

Conditional Probability, Independence and Bayes' Theorem

Class 3, 18.05

Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Know the definitions of conditional probability and independence of events.
2. Be able to compute conditional probability directly from the definition.
3. Be able to use the multiplication rule to compute the total probability of an event.
4. Be able to check if two events are independent.
5. Be able to use Bayes' formula to 'invert' conditional probabilities.
6. Be able to organize the computation of conditional probabilities using trees and tables.
7. Understand the base rate fallacy thoroughly.

2 Conditional Probability

Conditional probability answers the question 'how does the probability of an event change if we have extra information'. We'll illustrate with an example.

Example 1. Toss a fair coin 3 times.

(a) What is the probability of 3 heads?

answer: Sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

All outcomes are equally likely, so $P(3 \text{ heads}) = 1/8$.

(b) Suppose we are told that the first toss was heads. Given this information how should we compute the probability of 3 heads?

answer: We have a new (reduced) sample space: $\Omega' = \{HHH, HHT, HTH, HTT\}$.

All outcomes are equally likely, so

$$P(3 \text{ heads given that the first toss is heads}) = 1/4.$$

This is called **conditional probability**, since it takes into account additional conditions. To develop the notation, we rephrase (b) in terms of *events*.

Rephrased (b) Let A be the event 'all three tosses are heads' = $\{HHH\}$.

Let B be the event 'the first toss is heads' = $\{HHH, HHT, HTH, HTT\}$.

The **conditional probability** of A knowing that B occurred is written

$$P(A|B)$$

This is read as

'the conditional probability of A given B '

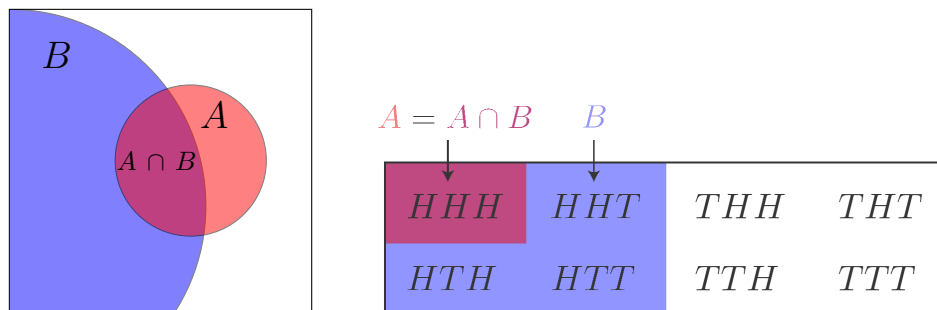
or

'the probability of A conditioned on B '

or simply

‘the probability of A given B ’.

We can visualize conditional probability as follows. Think of $P(A)$ as the proportion of the area of the *whole* sample space taken up by A . For $P(A|B)$ we restrict our attention to B . That is, $P(A|B)$ is the proportion of area of B taken up by A , i.e. $P(A \cap B)/P(B)$.



Conditional probability: Abstract visualization and coin example

Note, $A \subset B$ in the right-hand figure, so there are only two colors shown.

The formal definition of conditional probability catches the gist of the above example and visualization.

Formal definition of conditional probability

Let A and B be events. We define the conditional probability of A given B as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0. \quad (1)$$

Let's redo the coin tossing example using the definition in Equation (1). Recall A = ‘3 heads’ and B = ‘first toss is heads’. We have $P(A) = 1/8$ and $P(B) = 1/2$. Since $A \cap B = A$, we also have $P(A \cap B) = 1/8$. Now according to (1), $P(A|B) = \frac{1/8}{1/2} = 1/4$, which agrees with our answer in Example 1b.

3 Multiplication Rule

The following formula is called the multiplication rule.

$$P(A \cap B) = P(A|B) \cdot P(B). \quad (2)$$

This is simply a rewriting of the definition in Equation (1) of conditional probability. We will see that our use of the multiplication rule is very similar to our use of the rule of product in counting. In fact, the multiplication rule is just a souped up version of the rule of product.

We start with a simple example where we can check all the probabilities directly by counting.

Example 2. Draw two cards from a deck. Define the events: S_1 = ‘first card is a spade’ and S_2 = ‘second card is a spade’. What is the $P(S_2|S_1)$?

answer: We can do this directly by counting: if the first card is a spade then of the 51 cards remaining, 12 are spades.

$$P(S_2|S_1) = 12/51.$$

Now, let's recompute this using formula (1). We have to compute $P(S_1)$, $P(S_2)$ and $P(S_1 \cap S_2)$: We know that $P(S_1) = 1/4$ because there are 52 equally likely ways to draw the first card and 13 of them are spades. The same logic says that there are 52 equally likely ways the second card can be drawn, so $P(S_2) = 1/4$.

Aside: The probability $P(S_2) = 1/4$ may seem surprising since the value of first card certainly affects the probabilities for the second card. However, if we look at *all* possible two card sequences we will see that every card in the deck has equal probability of being the second card. Since 13 of the 52 cards are spades we get $P(S_2) = 13/52 = 1/4$. Another way to say this is: if we are not given value of the first card then we have to consider all possibilities for the second card.

Continuing, we see that

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = 3/51.$$

This was found by counting the number of ways to draw a spade followed by a second spade and dividing by the number of ways to draw any card followed by any other card). Now, using (1) we get

$$P(S_2|S_1) = \frac{P(S_2 \cap S_1)}{P(S_1)} = \frac{3/51}{1/4} = 12/51.$$

Finally, we verify the multiplication rule by computing both sides of (2).

$$P(S_1 \cap S_2) = \frac{13 \cdot 12}{52 \cdot 51} = \frac{3}{51} \quad \text{and} \quad P(S_2|S_1) \cdot P(S_1) = \frac{12}{51} \cdot \frac{1}{4} = \frac{3}{51}. \quad \text{QED}$$

Think: For S_1 and S_2 in the previous example, what is $P(S_2|S_1^c)$?

4 Law of Total Probability

The law of total probability will allow us to use the multiplication rule to find probabilities in more interesting examples. It involves a lot of notation, but the idea is fairly simple. We state the law when the sample space is divided into 3 pieces. It is a simple matter to extend the rule when there are more than 3 pieces.

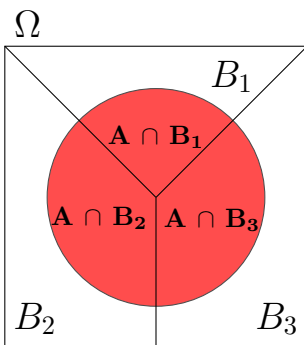
Law of Total Probability

Suppose the sample space Ω is divided into 3 disjoint events B_1 , B_2 , B_3 (see the figure below). Then for any event A :

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \quad (3)$$

The top equation says 'if A is divided into 3 pieces then $P(A)$ is the sum of the probabilities of the pieces'. The bottom equation (3) is called **the law of total probability**. It is just a rewriting of the top equation using the multiplication rule.



The sample space Ω and the event A are each divided into 3 disjoint pieces.

The law holds if we divide Ω into any number of events, so long as they are *disjoint* and *cover* all of Ω . Such a division is often called a *partition* of Ω .

Our first example will be one where we already know the answer and can verify the law.

Example 3. An urn contains 5 red balls and 2 green balls. Two balls are drawn one after the other. What is the probability that the second ball is red?

answer: The sample space is $\Omega = \{rr, rg, gr, gg\}$.

Let R_1 be the event ‘the first ball is red’, G_1 = ‘first ball is green’, R_2 = ‘second ball is red’, G_2 = ‘second ball is green’. We are asked to find $P(R_2)$.

The fast way to compute this is just like $P(S_2)$ in the card example above. Every ball is equally likely to be the second ball. Since 5 out of 7 balls are red, $P(R_2) = 5/7$.

Let’s compute this same value using the law of total probability (3). First, we’ll find the conditional probabilities. This is a simple counting exercise.

$$P(R_2|R_1) = 4/6, \quad P(R_2|G_1) = 5/6.$$

Since R_1 and G_1 partition Ω the law of total probability says

$$\begin{aligned} P(R_2) &= P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) \\ &= \frac{4}{6} \cdot \frac{5}{7} + \frac{5}{6} \cdot \frac{2}{7} \\ &= \frac{30}{42} = \frac{5}{7}. \end{aligned} \tag{4}$$

Probability urns

The example above used probability urns. Their use goes back to the beginning of the subject and we would be remiss not to introduce them. This toy model is very useful. We quote from Wikipedia: http://en.wikipedia.org/wiki/Urn_problem

In probability and statistics, an urn problem is an idealized mental exercise in which some objects of real interest (such as atoms, people, cars, etc.) are represented as colored balls in an urn or other container. One pretends to draw (remove) one or more balls from the urn; the goal is to determine the probability of drawing one color or another, or some other properties. A key parameter is whether each ball is returned to the urn after each draw.

It doesn't take much to make an example where (3) is really the best way to compute the probability. Here is a game with slightly more complicated rules.

Example 4. An urn contains 5 red balls and 2 green balls. A ball is drawn. If it's green a red ball is added to the urn and if it's red a green ball is added to the urn. (The original ball is not returned to the urn.) Then a second ball is drawn. What is the probability the second ball is red?

Interesting!!

answer: The law of total probability says that $P(R_2)$ can be computed using the expression in Equation (4). Only the values for the probabilities will change. We have

$$P(R_2|R_1) = 4/7, \quad P(R_2|G_1) = 6/7.$$

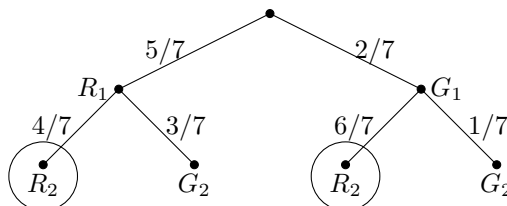
Therefore,

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|G_1)P(G_1) = \frac{4}{7} \cdot \frac{5}{7} + \frac{6}{7} \cdot \frac{2}{7} = \frac{32}{49}.$$

5 Using Trees to Organize the Computation

Trees are a great way to organize computations with conditional probability and the law of total probability. The figures and examples will make clear what we mean by a tree. As with the rule of product, the key is to organize the underlying process into a sequence of actions.

We start by redoing Example 4. The sequence of actions are: first draw ball 1 (and add the appropriate ball to the urn) and then draw ball 2.



You interpret this tree as follows. Each dot is called a **node**. The tree is organized by levels. The top node (**root node**) is at level 0. The next layer down is level 1 and so on. Each level shows the outcomes at one stage of the game. Level 1 shows the possible outcomes of the first draw. Level 2 shows the possible outcomes of the second draw starting from each node in level 1.

Probabilities are written along the branches. The probability of R_1 (red on the first draw) is $5/7$. It is written along the branch from the root node to the one labeled R_1 . At the next level we put in **conditional** probabilities. The probability along the branch from R_1 to R_2 is $P(R_2|R_1) = 4/7$. It represents the probability of going to node R_2 given that you are already at R_1 .

The multiplication rule says that the probability of getting to any node is just the product of the probabilities along the path to get there. For example, the node labeled R_2 at the far left really represents the event $R_1 \cap R_2$ because it comes from the R_1 node. The multiplication rule now says

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1) = \frac{5}{7} \cdot \frac{4}{7},$$

which is exactly multiplying along the path to the node.

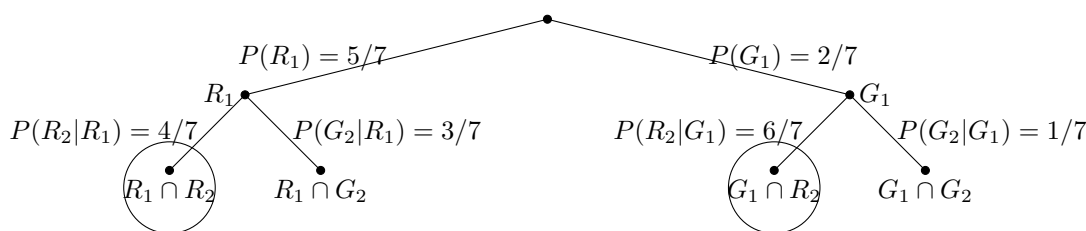
The law of total probability is just the statement that $P(R_2)$ is the sum of the probabilities of all paths leading to R_2 (the two circled nodes in the figure). In this case,

$$P(R_2) = \frac{5}{7} \cdot \frac{4}{7} + \frac{2}{7} \cdot \frac{6}{7} = \frac{32}{49},$$

exactly as in the previous example.

5.1 Shorthand vs. precise trees

The tree given above involves some shorthand. For example, the node marked R_2 at the far left really represents the event $R_1 \cap R_2$, since it ends the path from the root through R_1 to R_2 . Here is the same tree with everything labeled precisely. As you can see this tree is more cumbersome to make and use. We usually use the shorthand version of trees. You should make sure you know how to interpret them precisely.



6 Independence

Two events are independent if knowledge that one occurred does not change the probability that the other occurred. Informally, events are independent if they do not influence one another.

Example 5. Toss a coin twice. We expect the outcomes of the two tosses to be independent of one another. In real experiments this always has to be checked. If my coin lands in honey and I don't bother to clean it, then the second toss might be affected by the outcome of the first toss.

More seriously, the independence of experiments can be undermined by the failure to clean or recalibrate equipment between experiments or to isolate supposedly independent observers from each other or a common influence. We've all experienced hearing the same 'fact' from different people. Hearing it from different sources tends to lend it credence until we learn that they all heard it from a common source. That is, our sources were not independent.

Translating the verbal description of independence into symbols gives

$$A \text{ is independent of } B \quad \text{if} \quad P(A|B) = P(A). \quad (5)$$

That is, knowing that B occurred does not change the probability that A occurred. In terms of events as subsets, knowing that the realized outcome is in B does not change the probability that it is in A .

If A and B are independent in the above sense, then the multiplication rule gives $P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$. This justifies the following technical definition of independence.

Formal definition of independence: Two events A and B are independent if

$$P(A \cap B) = P(A) \cdot P(B) \quad (6)$$

This is a nice symmetric definition which makes clear that A is independent of B if and only if B is independent of A . Unlike the equation with conditional probabilities, this definition makes sense even when $P(B) = 0$. In terms of conditional probabilities, we have:

1. If $P(B) \neq 0$ then A and B are independent if and only if $P(A|B) = P(A)$.
2. If $P(A) \neq 0$ then A and B are independent if and only if $P(B|A) = P(B)$.

Independent events commonly arise as different trials in an experiment, as in the following example.

Example 6. Toss a fair coin twice. Let H_1 = 'heads on first toss' and let H_2 = 'heads on second toss'. Are H_1 and H_2 independent?

answer: Since $H_1 \cap H_2$ is the event 'both tosses are heads' we have

$$P(H_1 \cap H_2) = 1/4 = P(H_1)P(H_2).$$

Therefore the events are independent.

We can ask about the independence of any two events, as in the following two examples.

Example 7. Toss a fair coin 3 times. Let H_1 = 'heads on first toss' and A = 'two heads total'. Are H_1 and A independent?

answer: We know that $P(A) = 3/8$. Since this is not 0 we can check if the formula in Equation 5 holds. Now, $H_1 = \{HHH, HHT, HTH, HTT\}$ contains exactly two outcomes (HHT, HTH) from A , so we have $P(A|H_1) = 2/4$. Since $P(A|H_1) \neq P(A)$ these events are not independent.

Example 8. Draw one card from a standard deck of playing cards. Let's examine the independence of 3 events 'the card is an ace', 'the card is a heart' and 'the card is red'.

Define the events as A = 'ace', H = 'hearts', R = 'red'.

(a) We know that $P(A) = 4/52 = 1/13$, $P(A|H) = 1/13$. Since $P(A) = P(A|H)$ we have that A is independent of H .

(b) $P(A|R) = 2/26 = 1/13$. So A is independent of R . That is, whether the card is an ace is independent of whether it's red.

(c) Finally, what about H and R ? Since $P(H) = 1/4$ and $P(H|R) = 1/2$, H and R are not independent. We could also see this the other way around: $P(R) = 1/2$ and $P(R|H) = 1$, so H and R are not independent.

6.1 Paradoxes of Independence

An event A with probability 0 is independent of itself, since in this case both sides of equation (6) are 0. This appears paradoxical because knowledge that A occurred certainly

gives information about whether A occurred. We resolve the paradox by noting that since $P(A) = 0$ the statement ' A occurred' is vacuous.

Think: For what other value(s) of $P(A)$ is A independent of itself?

7 Bayes' Theorem

Bayes' theorem is a pillar of both probability and statistics and it is central to the rest of this course. For two events A and B **Bayes' theorem** (also called **Bayes' rule** and **Bayes' formula**) says

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}. \quad (7)$$

Comments: 1. Bayes' rule tells us how to 'invert' conditional probabilities, i.e. to **find** $P(B|A)$ from $P(A|B)$.

2. In practice, $P(A)$ is often computed using the law of total probability.

Proof of Bayes' rule

The key point is that $A \cap B$ is symmetric in A and B . So the multiplication rule says

$$P(B|A) \cdot P(A) = P(A \cap B) = P(A|B) \cdot P(B).$$

Now divide through by $P(A)$ to get Bayes' rule.

A common mistake is to confuse $P(A|B)$ and $P(B|A)$. They can be very different. This is illustrated in the next example.

Example 9. Toss a coin 5 times. Let H_1 = 'first toss is heads' and let H_A = 'all 5 tosses are heads'. Then $P(H_1|H_A) = 1$ but $P(H_A|H_1) = 1/16$.

For practice, let's use Bayes' theorem to compute $P(H_1|H_A)$ using $P(H_A|H_1)$. The terms are $P(H_A|H_1) = 1/16$, $P(H_1) = 1/2$, $P(H_A) = 1/32$. So,

$$P(H_1|H_A) = \frac{P(H_A|H_1)P(H_1)}{P(H_A)} = \frac{(1/16) \cdot (1/2)}{1/32} = 1,$$

which agrees with our previous calculation.

7.1 The Base Rate Fallacy

The base rate fallacy is one of many examples showing that it's easy to confuse the meaning of $P(B|A)$ and $P(A|B)$ when a situation is described in words. This is one of the key examples from probability and it will inform much of our practice and interpretation of statistics. You should strive to understand it thoroughly.

Example 10. The Base Rate Fallacy

Consider a routine screening test for a disease. Suppose the frequency of the disease in the population (**base rate**) is 0.5%. The test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

answer: We will do the computation three times: using trees, tables and symbols. We'll use the following notation for the relevant events:

$D+$ = 'you have the disease'

$D-$ = 'you do not have the disease'

$T+$ = 'you tested positive'

$T-$ = 'you tested negative'.

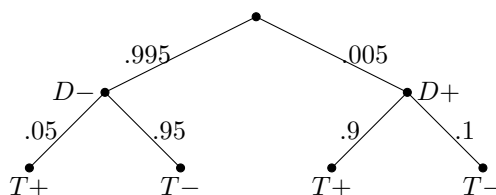
We are given $P(D+) = .005$ and therefore $P(D-) = .995$. The false positive and false negative rates are (by definition) conditional probabilities.

$$P(\text{false positive}) = P(T+ | D-) = .05 \quad \text{and} \quad P(\text{false negative}) = P(T- | D+) = .1.$$

The complementary probabilities are known as the true negative and true positive rates:

$$P(T- | D-) = 1 - P(T+ | D-) = .95 \quad \text{and} \quad P(T+ | D+) = 1 - P(T- | D+) = .9.$$

Trees: All of these probabilities can be displayed quite nicely in a tree.



The question asks for the probability that you have the disease given that you tested positive, i.e. what is the value of $P(D+ | T+)$. We aren't given this value, but we do know $P(T+ | D+)$, so we can use Bayes' theorem.

$$P(D+ | T+) = \frac{P(T+ | D+) \cdot P(D+)}{P(T+)}$$

The two probabilities in the numerator are given. We compute the denominator $P(T+)$ using the law of total probability. Using the tree we just have to sum the probabilities for each of the nodes marked $T+$

$$P(T+) = .995 \times .05 + .005 \times .9 = .05425$$

Thus,

$$P(D+ | T+) = \frac{.9 \times .005}{.05425} = 0.082949 \approx 8.3\%.$$

Remarks: This is called the base rate fallacy because the base rate of the disease in the population is so low that the vast majority of the people taking the test are healthy, and even with an accurate test most of the positives will be healthy people. Ask your doctor for his/her guess at the odds.

To summarize the base rate fallacy with specific numbers

95% of all tests are accurate does not imply 95% of positive tests are accurate

We will refer back to this example frequently. It and similar examples are at the heart of many statistical misunderstandings.



Other ways to work Example 10

Tables: Another trick that is useful for computing probabilities is to make a table. Let's redo the previous example using a table built with 10000 total people divided according to the probabilities in this example.

We construct the table as follows. Pick a number, say 10000 people, and place it as the grand total in the lower right. Using $P(D+) = .005$ we compute that 50 out of the 10000 people are sick ($D+$). Likewise 9950 people are healthy ($D-$). At this point the table looks like:

	$D+$	$D-$	total
$T+$			
$T-$			
total	50	9950	10000

Using $P(T+|D+) = .9$ we can compute that the number of sick people who tested positive as 90% of 50 or 45. The other entries are similar. At this point the table looks like the table below on the left. Finally we sum the $T+$ and $T-$ rows to get the completed table on the right.

	$D+$	$D-$	total
$T+$	45	498	
$T-$	5	9452	
total	50	9950	10000

	$D+$	$D-$	total
$T+$	45	498	543
$T-$	5	9452	9457
total	50	9950	10000

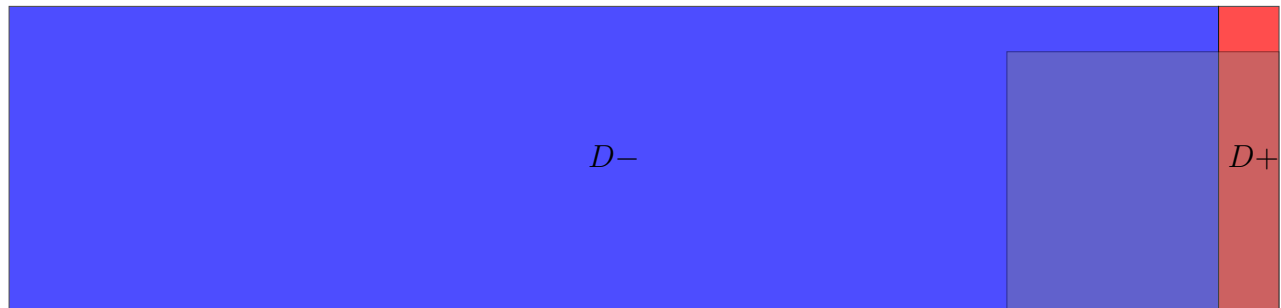
Using the complete table we can compute

$$P(D+|T+) = \frac{|D+ \cap T+|}{|T+|} = \frac{45}{543} = 8.3\%.$$

Symbols: For completeness, we show how the solution looks when written out directly in symbols.

$$\begin{aligned}
 P(D+|T+) &= \frac{P(T+|D+) \cdot P(D+)}{P(T+)} \\
 &= \frac{P(T+|D+) \cdot P(D+)}{P(T+|D+) \cdot P(D+) + P(T+|D-) \cdot P(D-)} \\
 &= \frac{.9 \times .005}{.9 \times .005 + .05 \times .995} \\
 &= 8.3\%
 \end{aligned}$$

Visualization: The figure below illustrates the base rate fallacy. The large blue area represents all the healthy people. The much smaller red area represents the sick people. The shaded rectangle represents the people who test positive. The shaded area covers most of the red area and only a small part of the blue area. Even so, the most of the shaded area is over the blue. That is, most of the positive tests are of healthy people.



7.2 Bayes' rule in 18.05

As we said at the start of this section, Bayes' rule is a pillar of probability and statistics. We have seen that Bayes' rule allows us to 'invert' conditional probabilities. When we learn statistics we will see that the art of statistical inference involves deciding how to proceed when one (or more) of the terms on the right side of Bayes' rule is unknown.