows that it would be a mistake to make a prediction by using the model that has the best AIC score while ignoring all the other models that were considered. After all, AIC is an estimator that is subject to error. This suggests that predictions should be made by *model averaging* (Burnham and Anderson 2002). If you want to predict the pressure that will result when you set your pressure cooker to a given temperature, you should consider the prediction made by the model with the best AIC score, the prediction made by the second best, and so on. You can average these different predictions by using *AIC weights* – giving more weight to predictions that come from models that have better AIC scores.

The scope of AIC

I have used the models LIN and PAR to explain what AIC amounts to, but this should not be taken to mean that AIC is relevant only to "curve-fitting problems." Philosophers sometimes disparage curve fitting as a kind of naïve inductive inference in which the hypotheses considered seek merely to identify patterns that hold among observational quantities. Model-selection criteria, including AIC, are not limited to such problems. They also apply to *causal models* that say that an effect term is influenced by the values of any number of input variables. In Chapters 3 and 4, we will see how model-selection ideas apply to problems in evolutionary biology.

Although I have argued that the dispute over AIC versus BIC is based on a failure to realize that they are estimators of different quantities, the fact remains that there are different model-selection criteria that all focus on the goal of estimating predictive accuracy. For example, there is a version of AIC derived by Sugiura (1978) that is better to use when some of the models under evaluation have a large number of parameters relative to the number of observations available; it is called AIC_c and imposes a larger penalty for complexity than AIC does.[44] There is also a criterion (TIC) derived by Takeuchi (1976). These criteria all compute the likelihood of the best fitting member of a model and then impose a penalty for complexity; they differ over what that penalty term is. I mentioned earlier

[44] Burnham and Anderson (2002: 50) recommend using AIC_c precisely when $n/k < 40$, where n is the number of observations and k is the number of parameters in the largest model under evaluation.

that AIC is equivalent to take-one-out cross-validation; this raises the question of what the statistical properties are of cross-validation methods that take more than one out, and of what use such methods are in different inference problems (Forster 2006, 2007). And there is also the question of what model-selection criteria are best when the goal is extrapolation, not interpolation. What I find striking in this diversity of problems and solutions is what they have in common. This is the Akaike framework, within which all these approaches are to be understood. We want to know how accurately a model will predict new data when it is fitted to old. How well the model fits the old data is relevant to this question, but so is the model's complexity (the number of adjustable parameters it contains). This framework helps explain why scientists should bother to test models that they know are false. If the goal were to decide which models are true, there would be little point in testing idealizations. But predictive accuracy is a different story, and it has its own epistemology. Bayesianism, likelihoodism, and the Neyman–Pearson framework each have their different drawbacks when applied to this kind of problem. The subject that Akaike initiated throws new light on these issues, and there is the promise of more light to come.

Realism and instrumentalism

Virtually everyone who follows professional basketball believes that players sometimes have "hot hands." When players are hot, their chance of scoring improves, and teammates try to feed the ball to them. Gilovich et al. (1985) tested this widespread belief by doing a statistical analysis of scoring patterns in the National Basketball Association. Their conclusion was that one cannot reject the null hypothesis that each player has a constant probability of scoring throughout the season; belief in hot hands, they say, is a "cognitive illusion."[45] Basketball mavens reacted to this statistical pronouncement with total incredulity. Placing this dispute in the Akaike framework allows it to make more sense. Scientists should not feel shy about admitting that the null hypothesis is false. The idea that each player never wavers in his probability of scoring *is* preposterous. But

[45] See Wardrop (1999) for a skeptical assessment of Gilovich et al.'s analysis. Wardrop argues that Gilovich et al. tested hypotheses about *correlation* (whether a player's probability of scoring on a given shot if he scored on earlier shots is greater than his probability of scoring if he missed previously), but did not assess the issue of *stationarity* (maybe a player's probability of scoring suddenly shifts from one value to another).

even if this silly hypothesis is false, there still may be a point to seeing how accurately it predicts new data. Perhaps the truth about basketball players is very complex; their scoring probabilities change as subtle responses to a large number of interacting causes. If so, players and coaches may make better predictions by relying on simplified models. Even if hot hands are a reality, trying to predict when players have hot hands may be a fool's errand.

The problem of evaluating how accurately models predict new data when fitted to old has a philosophically interesting property: a model known to be false will sometimes be more predictively accurate than a model known to be true. What is perhaps more surprising is that we can sometimes *estimate* which of them we should expect to be more predictively accurate and the methods available for assessing this sometimes favor false models over true ones. The Akaike framework thus breathes new life into an old philosophy. *Instrumentalism* is the view that the goal of scientific inference is to find theories that make accurate predictions, not to find theories that are true.[46] It stands opposed to *scientific realism*, which holds that the goal is to find true theories.

The debate between realism and instrumentalism can't be resolved by polling scientists as to what their goals are. Some scientists say that they want to find out what is true while others say that their object is to find theories that make accurate predictions; all may be sincerely reporting their personal goals, but that is not what is at issue. The philosophical debate concerns what *scientific inference* is able to attain, not what *scientists* yearn for. If the inference procedures used in science are able to discover which theories are true, or which are probably true, then realism is correct. If those procedures are capable only of discovering which theories will make the most accurate predictions, then instrumentalism is. Both philosophies need to be tempered by the fact that scientists rarely are able to examine a set of hypotheses that exhaust the possibilities (Stanford 2005). Scientists deal with the theories that have been developed thus far, and no one can foresee the novel theories that future innovators may put on the table. This sobering fact about the limitations that scientists perpetually face means that the best that scientists can do at any time is to render comparative judgments. Realism should be understood as the

[46] Instrumentalism is sometimes also formulated as a semantic thesis – that scientific theories are neither true nor false, but are merely instruments for making predictions. The proper response is that there is no reason to think that theories lack truth values, and no reason to burden an epistemological thesis with an outmoded philosophy of language (Sober 2002).

claim that scientific modes of inference indicate which of a set of competing hypotheses is the best candidate for being true; instrumentalists think that science is in a position only to say which of the competitors can be expected to make the most accurate predictions.

Instrumentalism and realism are usually formulated as *global* theses. They are claims about *all* the hypotheses that scientists investigate. It doesn't matter whether the hypotheses in question are models or fitted models, any more than it matters whether they are part of the subject matter of one science or another. The Akaike framework shows that this global formulation of the problem needs to be recast. The framework makes room for an instrumentalist philosophy of *models*. The fact that one model (M_1) has a better AIC score than another (M_2) is grounds to think that the first will be more predictively accurate; it is not grounds for thinking that M_1 is true, or more probably true, or better supported as a candidate for being true. However, this difference in the scores of the two models has another implication concerning the truth of the *fitted models* – Akaike's theorem can also be formulated as the thesis that the AIC score of a model M is an unbiased estimate of the closeness to the truth of the fitted model $L(M)$, where closeness is measured by the Kullback–Leibler distance.[47] With respect to the pressure cooker in your kitchen, there is a true but unknown curve that describes how temperature and pressure are related. Specific curves have different Kullback–Leibler distances to that true curve. Models are instruments for finding curves that are close to the truth and models are compared with each other to determine how well they advance that goal.[48] The Akaike framework therefore makes plausible a mixed philosophy: instrumentalism for models, realism for fitted models (Sober 2002b). When a false model F and a true model T are both fitted to the data, $L(F)$ will sometimes be closer to the truth than $L(T)$. AIC and other model-selection criteria seek to provide guidance as to when this is so.

[47] Suppose t is the true distribution $(p_1, p_2, \ldots, p_n)$ of a discrete random variable and c is a candidate distribution $(\pi_1, \pi_2, \ldots, \pi_n)$. The KL distance from the candidate c to the truth t is $I(t, c) = \sum p_i \log (p_i/\pi_i)$. Notice that the true distribution provides the weighting on the log of the ratio. KL is a "directed distance;" the distance from c to t (where t is true) doesn't have to be the same as the distance from t to c (where c is true). See Burnham and Anderson (1998) for further discussion.

[48] The relation of AIC to Kullback–Leibler distances provides an easy answer to the question of why one should care about AIC estimates if one has no interest in using fitted models to predict *new* data. One still might care about finding fitted models that are close to the truth when Kullback–Leibler distance is used to measure closeness.

One challenge to this limited form of instrumentalism begins with the idea that instrumentalism and realism should be thought of as claims about the *ultimate* goals of science. Maybe finding models that make accurate predictions is a mere tactic that science deploys in the larger campaign. A realist can grant that it is useful to find idealized models that make accurate predictions if such models are worth having because they help one get to the truth, and truth is the ultimate goal. A defense of this response requires more than the psychological fact that scientists often would *like* to find true theories. What is needed is an account of how scientific inference makes it possible to turn assessments of the predictive accuracy of models into claims about which theories are true. I've already mentioned that fitted models may be nearer or farther away from the truth, and that there is an intimate connection between M_1's being a better predictor than M_2 and $L(M_1)$'s being closer to the truth than $L(M_2)$. Perhaps the objection can then be put by saying that the real goal of science is to discover which fitted models are true and that models themselves are mere means to that end. Again, this may or may not be true as a psychological claim about what interests various scientists (though, in fact, scientists are often more interested in models than in fitted models). But how can it be justified as a claim about scientific inference, not about the psychology of scientists? If finding models that are accurate predictors and fitted models that are close to the truth go hand in hand, then it is hard to see that one is logically prior to the other. Given this, the mixed thesis of "instrumentalism for models, realism for fitted models" may be more satisfactory than either global realism or global instrumentalism.

What is a parameter?

AIC says that the complexity of a model is relevant to estimating its predictive accuracy; BIC says that a model's complexity is relevant to estimating its average likelihood. Both measure complexity by counting parameters. This raises an important question. A model is a *proposition*, distinct from the sentence in some language in which it happens to be expressed; the proposition that temperature is linearly related to pressure is no more a part of English than it is part of Chinese. Yet, the number of parameters in a model seems to be a syntactic feature of how the model happens to be described; by changing the language used, you seemingly can change the number of parameters the model contains. If so, how could the number of parameters be relevant to ascertaining these epistemically

relevant properties of the model itself – its predictive accuracy or its average likelihood?

This question can be fleshed out by way of our running example, the comparison of LIN and PAR. I've said that LIN has two parameters and PAR has three (ignoring, for the moment, the error term that each deploys). Any straight line of the form $y = mx + b$ can be represented as a point in a two-dimensional parameter space in which one axis is its slope (m) and the other is its y-intercept (b). A straight line in the x-y plane is just an ordered pair of numbers $<m,b>$ in this parameter space. In the nineteenth century, Georg Cantor discovered that the number of points in a plane is the same as the number of points on a line. This means that there is a one-to-one (injective) mapping from ordered pairs to single numbers. An example of this kind of mapping is provided by *interleaving*. Consider a plane whose possible m values run from 0 to 1 and whose b values do the same. Each point in this unit square can be expressed as an ordered pair, each of whose members is a decimal expansion of the form

$$\mathrm{m} = 0.\mathrm{m}_1\mathrm{m}_2\mathrm{m}_3 \cdots \quad \mathrm{b} = 0.\mathrm{b}_1\mathrm{b}_2\mathrm{b}_3 \cdots$$

By interleaving we can represent this pair of numbers as a single number

$$i = 0.\mathrm{m}_1\mathrm{b}_1\mathrm{m}_2\mathrm{b}_2\mathrm{m}_3\mathrm{b}_3 \cdots$$

Notice that there is a function from each $<m,b>$ pair to a single number i, and another function from each possible value of i back to that single $<m,b>$ pair. So, in what sense are there *two* parameters (m and b) in LIN? Why not say, instead, that there is just *one* (namely i)? And if LIN has just one parameter, so does PAR (since you can interleave triplets just as well as pairs). The difference in complexity of the two models seems to be an artifact of the notation we arbitrarily choose.

This question was important in nineteenth-century mathematics where the problem was to describe what *dimension* means. Is there a rigorous and linguistically invariant way to express the thought that a plane has two dimensions while a line has just one? The problem was solved in the twentieth century by Brouwer, who isolated a concept of dimension that is *topologically invariant* (Courant and Robbins 1959: 249–51; Dauben 1994). The idea of interleaving can be used to convey the intuitive idea.

Consider three straight lines (one of which is true); each is defined by its coordinates in the $<m,b>$ parameter space:

$$\text{Truth} = <1, 1> \qquad L_1 = <2, 1> \qquad L_2 = <1, 3>.$$

Notice that L_1 is closer to Truth than L_2 is. If we interleave each of the ordered pairs, we obtain:

$$\mathrm{I}(\text{Truth}) = 11 \qquad \mathrm{I}(L_1) = 21 \qquad \mathrm{I}(L_2) = 13.$$

Notice that $\mathrm{I}(L_2)$ is closer to I(Truth) than $\mathrm{I}(L_1)$ is. Although the mapping achieved via interleaving is injective, it is not *distance preserving*. The mapping does not have the property that points that are close together in the $<m,b>$ plane have images in the line that are always close together. There is more to the idea of topological invariance than that of a mapping that is distance-preserving, but the example of interleaving helps elucidate what a parameter is in model-selection theory. If a space has n dimensions, then there is no one-to-one, continuous, and distance-preserving mapping from that space to another space that has m dimensions, if $n \neq m$. Dimensionality is in this sense an invariant quantity.

What does this imply about the dimensionality of LIN? Is it two, or one, or some other number? By definition, it must be unique, the possibility of interleaving notwithstanding. To answer this question would lead us too far afield. But I hope the following two comments are helpful. First, consider the relationship of *LIN* to *PAR*. *LIN* is nested in *PAR*. This is a fact about the two propositions and has nothing to do with the language in which they happen to be expressed. It is a consequence of this nesting relationship that *LIN* cannot have a higher dimensionality than *PAR*. And since the fact about the nesting relationship is invariant, the same holds for the fact about dimensionality (Forster 1999). The second comment returns us to the content of Akaike's theorem. As noted, the theorem identifies an unbiased estimate of the predictive accuracy of a model M, or, equivalently, an unbiased estimate of the Kullback–Leibler distance from $L(M)$ to the true but unknown probability distribution T. Expressed in this second way, Akaike's theorem states that:

$$E[KL\text{–Closeness of } L(M) \text{ to } T] = [\text{Log-likelihood of } L(M)] - k.$$

The left-hand side describes a language-independent quantity, and the same is true of the first addend on the right. It follows that k must be

language independent as well. Again, this does not tell you how to determine what value of k a model has. But it does assure you that, whatever it is, it is not an artifact of notation.

Is AIC frequentist?

I have classified AIC as a type of frequentism; I now want to consider briefly whether this classification makes sense. I have emphasized that AIC isn't a criterion for acceptance and rejection and that it does not violate the principle of total evidence. What is more, the AIC score of a model does not depend on the stopping rule used. These properties of AIC separate it from significance tests and the Neyman–Pearson theory. If AIC is frequentist, it is a different kind of frequentism.

Akaike (1973) refers to his result as "an extension of the maximum likelihood principle," but this phrase should not lead us to conclude that AIC is a form of likelihoodism. AIC does not say that the best model is the one that has the highest average likelihood, nor does it say that model M_1 is better than model M_2 precisely when $L(M_1)$ has a higher likelihood than $L(M_2)$. It is even clearer that AIC is not Bayesian. In using AIC, you are not estimating the probability that a model is true, nor are you estimating the probability that one model will be more predictively accurate than another. To reach conclusions about such posterior probabilities, you would need prior probabilities, and these play no role in AIC.

The main reason that AIC is viewed as a frequentist construct is the character of Akaike's theorem, which establishes that this estimation procedure has the long-run operating characteristic of being unbiased. This is just the sort of property that frequentists care about. Of course, they recognize that other operating characteristics are relevant as well. Is a procedure statistically consistent? What is its variance? Is it admissible? As noted in §1.5, Bayesians and likelihoodists do not object to the evaluation of procedures; they find nothing amiss in comparing the Madison tuberculosis test with the one manufactured in Prairie du Chien. However, they insist that there is a further question that needs to be asked: How should one evaluate a given *estimate* (never mind what method of estimation was used to construct it)? Likelihoodists want to know how well supported the estimate is, where support is understood in terms of the law of likelihood. Bayesians want to know how probable it is that the estimate is true (or close to the truth). Frequentists deny that this second question makes any sense; they hold that *estimators* have long-run

operating characteristics, but there is nothing further to be said about the individual *estimates* that those estimators generate.

The fact that Akaike's *theorem* addresses a kind of question that frequentists think is important does not show that *AIC scores* are meaningless from a Bayesian or likelihoodist point of view. Of course it is possible for M_1 to have a better AIC score than M_2 even though M_1 has the lower average likelihood and even though $L(M_1)$ is less likely than $L(M_2)$. But the law of likelihood and AIC still could join hands in friendship if AIC scores provided evidence concerning the predictive accuracies of different models, where evidence is understood in terms of the law of likelihood. Think of AIC as a measurement device, like a thermometer; perhaps AIC scores are to predictive accuracy as thermometer readings are to temperature. If a thermometer assigns a higher number to one object than it does to another, we take that to be evidence that the first object has a higher temperature than the second. Perhaps the same is true of AIC scores. The relevant property of thermometers can be described as follows. Suppose the thermometer readings on objects O_1 and O_2, $R(O_1)$ and $R(O_2)$, are such that $R(O_1) - R(O_2) = x > 0$. This observation indicates that the best point estimate of the temperature difference is positive when

There exists a $y>0$ such that for all $z<0$,

$$Pr[R(O_1) - R(O_1) = x \mid Temp(O_1) - Temp(O_1) = y] > Pr[R(O_1) - R(O_2) = x \mid Temp(O_1) - Temp(O_2) = z].$$

What would it take for the same thesis to hold for AIC scores and their relationship to the predictive accuracies of different models? What would be true is that, when we observe that model M_1 has an AIC score that is x units larger than the AIC score of model M_2, that the best point estimate of the difference in predictive accuracies is positive. That is,

There exists a $y>0$ such that for all $z<0$,

$$Pr[AIC(M_1) - AIC(M_2) = x \mid PA(M_1) - PA(M_2) = y] > Pr[AIC(M_1) - AIC(M_2) = x \mid PA(M_1) - PA(M_2) = z].$$

This inequality does not follow from Akaike's theorem. And it may not hold for *all* values of x – e.g., when x is very close to zero (Forster and Sober, in preparation) – but when it *does* hold, Bayesians and likelihoodists should have no qualms about viewing AIC scores as evidence. AIC began life with a frequentist pedigree, with Akaike's theorem. But

AIC scores may be essentially tied to frequentism no more than thermometer readings are.

1.8 A SECOND TEST CASE: REASONING ABOUT COINCIDENCES

When Evelyn Marie Adams twice won the New Jersey lottery, the *New York Times* said that the odds of this happening by chance are 1 in 17 trillion; this is the probability that Adams would win both lotteries if she had purchased a single ticket for each and the drawings had been at random. In fact, the newspaper made a small mistake. If the goal is to calculate the probability of Adams' winning those two lotteries, the reporter should have taken into account the fact that Adams purchased multiple tickets; the newspaper's very low figure should therefore have been somewhat higher. However, the typical response of statistical sophisticates is that this modest correction misses the point. For sophisticates, the relevant event to consider is not that Adams won those two lotteries, but the fact that someone won two state lotteries at some time or other. Given the many millions of people who have purchased lottery tickets, this is "practically a sure thing" (Diaconis and Mosteller 1989: 859).

Was Adams' double win a mere coincidence? Or were these two lotteries rigged in her favor? Diaconis and Mosteller say that the relevant principle to use when reasoning about coincidences is the *law of truly large numbers*. This says that, "with a large enough sample, any outrageous thing is likely to happen." They cite Littlewood (1953) as having the same thought; with tongue in cheek, Littlewood defined a miracle as an event whose probability is less than 1 in 1 million. Using as an example the US population of 250 million people, Diaconis and Mosteller observe that if a miracle "happens to one person in a million each day, then we expect 250 occurrences a day and close to 100,000 such occurrences a year" (1989: 859). If the human population of the earth is used as the reference class, miracles can be expected to be even more plentiful.

How should the law of truly large numbers be applied to Adams' double win? One possibility is to change our description of the observations from

(1) Evelyn Marie Adams, having bought four tickets in each of two New Jersey lotteries, wins both.

to the logically weaker statement that

(2) Someone at sometime, having bought some number of tickets in two or more lotteries in one or more states, wins at least two lotteries in a single state.

If you are using probabilistic *modus tollens* (§1.4) to think about this problem, and if you believe that Adams' double win does not warrant rejecting the hypothesis that the lotteries were fair, then weakening the data description from (1) to (2) may be appealing. It provides a simple strategy for neutralizing the appeal of conspiracy theories. But even if this strategy leads to the conclusion about Adams' good fortune that you find intuitive, it raises the question of when and how much a description of the data should be weakened. Without some guidance on this issue, you run the risk of weakening the data whenever they go against your pet theories. This allows you to be complacent about what you already believe and skeptical about the hobbyhorses that others have chosen to ride – a satisfying state of mind perhaps, but one that cannot stand up to rational scrutiny.

A second approach, which abides by the principle of total evidence (§1.4), is Bayesian. It concedes that the hypothesis that the lotteries were fair has a much lower likelihood than the hypothesis that the two lotteries that Adams won were rigged in her favor, but then invokes prior probabilities to show that Adams' double win does not make it *probable* that the two lotteries were rigged. My objection to invoking priors here is not that they are subjective. After all, we may have evidence that lotteries are usually fair, though developing this point would require us to consider the fact that people who rig lotteries have a powerful incentive to insure that their chicanery remains secret. Rather, my reservation about this Bayesian reply is that it concedes that the observations favor the hypothesis that the two lotteries were rigged in Adams' favor. The law of likelihood, which is central to Bayesianism, obliges Bayesians to make this concession. I suggest that it is possible to show that the observations do not have this evidential significance. The model-selection framework allows this kind of argument to be developed, although it must be recognized that the goal has been changed; we no longer are trying to figure out which hypothesis is probably true or which has the highest likelihood; rather, we are aiming to discover which will be most predictively accurate.

The model-selection approach agrees with Bayesianism that data cannot be discarded. Rather, the right approach is to *add* observations. Instead of weakening the observations by discarding (1) and focusing on (2), we should include additional observations about the people who won and lost other lotteries and how many tickets they purchased. Once the data set is augmented, we can consider multiple models. One of them says that each lottery is fair:

(Fair) For each ticket i purchased in New Jersey lottery j,
$Pr(\text{ticket } i \text{ wins} \mid \text{ticket } i \text{ was purchased in lottery } j) = \frac{1}{n_j}$
(where n_j is the number of tickets purchased in lottery j).

This model has one parameter for each lottery. It is far simpler than the following model:

(Rigged) For any ticket i purchased in New Jersey lottery j by person k, Pr(ticket i wins | ticket i was purchased in lottery j by person k) $= p_{jk}$.

The (Rigged) model has a separate parameter for each person buying a ticket in each lottery. If the data on lottery winners and losers favors (Fair) over (Rigged), they do so not by showing that (Fair) is more probable than (Rigged), nor by showing that (Fair) has the higher likelihood, but by showing that (Fair) can be expected to be more predictively accurate than (Rigged).

(Fair) is a model that *unifies* the data far more than (Rigged) does. (Fair) says that all the tickets sold in a given lottery are subject to the same probabilistic process, whereas (Rigged) says that each person buying tickets in a given lottery is a law unto herself. Because AIC and other model-selection criteria value paucity of parameters, they offer an explanation of why a model that applies k parameters to an entire data set often has a leg up on a disunified model that subdivides the data into parts, supplying a different set of k parameters to each.

It is important to realize that whether a more unified model has a better AIC score than a less unified model depends on the data. There is no categorical imperative that says that unified models are always better. For example, it is not inevitable that Fair is superior to the following even simpler model:

(One) For each ticket i purchased in any New Jersey lottery, Pr(ticket i wins | ticket i was purchased in any New Jersey lottery) $= p$.

The (One) model lumps together all New Jersey lotteries; tickets purchased in different lotteries are said to have the same chance of winning. This model is even more unified than (Fair), but that does not guarantee that its estimated predictive accuracy will be greater.

Although the models just considered exhibit a virtue of the model-selection framework, there is a model not yet mentioned that exhibits one of its limitations. The conspiracy model (Rigged) gets lower marks than the (Fair) model, but what about the following (Mixed) model?

(Mixed) For each ticket k purchased by Evelyn Marie Adams, Pr(ticket k wins | ticket k was purchased by Evelyn Marie Adams) $= p$. For each other ticket i purchased in New Jersey lottery j, Pr(ticket i wins | ticket i was purchased in lottery j) $= \frac{1}{n_j}$ (where n_j is the number of tickets purchased by people other than Adams in lottery j).

Suppose, to make things simple, that Evelyn Marie Adams bought tickets only on the two lotteries that she ended up winning and bought a few tickets on each. This means that L(Mixed) fits the data far better than L(Fair). And (Mixed) has just one more parameter than (Fair). This means that (Mixed) may have a better AIC score than (Fair). If so what's wrong with this mixed model? The Bayesian has an answer: It has a lower prior probability. It is not obvious what the model selectionist can say here.

This question aside, there is a point here on which defenders of different statistical frameworks can agree. The human mind often imposes patterns where none exist. Repeatedly tossing a fair coin will inevitably produce runs of heads; it is tempting to think that the coin has suddenly become biased ("hot"). Part of what facilitates this kind of over-interpretation is that we tend to focus on observations that are vivid. We narrow the data set. We focus on the run of heads, and not on all the tosses. It is Adams' double win that excites our curiosity, not a boring compilation of all the winners and losers in all New Jersey lotteries. In all these cases, we need to embed what we find vivid in a more inclusive data set; we then need to formulate models that apply not just to what is vivid but to what is quotidian as well.

1.9 CONCLUDING COMMENTS

The claim that science aims to discover which theories are probably true may sound like a truism, but there are two reasons to pause over this formula. The first is that one must be wary of an equivocation. In ordinary English, to say that a theory is "probably true" just means that it is plausible or reasonable, given the evidence at hand; praising a theory in this way leaves open what relevance the mathematical theory of probability might have to such judgments. Bayesianism is a substantive epistemology, not a truism. The second reason for pausing is that scientists often work with idealized models that are known to be false. How can a model known to be false probably be true? There needs to be a place in our epistemology for comparisons of such theories.

Royall's three questions (§1.1) are different; questions about *evidence* must be separated from questions about *acceptance* and from questions about *action*. This threefold distinction will be important in what follows when we consider evidential questions such as the following:

- Are the imperfect adaptations that organisms exhibit evidence that they were not produced by an intelligent designer?

- Is the fact that bears in cold climates have longer fur than bears in warm climates evidence that fur length evolved by natural selection as an adaptive response to ambient temperature?
- Are the similarities that species exhibit evidence that they stem from a common ancestor?

Perhaps you find it obvious that the answer in all three cases is *yes*. If so, what's the point of taking on the job of figuring out why? The answer is that the book you are reading is a work of philosophy, not biology, and so the exploration of what seems obvious is of central importance. Even when a proposition strikes us as obvious, it is often not so obvious why the proposition is *true*. This is the occasion for philosophical exploration. One possible result is that what seems obvious turns out *not* to be true unrestrictedly, but is true only in a restricted set of circumstances. Another is a deeper grasp of the assumptions we tacitly make that underlie our convictions.

The law of likelihood is common ground for Bayesians and likelihoodists. It will provide the starting point for several of the questions about evidence and evolution that I will examine. Putting the law to work in the next chapter will require us to consider a new complication. The hypotheses we wish to test often do not have likelihoods when considered all by themselves; they need to be supplemented by additional information if they are to confer probabilities on the observations. An important question will be how this "additional information" should be obtained. There also will be a place in what follows for ideas about evidence that derive from a model-selection framework. Just as the readings of an unbiased scale can provide evidence as to which of two people is heavier, so AIC scores can provide evidence as to which of two models is apt to be more predictively accurate. The law of likelihood is central to understanding what evidence is, but it is not the only idea we will use. The law applies to simple statistical hypotheses and produces a verdict about whether the observations favor the hypothesis that H_1 is true over the hypothesis that H_2 is true; AIC and other model-selection criteria apply to composite statistical models and help us discern which models will be more predictively accurate. The law of likelihood and AIC are not in conflict, given their different goals and their different realms of applicability.

CHAPTER 2

Intelligent design

2.1 DARWIN AND INTELLIGENT DESIGN

The first edition of Darwin's *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1859) begins with quotations from two philosophers:

> But with regard to the material world, we can at least go so far as this – we can perceive that events are brought about not by insulated interpositions of Divine power, exerted in each particular case, but by the establishment of general laws. (W. Whewell, *Bridgewater Treatise*)

> To conclude, therefore, let no man out of a weak conceit of sobriety, or an ill-spirited moderation, think or maintain, that a man can search too far or be too well studied in the book of God's word, or in the book of God's works; divinity or philosophy; but rather let men endeavour an endless progress or proficience in both. (F. Bacon, *Advancement of Learning*)

William Whewell was Darwin's contemporary and rejected his theory of evolution, a result that Darwin probably anticipated when he wrote *The Origin of Species*.[1] Francis Bacon wrote more than 200 years earlier. The two quotations are interesting because of what they reveal about Darwin's views on the relationship of belief in God and belief in evolution.

Bacon's remark harks back to an old distinction between the Bible (God's word) and nature (God's work). Sacred texts and natural phenomena provide separate pathways for learning about God. This two-pathway picture was important in the formation of the Royal Society in

[1] The *Bridgewater Treatises* were a series of books that developed the argument for the existence of God that we will consider in detail in this chapter – the argument from design. In the 1833 book from which Darwin drew this quotation, Whewell embraced the view that the origin of species and the origin of languages are beyond the reach of present-day science and are likely to remain so; he argued that both require divine intervention. Darwin's quoting from Whewell does not mean that he expected Whewell to like how he used this passage. See Ruse (1979), Hodge (1991), Brooke (2003), and Snyder (2006) for different views of Darwin's relation to Whewell.