2 The probabilistic turn

Discussion of parsimony took a probabilistic turn in the twentieth century.[1] The project was to use probability theory to analyze and justify Ockham's razor. Not all of these efforts succeeded, but two of them did. I think there are two "parsimony paradigms" in which probability ideas show that parsimony is epistemically relevant. The two paradigms were developed within two different philosophical frameworks for understanding probability; one paradigm finds its home in Bayesianism, the other in frequentism. To set the stage for investigating probabilistic approaches to Ockham's razor, I'll start this chapter by providing a brief (and I hope accessible) primer on probability. But first I want to say a little about Bayesianism and frequentism.

Two philosophies of probability

Bayesianism is a philosophy of inference that traces back to a mathematical result (a theorem) obtained by Thomas Bayes (1701–1761). Bayes's (1764) theorem describes how the probability you assign to a hypothesis should be influenced by the new evidence you acquire. Bayesianism is now a general philosophy of scientific reasoning that has grown richer and more detailed than its eighteenth-century beginnings. This philosophy says that scientific reasoning has the attainable goal of figuring out how probable different scientific hypotheses are, given the evidence at hand. Or more modestly, it maintains that science is in the business of figuring out which hypotheses are more probable than which others, again in the light of the evidence. Either way, science crucially involves thinking about the probabilities of hypotheses.

[1] I borrow this phrase from Richard Rorty's influential anthology of 1967, *The Linguistic Turn*, which documented the emphasis on language as a philosophical subject in the previous eighty years. Rorty got the expression from Gustav Bergmann (1906–1987).

Bayesianism was not the dominant philosophy of probabilistic inference that scientists themselves embraced in the twentieth century. Rather, the dominant mode of thought was frequentism. Frequentism does not have the simple unity that Bayesianism exhibits; rather, it is a varied collection of ideas about how observations should be used to evaluate hypotheses. Frequentism uses probability ideas in this enterprise just as Bayesianism does, but its basic idea is different. The first commandment of frequentism is: *thou shalt not talk about the probabilities that hypotheses have!* The claim that science has the job of assessing how probable different theories are may sound like an unremarkable truism, but this innocent-sounding remark is something that frequentists categorically reject.

The difference between frequentism and Bayesianism is often characterized in terms of what each philosophy takes the concept of probability to mean. The standard picture is that Bayesians think that probability means rational degree of certainty whereas frequentists define probability in terms of frequency. When you think about your probability of getting lung cancer, given that you smoked lots of cigarettes over many years, Bayesians take this probability to represent how confident you should be that you'll get cancer, given your history of smoking, whereas frequentists take the probability to represent how frequently heavy smokers get lung cancer. Viewed in this way, Bayesianism is about something subjective (= in the mind of a rational subject) and frequentism is about something objective (= out there in the external world). If the two philosophies have different subject matters, why is there conflict between them?[2] Why can't these partisan schools see that probability has both a subjective and an objective meaning (as Carnap 1950 recognized) with each *ism* going its own way? Why can't people just get along? The answer is that Bayesianism and frequentism fundamentally disagree about what the goals of science ought to be. There is more to the debate than a question about the meaning of the word "probability." But even the idea that each school is wedded to a single interpretation of probability is too simple.

On the one hand, there are situations in which Bayesian inferences can be carried out by using probabilities that are as objective as any frequentist could wish. If I tell you what the frequency is of tuberculosis in Wisconsin, that Susan

[2] There is another usage of this terminology, as when people claim that various norms are objective. Here the thought is that the norms are correct and non-arbitrary. Many Bayesians are objectivists in this sense.

lives in that state, and that her tuberculosis test came out positive (where the test procedure produces erroneous results with a certain frequency), you can calculate the probability that Susan has tuberculosis, given her test result. We'll see in a moment how Bayesians do this calculation. The present point is that the probabilities used in this Bayesian calculation are all about objective matters of fact. Bayesians can go to work on frequencies!

On the other hand, there are good reasons why the probabilities that frequentists discuss often should not be interpreted as frequencies. Frequentists are happy to talk about the probability that a fair coin has of landing heads if it is tossed. Fairness means that the value of this probability is ½. But fair coins often fail to have frequencies that match this probability. For example, suppose you toss a fair coin three times and then destroy it. The frequency of heads in the short lifetime of this coin will not equal 50 percent. For this simple reason, you can't equate probability with actual frequency. You may reply that the relevant frequency idea is hypothetical long-run frequency. Although a fair coin won't land heads 50 percent of the time if it is tossed just once, the suggestion is that if a coin is fair, then the frequency of heads will converge on 50 percent if you toss the coin again and again. What's wrong with that? Let us consider what "converge" means. Here is one interpretation:

> A coin has a probability of landing heads of ½ precisely when the frequency of heads will get closer and closer to 50 percent as the coin is tossed repeatedly.

This is false. It is possible for a fair coin to produce two heads in the first four tosses and three heads in the first five. There need be no lockstep, monotonic approach to 50 percent. We can replace this flawed suggestion with something that is true. Consider any small positive number you please; call it ε ("epsilon").

> A coin has a probability of landing heads of ½ precisely when the probability approaches 1 that the frequency of heads will be within ε of 50 percent as the number of tosses approaches infinity.

This is one version of the *law of large numbers*. Notice that the concept of probability appears on both sides of this biconditional. This is not a proper definition; it is circular. For this reason, the law of large numbers, though true, does not provide an *interpretation* of probability in the required sense.[3]

[3] Here's a third suggestion for defining probability in terms of frequency: a coin has a probability of landing heads of ½ precisely when the coin would have to land heads

Despite its name, frequentism as a philosophy of scientific inference has no commitment to interpreting probability in terms of the idea of frequency – either actual or hypothetical.

Although defining Bayesianism and frequentism in terms of their different interpretations of probability is too simple, it does contain an ounce of truth. Bayesians *often* equate probability with rational degree of certainty and frequentists *always* want probability to be more objective than this. But the heart of the matter is that the two philosophies propose different epistemologies, not different semantics. Frequentists want assignments of values to probabilities to have an "objective justification." It should be possible to defend one's assignments by citing frequency data or an empirically justified theory, for example. It isn't good enough to say "well, my probability assignment simply reflects how certain I am in the proposition in question." When I talk about objectivity in what follows, I have in mind this epistemic usage.

A probability primer and the basics of Bayesianism

Before discussing the partisan worlds of Bayesianism and frequentism, I'll begin with the mathematical core of the probability concept itself. This is something on which Bayesians and frequentists agree.

Probability assignments always rest on assumptions. For example, if you assume that the deck of cards before you is standard and that the dealer is dealing you cards "at random," you can conclude that the probability that the first card you are dealt will be an ace of spades is $\frac{1}{52}$ and that the probability that the first card you receive is either an ace or a jack is $\frac{8}{52}$. Without the assumptions mentioned, these probability assignments can be incorrect. I will make the role of assumptions explicit in my description of probability by adding a subscript "A" to the canonical axioms of probability theory described

50 percent of the time if it were tossed an infinite number of times. Although this biconditional is not circular, there still is a problem. It is not impossible for a fair coin to land heads each time it is tossed, even if it is tossed an infinite number of times. True, the probability of the infinite sequence HHHH . . . is zero. However, you can't equate impossibility with a probability of zero. The probability of *any* infinite sequence (including the alternating sequence HTHTHT . . .) is zero if the coin is fair.

by Kolmogorov (1950):

$0 \leq \Pr_A(H) \leq 1$.

$\Pr_A(H) = 1$ if A logically entails H.

$\Pr_A(H \text{ or } J)$

$= \Pr_A(H) + \Pr_A(J)$ if A logically entails that H and J are incompatible.

$\Pr_A(H)$ represents the probability of the proposition H under the assumptions codified in the propositions A. Applying probability to a problem involves isolating a class of propositions that are to be evaluated. In the card example, the propositions concern the different cards you may be dealt, not whether it will rain tomorrow. Notice that probability in the above axioms is a mathematical *function*: it maps propositions onto numbers. Two different probability functions may assign different numbers to the same proposition. The model I just described says that the deck is standard and that cards are dealt at random, with the result that $\Pr_A$(the first card you are dealt will be an ace of spades) $= \frac{1}{52}$. If we thought the deck was made of fifty-two such aces, we would use a different probability function, $\Pr_B(-)$ according to which $\Pr_B$(the first card you are dealt will be an ace of spades) $= 1$.

Here are three consequences of the axioms just stated that do not depend on what assumptions go into A: (i) Tautologies have a probability of 1 and contradictions have a probability of 0; (ii) If propositions H and J are logically equivalent, then $\Pr_A(H) = \Pr_A(J)$; (iii) $\Pr_A(H) = \Pr_A(H\,\&J) + \Pr_A(H\,\&not J)$. This last equality follows from (ii) and the third axiom; it is called *the theorem of total probability*.

The third axiom describes how the probability of a disjunction is settled by the probabilities of the disjuncts if the disjuncts are incompatible with each other. But what if the disjuncts are not mutually exclusive? There is a general principle available here that you can visualize by thinking about probabilities in terms of the diagrams that John Venn (1834–1923) invented. Figure 2.1 shows a square in which each side has a length of one unit. Let's suppose that each point in the square represents a possible way the world might be. Each proposition that we might want to talk about can be associated with a set of points in the square – the set of possible situations in which the proposition is true. The area of the square is 1, which conveniently is also the maximum value that a probability can have. Tautologies are true in

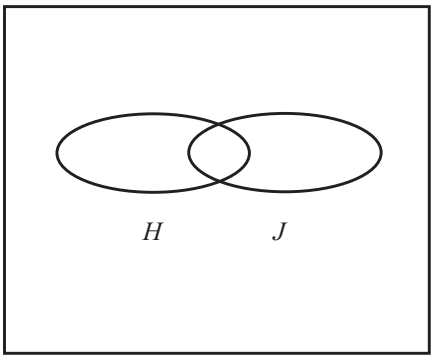

Figure 2.1

all possible situations; they fill the whole unit square. The figure represents propositions H and J as two ovals. The intersection of the two ovals – their area of overlap – represents the conjunction $H\&J$. Since there is a region of overlap, the two propositions are compatible with each other; there are situations in which both are true. I hope the Venn diagram makes it obvious that

$$\Pr(H \text{ or } J) = \Pr(H) + \Pr(J) - \Pr(H\&J).$$

The reason for subtracting $\Pr(H\&J)$ is to insure that the area of overlap is not double-counted. When $\Pr(H\&J) = 0$, the above equality reduces to the special case described in Axiom 3.

What can be said about the probability of conjunctions? This is where we need to define the concept of *probabilistic independence*:

> Propositions H and J are probabilistically independent in probability model A precisely when $\Pr_A(H\&J) = \Pr_A(H) \times \Pr_A(J)$.

When you flip a fair coin twice, the probability of getting a head on the first toss is $\frac{1}{2}$ and the probability of getting a head on the second is also $\frac{1}{2}$. The tosses are probabilistically independent; the probability of getting heads on both tosses is $\frac{1}{4}$. That is a contingent empirical fact about coin tossing; it is logically possible for tosses to be probabilistically dependent. Suppose we lived in a world in which there are two kinds of coins: 50 percent of the coins have two heads and 50 percent have two tails. You select a coin at random and toss it repeatedly. Under the assumptions stated, $\Pr_A$(Heads on the first toss) = $\Pr_A$(Heads on the second toss) = $\frac{1}{2}$. However, it's also true that

$\Pr_A$(heads on both the first and second tosses) $= \frac{1}{2}$. Independence fails. In this fanciful world, knowing the outcome on the first toss would give you information about what will happen on the second. In the real world, the tosses are independent; knowing the outcome of the first toss doesn't change the probability you assign to the second.

Probabilistic independence and logical independence are different. Propositions X and Y are logically independent precisely when all four conjunctions of the form $\pm X \& \pm Y$ are logically possible (i.e., non-contradictory). For example, "it is raining" and "you are carrying an umbrella" are logically independent of each other. However, if you follow the advice of accurate weather forecasts, these two propositions will be probabilistically dependent on each other. Consider any two propositions that are neither tautologies nor contradictions: if they are probabilistically independent, then they are logically independent, but the converse implication does not hold.

		color of boat on Tuesday		
		green (p = 0.2)	red (p = 0.3)	blue (p = 0.5)
color of boat on Monday	green (p = 0.2)			
	red (p = 0.3)			
	blue (p = 0.5)			

Here's a little exercise that involves thinking about how the probability of a conjunction is related to the probability of its conjuncts. It involves the example about sailboats mentioned in the previous chapter in the section on Copernicus and Ptolemy. My friend Susan saw a red sailboat on Lake Mendota on Monday, and on Tuesday she also saw a red sailboat. In the accompanying table I've listed probabilities for some sailboat colors on each of the two days. Note that the three probabilities for each day sum to one; I'm assuming that sailboats on the lake have no chance of being yellow. These probabilities are called *marginal probabilities* because they are written along the margins of the table. Now consider these hypotheses:

(ONE) Susan saw the same boat on both days.

(TWO) Susan saw one boat on Monday and a different boat on Tuesday.

The cells in the table represent conjunctions. For example, the cell in the upper right-hand corner represents the possibility that the sailboat seen on the first day is green *and* the one sighted on the second is blue. What probabilities does the TWO hypotheses dictate for the cells? What cell entries does ONE say are correct? Assume in both cases that sailboats don't change color from day to day. How does the concept of probabilistic dependence apply to what the two hypotheses say?

The truth value of a conjunction $H\,\&J$ is determined by the truth value of H and the truth value of J. The conjunction is true if H is true and J is true, and it is false otherwise. This is what logicians mean when they say that conjunction is a "truth-functional operator." We have just seen that the probability of the conjunction $H\,\&J$ isn't settled by the probability of H and the probability of J. If anything, it is the probabilities of conjunctions that settle the probability of a conjunct. Here I have in mind a fact I mentioned earlier, the theorem of total probability, which says that $\Pr(H) = \Pr(H\,\&J) + \Pr(H\,\&not\,J)$.

Another concept that will be useful in what follows is *mathematical expectation*. You've encountered this before when you've heard discussion of the "life expectancy" of a baby born this year. As a first pass, this quantity can be understood as an average. If you say that the life expectancy for a baby born this year in the United States is 80 years, this means that 80 years will be the average lifespan of the individuals born this year. Let's get more precise by talking about probabilities and coin tosses. If you toss a fair coin ten times, there are eleven possible outcomes (0 heads, 1 head, 2 heads, . . . , 10 heads) and each of these has its own probability. The expected number of heads is defined as follows:

$$\begin{aligned}
&\text{Expected}_{\text{A}}(\text{number of heads}) \\
&\quad = (0)\Pr_{\text{A}}(\text{exactly 0 heads}) + (1)\Pr_{\text{A}}(\text{exactly 1 head}) \\
&\qquad + (2)\Pr_{\text{A}}(\text{exactly 2 heads}) + \cdots + (10)\Pr_{\text{A}}(\text{exactly 10 heads}) \\
&\quad = \sum_{i=0}^{10} (i)\Pr_{\text{A}}(\text{exactly } i \text{ heads}).
\end{aligned}$$

Here A is the assumption that the coin is fair and you toss the coin ten times. It turns out that the expected value is 5. As you do this ten-toss experiment again and again, you can be more and more certain that the average number of heads across the different ten-toss repetitions is close to 5. This is the law of large numbers I mentioned earlier.

The expected number is often not the number you should expect. If you toss a fair coin three times, the expected number of heads is 1.5, but this doesn't mean that you should expect there to be 1.5 heads when you perform this experiment just once. In the experiment I described five paragraphs ago concerning a world in which all coins either have two heads or two tails, what is the expected frequency of heads if you toss a randomly chosen coin ten times? What is the frequency you should expect?

Although the axioms of probability that I have described always involve a relation between the assumptions that define the probability function and this or that proposition, I have yet to define the idea of "conditional probability." I have been talking about $\Pr_A(H)$, not about $\Pr_A(H \mid E)$. The latter is read as "the probability of H given E." Take care to understand what this means. It doesn't mean that E is true and that H therefore has a certain probability. Just as "if you toss the coin then it will land heads" does not assert that you toss the coin, so "$\Pr_A$(the coin lands heads | you toss the coin) $= \frac{1}{2}$" does not say that you actually toss the coin. What it means is this: suppose for the moment that you have tossed the coin. You then are asked how probable it is that the coin will land heads, given that supposition. The value of the conditional probability is the answer to this question.

The concept of conditional probability can be introduced by saying how it is related to the notion of unconditional probability that is defined by our axioms:

$$\Pr_A(H \mid E) = \frac{\Pr_A(H \,\&\, E)}{\Pr_A(E)} \text{ if } \Pr_A(E) > 0.$$

This is called the *ratio formula*. If A says that E has a probability of zero, this "definition" of conditional probability offers no advice on what conditional probability means. I put "definition" in scare quotes because a (full) definition should provide necessary and sufficient conditions; the above statement provides only the latter. Some think that the conditional probability $\Pr_A(H \mid E)$ has no meaning when $\Pr_A(E) = 0$. I disagree. A coin can be fair even if you lock it in an impregnable safe so that the coin can never be tossed. Here $\Pr_A$(the coin lands heads | you toss the coin) $= \frac{1}{2}$ even though $\Pr_A$(you toss the coin) $=$ 0 (Rényi 1970; Hájek 2003; Sober 2008b). There is a second qualification that needs to be registered in connection with the ratio formula, which I'll discuss later. But for now it's worth noting that if $\Pr_A(H \mid E)$, $\Pr_A(H \,\&\, E)$, and $\Pr_A(E)$ all have values and $\Pr_A(E) > 0$, then the ratio formula must hold.

To illustrate the idea of conditional probability, let's return to the example of the deck of cards. As before, I'll assume that the deck is standard and that you are dealt cards at random. What is the value of $\Pr_A$(the card you were just dealt is a heart | the card you were just dealt is red)? On the supposition that the card is red, the probability of its being a heart is $\frac{1}{2}$. The ratio formula delivers this result. Here's the argument:

$$\begin{array}{l} \Pr(\text{the card is a heart and the card is red}) = \frac{1}{4}. \\ \Pr(\text{the card is red}) = \frac{1}{2}. \\ \hline \text{Therefore, } \Pr(\text{the card is Heart} \mid \text{the card is red}) = \frac{1}{2}. \end{array}$$

As an exercise, I suggest that you draw a Venn diagram of this example. You need to have an area of the diagram representing "the card is red" and another area representing "the card is a heart." And of course you need to consider the intersection of these two areas, which represents the conjunction of the two propositions. When you consider the conditional probability, you focus on the area of the diagram in which the card is red and determine what proportion of that area is occupied by the card's being a heart. Notice how hard this would be if the card had a probability of zero of being red!

I have used the word "assumption" to describe probability functions and the word "supposition" to describe conditional probabilities. These two terms may sound like synonyms but I am using them to pick out different things. In the example just described, I assumed that the deck is standard and that cards are dealt at random. I did not assign probabilities to those assumptions. With those assumptions in place, I asked you to consider the conditional probability $\Pr_A$(the card you were just dealt is a heart | the card you were just dealt is red), which requires you to consider the supposition that the card is red. Assumptions define probability functions whereas suppositions come up within a given probability function when a conditional probability is being evaluated. We often use models that we believe are true and we often entertain suppositions that we think are false. I believe that the deck is standard and that the cards are dealt at random. In contrast, I do not believe that the card you were just dealt is red, though I wish to entertain that supposition in evaluating a conditional probability. Now that I have separated assumptions from suppositions, let me bring them back together. There is a numerical identity that connects the assumptions that a probability

model makes and the suppositions that one entertains within a model. It is this:

$$\Pr_{A\&B}(H) = \Pr_A(H \mid B).$$

The values are the same, but the epistemic status of B is subtly different.

The "definition" of conditional probability makes it clear why $\Pr_A(H\&J) \leq \Pr_A(J)$, no matter what A is. Assuming that $\Pr_A(J) > 0$, the inequality can be rewritten as $\Pr_A(H \mid J)\Pr_A(J) \leq \Pr_A(J)$. Since probabilities are numbers between 0 and 1, the product of two probabilities cannot be greater than the value of either of them. This fact is relevant to the razor of silence that I discussed in the previous chapter. If you consider a conjunction $H\&J$ and slice away H (not by denying that H is true, but simply by declining to assert or deny it), the probability of what remains (J) cannot be less than the probability of the conjunction with which you began. In fact, if $\Pr_A(H)$ and $\Pr_A(J)$ are both positive and $\Pr_A(H \mid J)$ is less than 1, the slicing away will increase probability. Silence reduces your risk of error. The razor of silence has a simple Bayesian rationale.

The fact that a conjunction can't be more probable than its conjuncts is anything but obvious to many people. In a much-cited psychology experiment, Tversky and Kahneman (1982) told their subjects the following story:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

The subjects then were asked which of the following statements is more probable:

- Linda is a bank teller.
- Linda is a bank teller and is active in the feminist movement.

Well over half of the subjects in the experiment said that the second statement is more probable than the first. This example is a warning: when you use the mathematical concept of probability, don't stumble into the mistake of committing "the conjunction fallacy"!

We now can use the ratio formula to derive Bayes's theorem. To simplify notation, I'll drop the subscript "A," but don't forget that a probability

function is always built on a set of assumptions. Whenever I talk about a conditional probability $\Pr(X \mid Y)$, I'll assume that $\Pr(Y) > 0$. So let's start by describing each of $\Pr(H \mid E)$ and $\Pr(E \mid H)$ in terms of ratios of unconditional probabilities:

$$\Pr(H \mid E) = \frac{\Pr(H \,\&\, E)}{\Pr(E)} \qquad \Pr(E \mid H) = \frac{\Pr(E \,\&\, H)}{\Pr(H)}$$

These two equations can be rearranged to yield:

$$\Pr(H \,\&\, E) = \Pr(H \mid E)\Pr(E) \qquad \Pr(E \,\&\, H) = \Pr(E \mid H)\Pr(H)$$

The left-hand sides of these two equations are equal, since $H\&E$ is logically equivalent to $E\&H$, which means that the right-hand sides must be equal to each other. Setting the right-hand sides of these equations equal and performing a little algebra yields Bayes's theorem:

$$\Pr(H \mid E) = \frac{\Pr(E \mid H)\Pr(H)}{\Pr(E)}.$$

Although the derivation of Bayes's theorem works for any propositions H and E you please, the typical application involves H's being a "hypothesis" and E's being "observational evidence." Bayes's theorem shows how the conditional probability $\Pr(H \mid E)$ can be computed from three other quantities. Notice that one of them is the unconditional probability $\Pr(H)$.

At the start of this chapter, I said that Bayesianism and frequentism are different philosophies of scientific inference. Does accepting Bayes's theorem place you knee deep in the former philosophy? Not so! The theorem is a mathematical truth – it follows from the axioms of probability and the "definition" of conditional probability. Frequentists do not challenge the correctness of this derivation. Rather, they challenge its *usefulness*. Frequentists think that it often isn't possible to think about $\Pr(E \mid H)$, $\Pr(H)$, and $\Pr(E)$ as objective quantities. However, they agree that *if* these three probabilities had objective values, the value of $\Pr(H \mid E)$ could be calculated by using Bayes's theorem. They also agree that if lions could fly, then zebras would need to watch out for aerial lion attacks.

In thinking about the probabilities that figure in Bayes's theorem, it is important to recognize that $\Pr(H \mid E)$ and $\Pr(E \mid H)$ are different quantities and therefore may have different values. Much heartache will be avoided by attending to this difference! In logic it is a familiar idea that a conditional

and its converse are different, and that one can be true while the other is false. For example, consider

- If noisy gremlins are bowling in your attic, then you hear noise coming from your attic.
- If you hear noise coming from your attic, then noisy gremlins are bowling in your attic.

It is obvious that the first can be true while the second is false. In just the same way the following two conditional probabilities can have different values:

- Pr(you hear noise coming from your attic | noisy gremlins are bowling in your attic)
- Pr(noisy gremlins are bowling in your attic | you hear noise coming from your attic)

Personally, I think that the first probability has a high value and the second has a low one.

The quantity $\Pr(E)$ on the right-hand side of Bayes's theorem deserves a comment. $\Pr(E)$ is the unconditional probability of the evidence E. In our gremlin example, E is the proposition that you hear noises coming from your attic. You might think that $\Pr(E)$ should be high if there frequently are noises coming from up there and that it should be low if such noises are rare. I agree that frequencies often provide *evidence* that is relevant to estimating the value of $\Pr(E)$. But, as noted earlier, I don't want to *define* probability as frequency. So what does the unconditional probability of E mean? The theorem of total probability tells us that

$$\Pr(E) = \Pr(E\&H) + \Pr(E\,\&notH).$$

Using the "definition" of conditional probability (and assuming that all the relevant probabilities are positive), we can rewrite this as

$$\Pr(E) = \Pr(E \mid H)\Pr(H) + \Pr(E \mid notH)\Pr(notH).$$

This characterization of $\Pr(E)$ shows that the value of this quantity will sometimes be very different from the frequency with which E is true. The example of the world in which half the coins have two heads and half have two tails provides an example. You choose a coin at random and toss it repeatedly. Use the above equality to convince yourself that $\Pr(\text{Heads}) = 0.5$. Yet, when

you do the experiment, you obtain either 100 percent heads or 100 percent tails.[4]

When I explained earlier what probabilistic independence means, I did so by describing a relation among unconditional probabilities. Now that the idea of conditional probability has been introduced, I can define the idea of *conditional* independence. It parallels the unconditional concept already explained:

> X and Y are probabilistically independent of each other conditional on C in probability function $\Pr_A(-)$ if and only if $\Pr_A(X \& Y \mid C) = \Pr_A(X \mid C)\Pr_A(Y \mid C)$.

Here's an example from Mendelian genetics: the genotypes of two full siblings are independent of each other, conditional on the genotypes of their parents. For example,

> $\Pr_M$(sib 1 has *AA* & sib 2 has *AA* | mom has *AA* & dad has *Aa*) = $\Pr_M$(sib 1 has *AA* | mom has *AA* & dad has *Aa*) $\Pr_M$(sib 2 has *AA* | mom has *AA* & dad has *Aa*).

The "M" subscript on the probability function indicates that probabilities are assigned on the basis of the usual Mendelian model of inheritance. The probability on the left has a value of ¼ and the two probabilities on the right each have a value of ½. Notice that the above equality holds more generally; it holds for *any* genotypes that the two siblings, mom, and dad might have:

> For any genotypes G_1, G_2, G_3, and G_4, $\Pr_M$(sib 1 has G_1 & sib 2 has G_2 | mom has G_3 & dad has G_4) = $\Pr_M$(sib 1 has G_1 | mom has G_3 & dad has G_4) $\Pr_M$(sib 2 has G_2 | mom has G_3 & dad has G_4).

When this more general relation obtains, the parental genotype is said to *screen-off* each offspring genotype from the other. We can generalize from this genetics example and define screening-off as a relation that might obtain among any three variables X, Y, and Z:

[4] In this example, the unconditional probability of the evidence involves two possibilities – either H is true or *notH* is. But suppose there are n possible hypotheses $H_1, H_2, \ldots, H_n$, which are mutually exclusive and collectively exhaustive. What would $\Pr(E)$ mean in that case? The answer is a generalization of what I just said for the dichotomous case:

$$\begin{aligned}\Pr(E) &= \Pr(E \mid H_1)\Pr(H_1) + \Pr(E \mid H_2)\Pr(H_2) + \cdots + \Pr(E \mid H_n)\Pr(H_n) \\ &= \sum_{i=1}^{n} \Pr(E \mid H_i)\Pr(H_i).\end{aligned}$$

Z screens-off X from Y precisely when, for any values i, j, k,

$$\Pr(X = i \,\&\, Y = j \mid Z = k) = \Pr(X = i \mid Z = k)\Pr(Y = j \mid Z = k).$$

Here I'm using a notation that is standard in probability theory; "$X = i$" means that the variable X has the value i. As the genetics example suggests, Z screens-off Y from X precisely when Z screens-off X from Y. There is another, equivalent, definition of screening-off that you should know about. It says that Z screens-off Y from X if and only if $\Pr(X = i \mid Z = k) = \Pr(X \mid Z = k \,\&\, Y = j)$, for all i, j, k. You can prove the equivalence of these two ways of describing screening-off by using the "definition" of conditional probability (and assuming that all the conditioning propositions have positive probabilities). Just as parental genotype screens-off offspring genotypes from each other, it's also true that parental genotype screens-off grandparental genotype from offspring genotype. Screening-off often applies when a common cause has two (or more) effects, and it often comes up when you talk about a causal chain from Y to Z to X. "Often" does not mean *always*. Mom's genotype is a common cause of the genotypes of her two offspring. However, her genotype does not screen-off each from the other. See if you can figure out why this is so. And see if you can think of an example of a causal chain in which the proximate cause doesn't screen-off the distal cause from the effect.

Screening-off can be described informally in informational terms. If you know the parental genotype, the probability you assign to one offspring's genotype shouldn't be affected by learning the genotype of the other. And if you know the parental genotype, the probability you assign to an offspring's genotype shouldn't be affected by your learning the genotypes of the grandparents.

Conditional independence and unconditional dependence may sound like they are incompatible, but in fact they are not. Once again, genetics furnishes an example. As noted, the two offspring genotypes are independent of each other, conditional on the parental genotype. However, the fact that the two siblings have the same parents means that their genotypes will be unconditionally dependent:

$$\Pr(\text{sib 1 has } AA \,\&\, \text{sib 2 has } AA) > \Pr(\text{sib 1 has } AA)\Pr(\text{sib 2 has } AA).$$

Notice that this inequality makes no mention of what the parental genotype is. If you conditionalize on the parental genotype, the inequality turns

into an equality! Not only are unconditional dependence and conditional independence not in conflict; I'll explain later in this chapter how conditional independence can be part of the explanation of unconditional dependence.

Translating the last displayed inequality into the language of conditional probability, we get both of the following:

$$\Pr(\text{sib 1 has } AA \mid \text{sib 2 has } AA) > \Pr(\text{sib 1 has } AA)$$
$$\Pr(\text{sib 2 has } AA \mid \text{sib 1 has } AA) > \Pr(\text{sib 2 has } AA).$$

This is what it means for the two genotypes to be *correlated*. Bayesians take these two inequalities to have an important epistemological significance. They gloss these inequalities by saying that the *AA* genotype of each sibling provides *confirmation* that the other sibling has the *AA* genotype. Bayesians define confirmation as follows:

Observation E confirms hypothesis H if and only if $\Pr(H \mid E) > \Pr(H)$.[5]

Disconfirmation gets defined in tandem:

Observation E disconfirms hypothesis H if and only if $\Pr(H \mid E) < \Pr(H)$.

If confirmation means probability raising and disconfirmation means probability lowering, then *evidential irrelevance* means that the observation leaves the probability of the hypothesis unchanged.[6] The Bayesian ideas of confirmation and disconfirmation entail that there is a symmetry between confirmation and disconfirmation:

E confirms H if and only if *notE* disconfirms H.

Convince yourself that this biconditional is correct when confirmation and disconfirmation are given Bayesian interpretations. And then convince

[5] Following Carnap (1950), Bayesians sometimes contrast the "incremental" concept of confirmation just described with one that is "absolute." The idea is that E absolutely confirms H precisely when $\Pr(H \mid E)$ is high. Notice that this can be true when E incrementally disconfirms H or is evidentially irrelevant to it in the incremental sense. I think it is unfortunate that the word "confirm" is used to denote a high value for $\Pr(H \mid E)$. I won't do so in what follows.

[6] In view of the fact that assigning a value to $\Pr(H \mid E)$ does not require that E be true, it is better to read the Bayesian definition of confirmation as explicating the following proposition: E, if true, would confirm H. A parallel point holds for disconfirmation.

yourself, by using Bayes's theorem, that confirmation is a symmetrical relation – if X confirms Y, then Y confirms X.[7]

The Bayesian definition of confirmation can be used to underscore my earlier point that $\Pr(E)$ should not be defined as the frequency with which E is true. Suppose Susan takes a tuberculosis test several times and it comes out positive every time. This might lead you to think that $\Pr(E) = 1$, where E says that Susan's test outcome is positive. However, if $\Pr(E) = 1$, E cannot confirm the hypothesis T, which says that Susan has tuberculosis. This entailment can be verified by consulting Bayes's theorem. To get things right, you need to see that $\Pr(E)$ is an average whose value is described by the theorem of total probability:

$$\Pr(E) = \Pr(E \mid T)\Pr(T) + \Pr(E \mid notT)\Pr(notT).$$

Doing so allows you to see that $\Pr(E)$ is less than one, which means that E can confirm H.

The next distinctively Bayesian idea I need to describe concerns how agents should change their probability assignments as new evidence rolls in. All the probabilities described in Bayes's theorem use the same probability function $\Pr_A(-)$. The assumptions in A can be thought of as the assumptions that an agent is prepared to make at a given time. Suppose the agent learns (with certainty) that a proposition N is true, where N isn't something she already believed; N is news to her. Her set of assumptions has thereby been augmented. We need a rule that describes how the probabilities she assigned under her earlier probability function $\Pr_A(-)$ are related to the probabilities she should assign under her later probability function $\Pr_{A\&N}(-)$. A rule that describes this relationship is called an updating rule.

Before you are dealt a card from the standard deck of cards that I keep talking about, you assign $\Pr_A$(the card will be the Ace of hearts | the card is red) $= \frac{1}{26}$. Suppose you then learn that the card is red. Call this new piece of information N; N gets added to what you already assumed, namely A. So

[7] Bayesians have proposed different measures of *degree of confirmation*. These agree with the Bayesian definition of confirmation just described, but go beyond it, in that they assign numbers to represent how much the evidence confirms the hypothesis. These measures disagree with each other in that they are *ordinally non-equivalent* (Fitelson 1999). This means that there are Bayesian measures X and Y of degree of confirmation that have this property: $X(H_1, E) > X(H_2, E)$ while $Y(H_1, E) \leq Y(H_2, E)$.

what value should you assign to $\Pr_{A\&N}$(the card will be the Ace of Hearts)? The *rule of updating by strict conditionalization* says that your new unconditional probability should be $\frac{1}{26}$. More generally, the idea is this:

> *The Rule of Updating by Strict Conditionalization:* $\Pr_{t2}(H) = \Pr_{t1}(H \mid N)$ if the totality of what you learned between t_1 and t_2 is that N is true.

This updating rule has two major limitations. First, it characterizes learning as the discovery that some proposition N is true. However, if I tell you that N has a probability of, say, 0.6, that new information isn't something that the machinery of strict conditionalization tells you how to take into account.[8] Second, strict conditionalization describes how you should change your assignments of probability when you add a proposition to your assumptions, but it doesn't tell you what to do if something you previously assumed turns out to be false. The rule of strict conditionalization represents learning as gaining certainties, where a certainty, once gained, can never be lost.[9]

The simple updating rule just described allows me to explain some standard vocabulary that is used in connection with Bayes's theorem. I have described $\Pr_A(H \mid E)$ and $\Pr_A(H)$ as the conditional and the unconditional probability of H, but it is customary to describe the former as H's *posterior* probability and the latter as H's *prior* probability. This temporal terminology is a bit misleading; it suggests that $\Pr_A(H \mid E)$ is a probability assignment made later (after you've learned that E is true) whereas $\Pr_A(H)$ is a probability assignment made earlier (before you learn that E is true). In fact, the A subscript shows that both these probability assignments hold true under a single set of assumptions – the assumptions that an agent makes at a given time. And remember that you don't need to think that E is true to consider the value of $\Pr_A(H \mid E)$! What is true is that the old conditional probability $\Pr_A(H \mid E)$ is where the new unconditional probability $\Pr_{A\&E}(H)$ comes from (under the rule of strict conditionalization) when you learn that E is true. The old conditional probability gives rise to a new unconditional probability. Don't let the temporal labels "posterior" and "prior" confuse you. $\Pr_A(H \mid E)$ is the posterior probability of H in the sense that its value is the same as the value of $\Pr_{A\&E}(H)$, once you learn that E is true. Notice that the former

[8] Jeffrey (1965) develops a theory of updating for this more general notion of learning.
[9] Titelbaum (2013) develops a Bayesian model for losing certainties.

probability doesn't involve the assumption that E is true, but the latter one does.

Bayes's theorem allows you to compute how gaining new evidence E should lead you to change your degree of confidence in the hypothesis H. The posterior probability may have a different value from the prior. However, there are two cases in which no such change is possible. If $\Pr(H) = 1$ (or 0), then $\Pr(H \mid E) = 1$ (or 0), no matter what E is. The two extreme probability values (0 and 1) are "sticky." This is why Bayesians advise you to be extremely circumspect about assigning a hypothesis a prior of 0 or 1. In doing so, you are saying that no future experience could make it reasonable for you to change how confident you are in H.

One more bit of terminology can now be introduced. I used the gremlin example to illustrate the difference between $\Pr(H \mid E)$ and $\Pr(E \mid H)$. We are now calling the first of these H's posterior probability. The second also has a name – it is called H's *likelihood*. This terminology, due to R. A. Fisher, is unfortunate. In ordinary English, talking about the probability of H and the likelihood of H are two ways of saying the same thing. In the technical parlance that is now canonical, the two come apart. To avoid confusing them, keep gremlins firmly in mind; when you hear the noise in your attic, the gremlin hypothesis has a high likelihood but a low probability. In what follows, when I say "likelihood," I will be using the term's technical meaning.

I now can complete my sketch of Bayesianism by describing an important consequence of Bayes's theorem. Suppose hypotheses H_1 and H_2 are competing hypotheses. We have some observational evidence E and we want to know which of these hypotheses has the higher posterior probability. If you write Bayes's theorem for each of these hypotheses (please do so!), you can derive the following equation, which is called *the odds formulation of Bayes's theorem*:

$$\frac{\Pr(H_1 \mid E)}{\Pr(H_2 \mid E)} = \frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)} \frac{\Pr(H_1)}{\Pr(H_2)}.$$ [10]

This says that the ratio of posterior probabilities equals the likelihood ratio multiplied by the ratio of prior probabilities. Notice that the unconditional probability of the observations, $\Pr(E)$, has dropped out. Notice also that this version of Bayes's theorem says that there is exactly one way that an observation E can lead you to change how confident you are in H_1 as compared

[10] "Odds" is a word from gambling; it refers to the ratio of posterior probabilities. If this ratio is, say, 20-to-1, that means that H_1 is twenty times as probable as H_2.

with H_2. If the ratio of posteriors is to differ from the ratio of priors, this must be because the likelihoods differ. And the more the likelihood ratio departs from 1, the more the ratio of posterior probabilities departs from the ratio of priors.

The odds formulation of Bayes's theorem makes it easy to see how a hypothesis with a very low prior probability can have its probability driven above 0.5 by several favorable observations, even when one such observation is not enough to push the hypothesis over the top. Consider Susan and her positive tuberculosis test. Suppose the prior probability of Susan's having tuberculosis is 0.001. She then takes a tuberculosis test that has the following property:

$$\Pr(\text{positive test outcome} \mid \text{Susan has tuberculosis}) = 0.96$$

$$\Pr(\text{positive test outcome} \mid \text{Susan does not have tuberculosis}) = 0.02$$

The odds formulation of Bayes's theorem allows you to compute the ratio of posterior probabilities from the numbers we have at hand. The likelihood ratio is 48. The ratio of the priors is $\frac{1}{999}$. So the ratio of the posterior probabilities is $\frac{48}{999}$. This last number means that the posterior probability of Susan's having tuberculosis is $\frac{48}{999+48}$. This is way less than $\frac{1}{2}$, but it is bigger than $\frac{1}{1000}$. The positive test result has increased Susan's probability of having tuberculosis, but not by a whole lot. Now suppose that Susan takes the test a second time and again gets a positive result. Since the two test results are independent of each other, conditional on each of the two hypotheses, the odds formulation of Bayes's theorem will take the following form:

$$\frac{\Pr(H_1 \mid E_1 \& E_2)}{\Pr(H_2 \mid E_1 \& E_2)} = \frac{\Pr(E_1 \mid H_1)}{\Pr(E_1 \mid H_2)} \frac{\Pr(E_2 \mid H_1)}{\Pr(E_2 \mid H_2)} \frac{\Pr(H_1)}{\Pr(H_2)}.$$

The product of the two likelihood ratios is (48)(48) = 2304. Given the ratio of the priors, the ratio of the posterior probabilities is now $\frac{2304}{999}$, so the probability of tuberculosis is now $\frac{2304}{999+2304}$, which is about 0.69. The *single* positive test outcome doesn't entail that Susan probably has the disease, but the *two* positive outcomes together have that implication. People often think that if they take a reliable tuberculosis test and get a positive outcome, then they probably have tuberculosis. Kahneman and Tversky (1985) call this the *base rate fallacy*; the mistake is the failure to take account of the prior probability of tuberculosis.

When I introduced the odds version of Bayes's theorem, I mentioned that the likelihood ratio represents the sole vehicle in the Bayesian framework

whereby new evidence can modify your relative confidence in competing hypotheses. It will be useful to have a principle that isolates this unique role. Hacking (1965) calls the following *the law of likelihood*:

> Evidence E favors hypothesis H_1 over hypothesis H_2 if and only if $\Pr(E \mid H_1) > \Pr(E \mid H_2)$.

When the evidence favors H_1 over H_2 in this sense, the ratio of posterior probabilities exceeds the ratio of priors.

The law of likelihood isn't a deductive consequence of the odds formulation of Bayes's theorem. The theorem is a mathematical fact, but the law is not a truth of mathematics; *favoring* isn't a concept that gets used in the axioms of probability. So perhaps we should regard the law as a proposed explication of the ordinary language concept of favoring. If we do so, we must conclude that the law is flawed. Suppose a talented weather forecaster looks at today's weather conditions and concludes that there probably will be snow tomorrow. The forecaster might summarize this finding by saying that the present weather conditions favor snow tomorrow over no snow tomorrow. Here the word "favoring" is being used to describe an inequality between *probabilities*, not a *likelihood* inequality; what is being claimed is that Pr(snow tomorrow | today's weather conditions) > Pr(no snow tomorrow | today's weather conditions). So if the law of likelihood is an explication of the word "favoring," it is flawed (Sober 2008b). An alternative interpretation of the law of likelihood is better. We can regard the law as a stipulation; the term "favoring" is being used to mark the fact that likelihood inequalities have a special epistemic significance, with no pretense that the law captures every proper use of the word "favoring" in ordinary English.[11] It is not for nothing that Bayesians have come to call the likelihood ratio "the Bayes factor."[12]

It is a consequence of the law of likelihood that the evidence at hand may favor an implausible hypothesis over a sensible one. When you observe

[11] Stipulations are often said to be "arbitrary." But within the Bayesian framework, there is nothing arbitrary about the claim that the likelihood ratio plays a special epistemic role. What is arbitrary is using the word "favoring" to name that role. This, by the way, reveals one limitation of the idea that philosophy's sole aim is to explicate concepts that already have names in ordinary language.

[12] Fitelson (2011) argues that Bayesians should reject the law of likelihood; I reply in Sober (2011d).

that the card you are dealt is an ace, the law says that this observation favors the hypothesis that the deck is made entirely of aces over the hypothesis that the deck is normal (since $1 > 4/52$). This may sound like an objection to the law, but there is a reply. Your doubts about the first hypothesis stem from information you had before you observed the ace, not from what you just observed (Edwards 1972). Likelihood comparisons are supposed to isolate what the evidence says, not to settle which hypotheses are more probable than which others.

Another feature of the law of likelihood is that it says that an observation can favor one hypothesis over another even when neither hypothesis predicts the observation. Suppose $\Pr(E \mid H_1) = 0.001$ and $\Pr(E \mid H_2) = 0.000001$. Neither hypothesis "predicts" E in the sense of saying that E is more probable than not, but the fact remains that E discriminates between the two hypotheses. Asking what a hypothesis "predicts" is a highly imperfect guide to interpreting evidence.

To keep things simple, I have treated the law of likelihood as a Bayesian idea, and I have talked about two philosophies of scientific inference – Bayesianism and frequentism. In fact there is a third camp; there are non-frequentists who are critical of Bayesianism (Edwards 1972; Royall 1997). These "likelihoodists" think that the law of likelihood stands on its own; they think that its justification does not depend on the role that likelihoods play in the odds formulation of Bayes's theorem. The motivation for likelihoodism is illustrated by the following example. When Arthur Stanley Eddington observed the bending of light during a solar eclipse, this was widely regarded as strong evidence favoring Einstein's general theory of relativity over the classical physics of Newton. Likelihoodists represent this in terms of the relationship between two likelihoods:

$$\begin{aligned}&\Pr(\text{Eddington's data on the solar eclipse} \mid \text{general relativity theory})\\ &\quad > \Pr(\text{Eddington's data on the solar eclipse} \mid \text{classical mechanics}).\end{aligned}$$

You don't need to think about the prior probability of either theory to see that this inequality is true.[13]

[13] You can see here why likelihoodists don't like the "definition" of conditional probability as a ratio of unconditional probabilities. Likelihoodists want likelihoods to "make sense" even when priors do not.

Although likelihoodists aren't Bayesians, there is a formal connection between the law of likelihood and Bayesian confirmation theory:

$$\Pr(E \mid H) > \Pr(E \mid notH) \text{ if and only if } \Pr(H \mid E) > \Pr(H).$$

E favors *H* over *notH* (in the sense of the law of likelihood) precisely when *E* confirms *H* (in the sense of Bayesianism). I suggest that you prove this biconditional. Doesn't this formal connection of the two *ism*'s force likelihoodists to admit that they are Bayesians under the skin? Not really. Besides eschewing prior probabilities, likelihoodists think that assigning a value to $\Pr(E \mid notH)$ often lacks an objective justification. It is clear enough what the probability was of Eddington's observations of the solar eclipse, given general relativity. However, the probability of those observations, given the negation of general relativity, is more opaque. The negation of general relativity is a vast disjunction, covering all possible alternatives to general relativity, even ones that have not yet been formulated. The likelihood of *notGTR* therefore takes the following form:

$$\Pr(O \mid notGTR) = \sum_i \Pr(O \mid A_i) \Pr(A_i \mid notGTR).$$

The likelihood of *notGTR* is a weighted average of the likelihoods of all the alternatives (the A_i's) to *GTR*; to compute this average, you need to know how probable each A_i is, given *notGTR*. The negation of the general theory of relativity is an example of what philosophers of science call a "catchall hypothesis." Likelihoodists restrict their epistemology to "specific" theories – to general relativity and Newtonian mechanics, for example. So there are *two* reasons why likelihoodists aren't Bayesians: they don't want to talk about the prior and posterior probabilities of theories, *and* they don't want to talk about the likelihoods of catchalls (Sober 2008b).[14]

[14] Can the objection to Bayesianism that focuses on its need for prior probabilities be dealt with by appealing to various theorems concerning "the washing out of priors"? The idea here is that agents who start with very different prior probabilities and then interpret the evidence in the same way (because they agree on the values of the likelihoods) will end up agreeing on the posterior probabilities; their different starting points don't matter in the long run. The mathematical arguments being appealed to here are correct, but the problem is that they are asymptotic. When agents who have different priors confront a finite data set, they will disagree about the posteriors, often dramatically; what would happen in the infinite long run doesn't change that point. Think about Susan's single tuberculosis test.

It may be helpful to think of the difference between Bayesianism and likelihoodism in terms of the distinction between *private* and *public*. Bayesianism is a philosophy for individual agents who each want to decide how confident they should be in various hypotheses. Likelihoodism is an epistemology for the public world of science; it aims to isolate something objective on which agents can agree despite the fact that they differ in terms of their prior degrees of confidence in the hypotheses under consideration. Agents need prior and posterior probabilities to live their lives, but science needs something that in an important way transcends individual differences. This suggestion does not deny that scientists are agents.

Bayesianism comes in many forms, but to organize ideas let's lump these variants together under a single banner: *computing the posterior probabilities of hypotheses is always an attainable goal*. Likelihoodists claim that this is often impossible to achieve. When Bayesianism fails, likelihoodists hold that discovering which of several specific hypotheses the evidence favors is an attainable goal. Likelihoodism's goal is more modest than Bayesianism's. In a sense, likelihoodism is an *attenuated* Bayesianism; likelihoodism describes what remains of Bayesianism when some of it is stripped away. I have yet to describe what frequentism embraces as its attainable goal. In fact, I think there is no such thing; frequentism is too big a tent for that to be true. However, I have mentioned that frequentists have something negative in common. They want to use probabilistic tools in scientific inference without ever assigning probabilities to hypotheses. Later in this chapter I'll discuss one variety of frequentism and outline its goals and methods.

The present lay of the land is that most theorists about inference are *monists*; they sign up under a single *ism* and swear allegiance to it 100 percent of the time. I am inclined to be more pluralistic. I think that Bayesian, likelihoodist, and frequentist ideas all have their place. The attainable goals in scientific inference vary from problem to problem.

Ockham's razor for Bayesians

The odds formulation of Bayes's theorem has great significance for our investigation of Ockham's razor. For Bayesians, parsimony is not rock bottom; rather it is Bayes's theorem that is fundamental. This means that if Bayesians are going to show that a simpler theory S has a higher posterior probability than a theory C that is more complex, they must show that S has the higher

likelihood or that it has the higher prior probability (or both). In saying this, I am not endorsing Bayesianism as a true and complete epistemology of science. Rather, I am stating an *if*: *if* you are a Bayesian and you think that simplicity is epistemically relevant, there are just two stories you can tell about why this is so. Bayesians of course have the option of scoffing at the relevance of parsimony, and some have done so.

Two kinds of prior probability

When Bayesians talk about prior probabilities, this often involves substantive empirical background assumptions. For example, when Susan takes a tuberculosis test and the test comes out positive, you may want to figure out how probable it is that she has tuberculosis, given this observation. Calculating this posterior probability requires that you assign a value to a prior probability. What probability should you assign to her having tuberculosis before you take account of her test outcome?

As already noted, what is a prior probability at one time is often identical in value to an earlier posterior probability. The prior probability at time t_2, $\Pr_{t2}(S \text{ has tuberculosis})$, will have the same value as the posterior probability at the earlier time t_1, $\Pr_{t1}(S \text{ has tuberculosis} \mid S \text{ lives in Wisconsin})$ if the only relevant fact you learn between t_1 and t_2 is that Susan lives in Wisconsin. If you also know that the frequency of tuberculosis in the state this year is about 0.00001 (approximately 60 cases in a population of about 6000000), you may want to assign your prior at t_2 a value of 0.00001. This sort of prior probability is different from what Bayesians term a "first prior." The first prior of Susan's having tuberculosis must be based on no empirical evidence at all. Some Bayesians think that a proper theory of scientific inference requires that one assign first priors to hypotheses. Here the assumptions that constitute the probability function $\Pr_A(-)$ are merely the logical truths. Although it isn't weird to assign a prior probability to Susan's having tuberculosis on the assumption that Susan lives in a state where the frequency of tuberculosis is 0.00001, it is hard to understand how a prior probability can be assigned to this proposition on the assumption of tautologies alone. Yet, many Bayesians think this is necessary.

The traditional solution to this problem, which many Bayesians now reject, is to appeal to the *principle of indifference*. This principle says that if there are n exclusive and exhaustive propositions (called a *partition*), and you have no