Introduction to Bayesian Epistemology

Elliott Sober

1. Deriving Bayes' Theorem

Bayes' theorem is a piece of mathematics. It is called a theorem because it is derivable from a simple definition in probability theory. As a piece of mathematics, it is not controversial. Bayesianism, on the other hand, is a controversial philosophical theory in epistemology. It proposes that the mathematics of probability theory can be put to work in explicating various concepts connected with issues about evidence, confirmation, and rational belief.

Before I make an observation, I assign a probability to the hypothesis H; this probability may be high, medium, or low. After I make the observation, thereby learning that some observation statement O is true, I want to update the probability I assigned to H, to take account of what I have just learned. The probability that H has before the observation is called its *prior probability*; it is represented by Pr(H). The probability that H has in the light of the observation O is called its *posterior probability*; it is represented by the conditional probability Pr(H│O) (read this as "the probability of H, given O"). Bayes' theorem shows how the prior and the posterior probability are related.

The conditional probability Pr(A│B) is defined as follows:

$$\Pr(A \mid B) = \frac{\Pr(A\&B)}{\Pr(B)}\ .$$

This definition (due to Kolmogorov) is intuitive. What is the probability that a card drawn at random from a standard deck is a heart, given that it is red? Well, the probability that it is a red heart is 1/4; the probability that it is red is ½. Thus, the answer is: ½.

By switching A's and B's with each other, it will also be true that

$$\Pr(B \mid A) = \frac{\Pr(A\&B)}{\Pr(A)}.$$

These two equalities allow the probability of the conjunction (A&B) to be expressed in two different ways:

$$\Pr(A\&B) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A).$$

From this last equality, we can obtain Bayes' theorem:

$$\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

2. Bayesian Definitions of Confirmation and Disconfirmation

Let's rewrite Bayes' theorem with some new letters. We want to figure out what the probability of a hypothesis H is in the light of observations O. According to Bayes' theorem, this conditional probability, Pr(H | O), can be expressed as follows:

$$\Pr(H \mid O) = \frac{\Pr(0|H)\Pr\,(H)}{\Pr(O)}.$$

This expression can be rewritten as an equality between two ratios:

$$(*) \qquad \frac{\Pr(H|O)}{\Pr(H)} = \frac{\Pr(O|H)}{\Pr(O)}$$

When an observation O is obtained, it may have three different kinds of significance for the question of whether the hypothesis H is true. O may confirm H, O may disconfirm H, or O may be evidentially irrelevant to H. Bayesian theory says that each of these ideas can be understood in terms of a relationship between the prior and posterior probabilities of H. Here is the Bayesian proposal:

(1) O confirms H iff Pr(H | O) > Pr(H)

O disconfirms H iff Pr(H | O) < Pr(H)

O is evidentially irrelevant to H iff Pr(H | O) = Pr(H).

Notice that these proposals have implications for whether the left-hand ratio in (*) is greater than, less than, or equal to unity. For example, if O confirms H, $\frac{\Pr(O|H)}{\Pr(O)}$ will be greater than unity. Let's consider some implications of this. First, suppose that H deductively implies O. If so, Pr(O | H)=1. In this case, notice that $\frac{\Pr(O|H)}{\Pr(O)}$can't be less than unity. It follows via (*) that $\frac{\Pr(H|O)}{\Pr(H)}$can't be less than unity. This makes sense of the following idea: when you deduce an observational prediction from a hypothesis, and the prediction comes true, this result can't disconfirm the hypothesis. The hypothesis may go up in probability or it may stay the same, but it can't decline in probability.

If Pr(O | H)=1, how could the observation that O is true fail to confirm the hypothesis? This will happen if Pr(O) = 1. That is, if you were certain that O would be true before you made the observation, the fact that the observation comes true does not confirm H. If H deductively implies O and the truth of O is to confirm H, then Pr(O)<1. True predictions that are totally unsurprising fail to confirm.

In thinking about the idea of confirmation, there is a simple fact that you want to bear clearly in mind. An observation may confirm a hypothesis, even though the hypothesis is still very improbable in the light of the observation. Suppose Pr(H | O) = 1/1,000 and Pr(H) = 1/1,000,000. In this case, the observation increased the probability of the hypothesis a thousand fold. Even so, H remains very improbable. Bayesians use the term “confirm” to mean “probability raising;” a “confirmed” hypothesis may not be worthy of belief.

3. Comparing Posterior Probabilities

According to Bayesianism, confirmation is a “diachronic” relation -- it involves a before-and-after comparison. However, we sometimes are interested in making a distinct, synchronic, comparison -- we want to say whether an observation makes one hypothesis more probable than another hypothesis. Here we are comparing two posterior probabilities – Pr(H1 │ O) and Pr(H2 │ O). Two applications of Bayes’ theorem yields

$$\mathrm{Pr}(\mathrm{H1}\,|\,\mathrm{O}) = \frac{\mathrm{Pr}(0|H1)\mathrm{Pr}\,(H1)}{\mathrm{Pr}(O)}.$$

$$\mathrm{Pr}(\mathrm{H2}\,|\,\mathrm{O}) = \frac{\mathrm{Pr}(0|H2)\mathrm{Pr}\,(H2)}{\mathrm{Pr}(O)}.$$

From these two statements, we obtain:

(2) Pr(H1│O) > Pr(H2│O) iff Pr(O│H1)Pr(H1) > Pr(O│H2)Pr(H2).

Which hypothesis has the higher posterior probability depends on two considerations -- the prior probabilities of the hypotheses, and the probabilities that each hypothesis confers on the observations.

It follows from (2) that H1 might have the higher posterior probability even though H1 says that the observations were very improbable, whereas H2 says that they were very probable. It also is possible for O to raise the probability of H1 (a diachronic result), even though, synchronically, Pr(H1│O) < Pr(H2│O).

Here is a case that illustrates these possibilities: Suppose I sample three balls with replacement from an urn. That is, I take a ball out, note its color, and then return it to the urn, which I then shake; I then draw another. Suppose my observation (O) is that the three balls I've drawn are all green. There are two hypotheses I want to consider. These are:

H1 : All the balls in the urn are green.
H2 : 50% of the balls in the urn are green.

We want to answer two separate questions. Does O make H1 more probable than H2? And does O make H1 more probable than it was before?

To answer these questions, I need to provide more details. To use Bayes' Theorem, we need to make sense of the prior probabilities of H1 and H2. That is, we need to be able to say how probable it is that the urn had one composition rather than another, before the sampling from the urn was performed. Let's imagine that the urn was composed by the following process. There were a hundred buckets, each containing balls. A bucket was chosen at random and then dumped into the urn. In bucket #1, all the balls are green; in buckets 2 through 100, half the balls are green. Given the process just described, we can assign prior probabilities as follows:

Pr(H1) = 1/100 Pr(H2) = 99/100

The next step is to consider how probable the observation O would have been, if H1 had been true. Clearly, if all the balls in the urn are green, then the probability that the three sampled balls should have been green is unity. On the other hand, if H2 were true, then the probability of obtaining three green balls in three draws is (½)(½)(½) = 1/8. Thus, the probability of the observation, conditional on each of the two hypotheses, is:

Pr(O | H1) = 1 Pr(O | H2)= 1/8.

The last probability we need to figure out is the "unconditioned probability of the observations"
-- the quantity Pr(O). But how can we figure out how probable it was that three green balls should have been drawn without knowing which bucket was the one that filled the urn? Well, we know that there was a 1/100 chance that the urn was filled from bucket #1; in that case the probability that the three sampled balls should have been green would be 1. On the other hand, there is a 99/100 chance that the urn was filled from one of the other ninety-nine buckets, in which case the probability that the three balls should have been green would have been 1/8. The probability of O takes both these possibilities into account, as follows:

Pr(O) = Pr(O | H1)Pr(H1) + Pr(O | H2)Pr(H2)

= (1)(1/100) + (1/8)(99/100) ≈ 0.13.

This is an approximate value for Pr(O). We now can use Pr(H1), Pr(O | H1), and Pr(O) to compute Pr(H1 | O), by Bayes's theorem:

$$\Pr(H1 \mid O) = \frac{\Pr(0|H1)\Pr(H1)}{\Pr(O)} = \frac{(1)(\frac{1}{100})}{0.13}$$

So Pr(H1 | O) is about 0.08.

Notice that the observation makes H1 far more probable than it was initially. H1 enjoyed an eight-fold increase in its probability. Yet, this fact of confirmation does not mean that H1 became very probable. Indeed, it did not do so: Pr(H1 | O) is far less than Pr(H2 | O); 0.08 < 0.92.

According to proposition (2), there are *two* factors that affect whether one hypothesis will be more probable in the light of the evidence than another. First, there is the question of how good a job the hypotheses do in predicting the evidence at hand. This issue is represented by the quantities Pr(O | H1) and Pr(O | H2). Second, there is the question of how plausible the hypotheses were before the present evidence was obtained; this is represented by the quantities Pr(H1) and Pr(H2). In this example, H1 gets a high mark on the first consideration, but a low one on the second.

4. Likelihood

Some terminology: "$\Pr(O \mid H)$" is sometimes called the "likelihood" of H. This is a technical usage; it is apt to be misleading, since "x is likely" and "x is probable" are synonyms in ordinary English. However, the likelihood of a hypothesis [$\Pr(O \mid H)$] and its probability [$\Pr(H \mid O)$] can have very different values, as the urn example illustrates.

I so far have defined two ideas about evidence in probabilistic terms. First, there was the before-and-after notion that I called confirmation, described in proposition (1). Second, there was a comparison of the probabilities of two hypotheses in the light of the same evidence, described in proposition (2) . Now it's time for a third. We may ask whether an observation supports one hypothesis better than another. Here we're not interested in whether the one hypothesis has a higher prior probability than the other; we want to isolate what the impact of the observation is. I suggest that this idea can be understood as follows: one hypothesis is better supported by an observation than another is if and only if the first hypothesis makes the observation more probable than the other hypothesis does:

(3) H1 is better supported than H2 by O iff $\Pr(O \mid H1) > \Pr(O \mid H2)$.

Here we have the idea that differential support is measured by *likelihood*. If one hypothesis says that what I observed was to be expected, whereas the other hypothesis says that what I observed was very improbable, it is the first that is better supported by the observation. Proposition (3) is often called the *Law of Likelihood*.

5. An Exercise

Here is a problem that you should be able to solve by reasoning in a way parallel to the urn problem: Suppose that a disease is found in 1/100 people in the US. We select a US individual at random and then give the individual a diagnostic test, which is 90% reliable, by which I mean the following: if an individual has the disease, the probability that the test will come out positive is 0.9, and if an individual does not have the disease, the probability is 0.9 that the test will come out negative. Suppose the test comes out positive. What is the probability that the individual has the disease, given this positive test result? How does this probability compare with the probability that the individual does not have the disease, given that the test result was positive? Calculate the relevant quantities and then plug them into Bayes' theorem.

Psychologists have found that people often do better at problems like this when they formulate them as problems about frequencies in a population of known size. So suppose that the population contains 1,000 people, and that 10 of them have the disease, while 990 do not. What would happen if you gave the test to all 1,000 people? Fill in the following 2-by-2 table with the approximate numbers you'd expect to find in each cell, given that the test is 90% reliable:

	10 have the disease	990 do not have the disease
number of positive outcomes		
number of negative outcomes		

Now suppose someone in the population has a positive test result. What is the probability that he or she has the disease?

Two observations: (i) Notice that I described the reliability of the test procedure by describing conditional probabilities of the form $\Pr(\pm\text{test} \mid \pm\text{disease})$. These numbers do not settle the values of probabilities of the form $\Pr(\pm\text{disease} \mid \pm\text{test})$. This illustrates how likelihood and posterior probabilities are different. (ii) Although it is useful in this problem to think of probabilities in terms of actual frequencies, it isn't true that the probability of an event and its actual frequency must be the same. A fair coin can be tossed an odd number of times and then destroyed. Still, it sometimes makes problems easier to solve if you think of probabilities in terms of the (approximate) actual frequencies you'd expect to find.

6. The Dispute about Bayesianism

Philosophers who criticize Bayesianism do so mainly because they believe that prior probabilities cannot be interpreted objectively, and when they merely reflect subjective degrees of belief, they have no relevance to how evidence ought to interpreted. In our urn example, it *did* make sense to talk about the objective prior probabilities of the hypotheses. The reason was that we viewed the two hypotheses H1 and H2 as possible outcomes of a chance process. I said that the urn was filled by choosing at random from 100 buckets. It was on the basis of this story that we assigned priors in the way we did.

Consider the fact that many of the hypotheses that scientists wish to test do *not* describe possible outcomes of a chance process. For example, take Newton's law of gravitation (G). It makes sense to talk about what G predicts about observations. Perhaps G says that some observations are probable while others are improbable. This will allow us to make sense of the *likelihood* of G. $\Pr(O \mid G)$ will make sense. However, what is the *probability* that Newton's law is true? In particular, we need to make sense of the idea that G has a *prior* probability. Before we do any observational tests of the theory, what probability should we assign to it?

Suppose God had chosen the laws that govern our universe by sampling balls from an urn. On each ball is written a set of laws. If Newton's law was written on just one ball and there were 1000 in the urn, then the prior probability of the law would be 1/1,000. However, no one believes this story about the process that gave our universe the laws it possesses. In the absence of any alternative and plausible process model, critics of Bayesianism decline to assign probabilities to Newton's law. If probabilities aren't objective, these critics don't want to use them in the evaluation of scientific theories. The reason is that if probabilities merely reflect an agent's *degree of belief*, then different agents may have different degrees of confidence in the hypothesis in question; one won't be able to say which assignment of probabilities is correct and which others are incorrect.

To this, Bayesians often reply with the "swamping of priors" argument. They point out that it often doesn't matter what prior probabilities one assigns. Once a reasonable amount of evidence becomes available, people will end up assigning nearly the same posterior probabilities, even if they started out with very different priors. Critics sometimes reply that the swamping of priors argument does not show that the idea of prior probabilities is legitimate.

Another Bayesian strategy is to try to show how objectively correct prior probabilities can be assigned even though one has no information about what processes (if any) influence which hypothesis is true. Bayesians who go this route try to formulate a plausible version of the *Principle of Indifference* (PI). The PI says, roughly, that if you have no reason to assign H1 and H2 different probabilities, then you should assign them the same probability. Stated with a bit more generality, the PI says that if you have no information that would allow you to say which of n exclusive and exhaustive options will come true, you should assign each a probability of 1/n. This principle, if correct, would allow one to obtain knowledge of probabilities from the fact that one is ignorant.

The problem with the PI is that possibilities can be sliced up in different ways, and these different ways of dividing the pie generate probability assignments that are incompatible with each other. If you don't know anything about the color of my favorite sweater, should you assign equal probabilities to the four options [Green, Blue, Red, Black] or equal probabilities to the five options [Light Green, Dark Green, Blue, Red, Black]?

A quantitative example exhibits the same problem. Suppose you know that some particular object has a length (L) that is somewhere between 2 and 4 meters. Applying the PI, you might reason that

(A) Pr(L is between 2 and 3) = Pr(L is between 3 and 4).

But now consider the value of the quantity L^2. You know that L^2 is somewhere between 4 and 16. If you apply the PI to the range of values that L^2 might take, you might end up saying that

Pr(L^2 is between 4 and 10) = Pr(L^2 is between 10 and 16).

But this assignment of probabilities to L^2 entails something about the probabilities of L's values. It entails that

(B) Pr(L is between 2 and $\surd 10$) = Pr(L is between $\surd 10$ and 4).

Note that (A) and (B) are incompatible. Why should one apply the PI to L rather than to L^2? In fairness, I should point out that there are Bayesians who try to refine the PI so that it doesn't generate contradictions.

In terms of the concepts we defined before, critics of Bayesianism will not accept the Bayesian definition of confirmation and disconfirmation, though they may be happy to talk about the likelihoods of hypotheses as measures of how well supported they are.

Side-by-side with this philosophical dispute about the use of prior probabilities, there is a parallel disagreement that arises in connection with likelihoods. What statisticians call "simple hypotheses" specify the probabilities of different observational outcomes. Here's an example from Mendelian genetics:

Pr(offspring is an Aa heterozygote │ Mom is AA and Dad is Aa) = ½ .

The hypothesis about Mom and Dad's genotype is "simple" (another technical term) because it tells you what the probabilities are for the offspring's having different possible genotypes. However, now let's consider the hypothesis that Mom is AA. What is the value of Pr(offspring is an Aa heterozygote │ Mom is AA)? This is a weighted average:

Pr(offspring is Aa │ Mom is AA) =

Pr(offspring is an Aa │ Mom is AA and Dad is AA)Pr(Dad is AA │ Mom is AA) +
Pr(offspring is an Aa │ Mom is AA and Dad is Aa)Pr(Dad is Aa │ Mom is AA) +
Pr(offspring is an Aa │ Mom is AA and Dad is aa)Pr(Dad is aa │ Mom is AA).

The rules of Mendelism entail that this equals

(0)Pr(Dad is AA │ Mom is AA) + (1/2)Pr(Dad is Aa │ Mom is AA) +
(1)Pr(Dad is aa │ Mom is AA).

If we had data on how often males and females form different mating pairs, we could estimate the values of probabilities that have the form Pr(Dad's genotype│Mom's genotype) and complete our calculation of the likelihood of the hypothesis that Mom is AA.

However, there are many hypotheses in science whose likelihoods cannot be handled in this way. For example, the negation of a scientific theory is usually not amenable to this treatment. When Eddington tested the General Theory of Relativity (GTR) by measuring the degree to which light from the sun was bent during an eclipse, he was considering a quantity of the form:

Pr(light is bent to such-and-such a degree │ GTR & auxiliary assumptions).

But what is the value of

Pr(light is bent to such-and-such a degree │ not-GTR & auxiliary assumptions)?

As the example concerning Mom's genotype illustrates, this will be a weighted average over the likelihoods of the different specific alternatives A1, A2, … An to the GTR, where the weighting terms are

Pr(A1 │ not-GTR), Pr(A2 │ not-GTR), …, Pr(An │ not-GTR).

Here the alternatives must cover *all* the alternatives, even ones that have not been conceptualized yet.

The negation of GTR is a “composite hypothesis” (“composite” is the opposite of “simple”). Anti-Bayesians despair of giving an objective interpretation to the likelihoods of composite hypotheses, just as they despair of assigning objective prior probabilities to scientific theories like Newton’s theory of gravitation.

Bayesian statisticians, as opposed to Bayesian philosophers, rarely talk about the prior probabilities of a theory or about the likelihoods of a theory’s negation (when these can’t be interpreted objectively). However, they do treat the likelihoods of some composite hypotheses by inventing and using models of the probabilities that different values of a parameter might take. I say “inventing” here because the models aren’t justified in terms of frequency data. For example, consider the problem described earlier of computing Pr(Offspring is Aa | Mom is AA). Suppose one doesn’t have data on the frequencies of different mating pairs from which to estimate quantities of the form Pr(Dad’s genotype | Mom’s genotype). A Bayesian might nonetheless propose that we assume that

$$\begin{aligned}&\mathrm{Pr}(\text{Dad is AA} \mid \text{Mom is AA}) = \mathrm{Pr}(\text{Dad is Aa} \mid \text{Mom is AA}) = \\ &\qquad \mathrm{Pr}(\text{Dad is aa} \mid \text{Mom is AA}) = 1/3,\end{aligned}$$

or that we consider a range of different assignments and see how those assignments affect our evaluation of Pr(Offspring is Aa | Mom is AA). Anti-Bayesians reject this way of handling the problem, claiming that these assignments merely reflect the researcher’s subjective degree of belief.