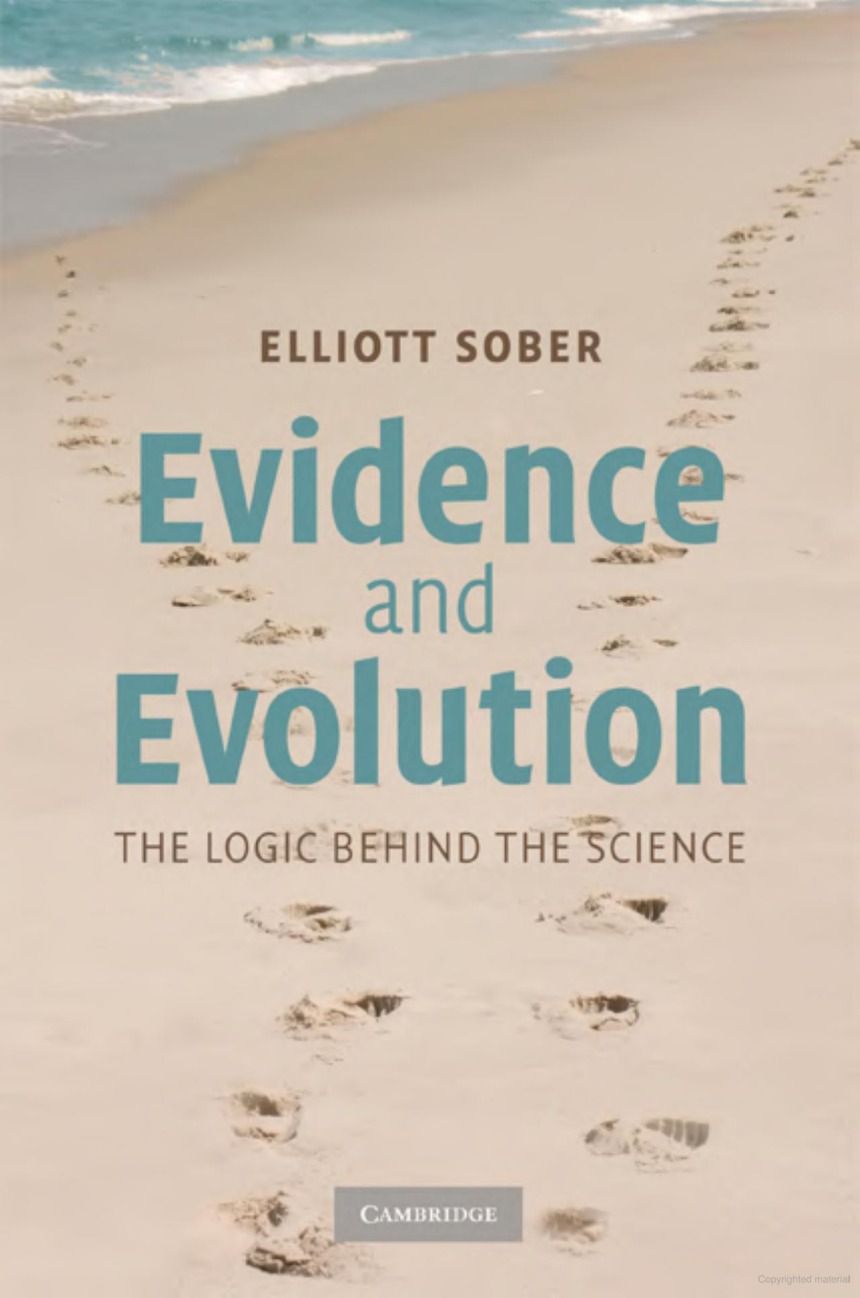

EVIDENCE AND EVOLUTION

The logic behind the science

ELLIOTT SOBER

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521871884

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

Sober, Elliott.
Evidence and evolution : the logic behind the science / Elliott Sober.
p. cm.
Includes bibliographical references.
ISBN 978-0-521-87188-4 (hardback : alk. paper) – ISBN 978-0-521-69274-8 (pbk. : alk. paper)
1. Evolution (Biology)–Philosophy. 2. Natural selection–Philosophy. 3. Evidence.
4. Probabilities–Philosophy. I. Title.
QH360.5.S625 2008
576.8–dc22
2007051438

ISBN 978-0-521-87188-4 hardback
ISBN 978-0-521-69274-8 paperback

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?

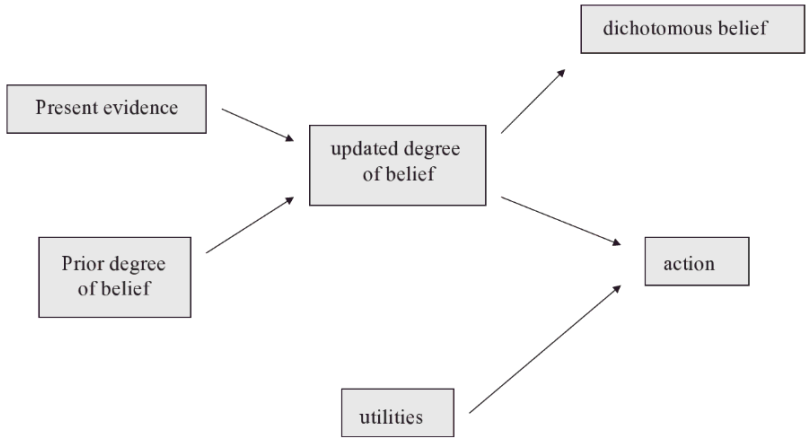

Figure 1.1 Present evidence and its downstream consequences.

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).

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 $Pr(\text{heart \& red})/Pr(\text{red})$. 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 *H*s and *O*s 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 $Pr(S \text{ has tuberculosis} \mid \text{the test is positive})$ and $Pr(S \text{ has tuberculosis} \mid \text{the test is negative})$. 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.

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

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

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:

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

[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]

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

$\mathrm{E}(x \mid y)$ 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.

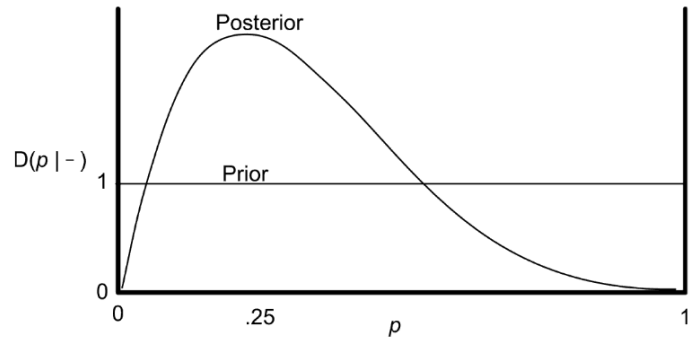

Figure 1.3 A flat prior density distribution for p and the non-flat posterior density occasioned by observing one head in four tosses. The prior expected value of p is 0.5; given this prior, the posterior expected value of p is 0.33.

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 $p = Pr(\text{the coin lands heads} \mid \text{the coin is tossed})$ 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

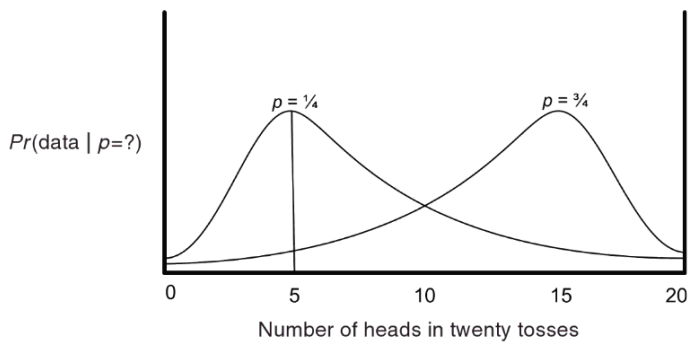

Figure 1.4 When the coin lands heads in five of twenty tosses, the maximum likelihood estimate of $p = Pr(\text{the coin lands heads} \mid \text{the coin is tossed})$ is $p = \frac{1}{4}$. The likelihood of the estimate $p = \frac{3}{4}$ is lower.

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 $Pr(+ \text{ result} \mid \text{tuberculosis})$ and $Pr(+ \text{ result} \mid \text{no tuberculosis})$. 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

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(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(+ test result), since this quantity is defined as Pr(+ test result | tuberculosis)Pr(tuberculosis) + Pr(+ test result | no tuberculosis)Pr(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) $\approx$ 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

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

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 $Pr(E|H) > Pr(E|notH)$. 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 $Pr(H|E) > Pr(H)$.

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, ... O_n, and H_1, H_2, ... 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 $Pr(O \mid not H_1)$, 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 $Pr(O \mid not H_1)$, 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.

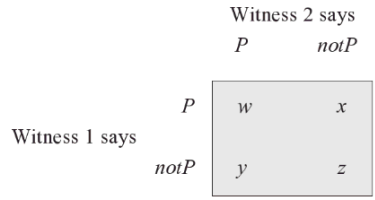

Figure 1.5 When two independent and reliable witnesses each report on whether proposition P is true, yeses provide stronger evidence for P than one, and one yes provides stronger evidence than zero. Each cell represents the likelihood ratio $Pr(\text{testimony} \mid P)/Pr(\text{testimony} \mid notP)$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

witnesses who agree that P is true provide stronger evidence in favor of P than either witness does alone.[23]

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.

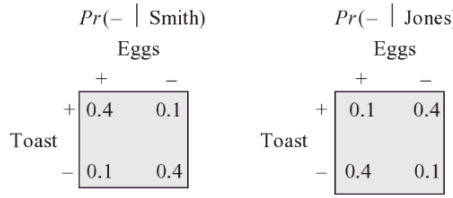

Figure 1.6 Smith and Jones differ in their inclinations to place different orders for breakfast. The breakfast order of toast and eggs provides evidence about which of them placed the order, although the fact that the order included toast does not, and neither does the fact that the order included eggs.

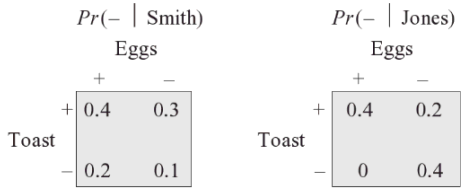

Figure 1.7 A new set of breakfast inclinations for Smith and Jones. Now the breakfast order of toast and eggs provides no evidence about which of them placed the order, though each part of the order favors Smith over Jones.

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 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 $w_A = w_a$,

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

(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)

$$\frac{\begin{array}{l} Pr(O \mid H) \text{ is very high.} \\ notO \end{array}}{notO \text{ is evidence against } H.}$$

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 probability of a disjunction in

[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.

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:

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

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 strong*er* 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

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 no

[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.

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

		Possible states of the world	
		$H = S$ has tuberculosis	S does not
Possible decisions	reject H	e_1	$1 - e_2$
	accept H	$1 - e_1$	e_2

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).