1.4 Conditional Probability and Independence

The first few examples of this section illustrate the idea of conditional probability in a setting of equally likely outcomes.

Example 1. Three coin tosses.

If you bet that 2 or more heads will appear in 3 tosses of a fair coin, you are more likely to win the bet given the first toss lands heads than given the first toss lands tails. To be precise, assume the 8 possible patterns of heads and tails in the three tosses, $\{hhh, hht, hth, htt, thh, tht, tth, ttt\}$, are equally likely. Then the *overall* or *unconditional* probability of the event

$$A = (2 \text{ or more heads in 3 tosses}) = \{hhh, hht, hth, thh\}$$

is $P(A) = 4/8 = 1/2$. But given that the first toss lands heads (say H), event A occurs if there is at least one head in the next two tosses, with a chance of $3/4$. So it is said that the conditional probability of A given H is $3/4$. The mathematical notation for the conditional probability of A given H is $P(A|H)$, read "P of A given H". In the present example

$$P(A|H) = 3/4$$

because $H = \{hhh, hht, hth, htt\}$ can occur in 4 ways, and just 3 of these outcomes make A occur. These 3 outcomes define the event $\{hhh, hht, hth\}$ which is the *intersection* of the events A and H, denoted A *and* H, $A \cap H$, or simply AH. Similarly, if the event H^c = "first toss lands tails" occurs, event A happens only if the next two tosses land heads, with probability $1/4$. So

$$P(A|H^c) = 1/4$$

Conditional probabilities can be defined as follows in any setting with equally likely outcomes.

Counting Formula for $P(A|B)$

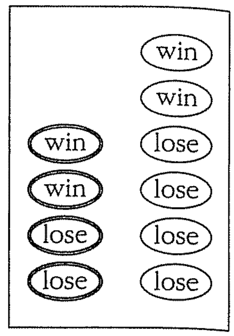
For a finite set Ω of equally likely outcomes, and events A and B represented by subsets Ω, the *conditional probability of A given B* is

$$P(A|B) = \frac{\#(AB)}{\#(B)}$$

the proportion of outcomes in B that are also in A. Here $AB = A \cap B = A$ *and* B is the *intersection* of A and B.

Example 2. **Tickets.**

Problem. A box contains 10 capsules, similar except that four are black and six are white. Inside each capsule is a ticket marked either *win* or *lose*. The capsules are opaque, so the result on the ticket inside cannot be read without breaking open the capsule. Suppose a capsule is drawn at random from the box, then broken open to read the result. If it says *win*, you win a prize. Otherwise, you win nothing. The numbers of winning and losing tickets of each color are given in the diagram, which shows the tickets inside the capsules. Suppose that the capsule has just been drawn, but not yet broken to read the result. The capsule is black. Now what is the probability that you win a prize?



Solution. This conditional probability is the proportion of winners among black capsules:

$$P(\text{win}|\text{black}) = \frac{\#(\text{win and black})}{\#(\text{black})} = \frac{2}{4} = 0.5$$

Compare with the unconditional probability $P(\text{win}) = 4/10 = 0.4$

Example 3. **Two-sided cards.**

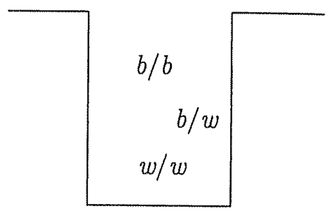
Problem. A hat contains three cards.

One card is black on both sides.

One card is white on both sides.

One card is black on one side and white on the other.

The cards are mixed up in the hat. Then a single card is drawn and placed on a table. If the visible side of the card is black, what is the chance that the other side is white?



Solution. Label the faces of the cards:

b_1 and b_2 for the black–black card;

w_1 and w_2 for the white–white card;

b_3 and w_3 for the black–white card.

Assume that each of these six faces is equally likely to be the face showing uppermost. Experience shows that this assumption does correspond to long-run frequencies, provided the cards are similar in size and shape, and well mixed up in the hat. The outcome space is then the set of six possible faces which might show uppermost:

$$\{b_1, b_2, b_3, w_1, w_2, w_3\}$$

The event {black on top} is identified as

$$\{\text{black on top}\} = \{b_1, b_2, b_3\}$$

Similarly,

$$\{\text{white on bottom}\} = \{b_3, w_1, w_2\}$$

Given that the event {black on top} has occurred, the face showing is equally likely to be b_1, b_2, or b_3. Only in the last case is the card white on the bottom. So the chance of white on bottom given black on top is

$$P(\text{white on bottom}|\text{black on top}) = \frac{\#(\text{white on bottom and black on top})}{\#(\text{black on top})} = \frac{1}{3}$$

Discussion. You might reason as follows: The card must be either the black–black card or the black–white card. These are equally likely possibilities, so the chance that the other side is white is $1/2$. Many people find this argument convincing, but it is basically wrong. The assumption of equally likely outcomes, given the top side is black, is not consistent with long-run frequencies. If you repeat the experiment of drawing from the hat over and over, replacing the cards and mixing them up each time, you will find that over the long run, among draws when the top side is black, the bottom side will be white only about $1/3$ of the time, rather than $1/2$ of the time.

Frequency interpretation of conditional probability. This is illustrated by the previous example. If $P(A)$ approximates to the relative frequency of A in a long series of trials, then $P(A|B)$ approximates the relative frequency of trials producing A among those trials which happen to result in B. A general formula for $P(A|B)$, consistent with this interpretation, is found as follows. Start with the counting formula for $P(A|B)$ in a setting of equally likely outcomes, then divide both numerator and denominator by $\#(\Omega)$ to express $P(A|B)$ in terms of the unconditional probabilities $P(AB) = \#(AB)/\#(\Omega)$ and $P(B) = \#(B)/\#(\Omega)$:

$$P(A|B) = \frac{\#(AB)}{\#(B)} = \frac{\#(AB)/\#(\Omega)}{\#(B)/\#(\Omega)} = \frac{P(AB)}{P(B)}$$

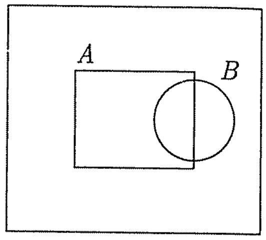
General Formula for $P(A \mid B)$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

If probabilities $P(A)$ are specified for subsets A of an outcome space Ω, then conditional probabilities given B can be calculated using this formula. This restricts the outcome space to B and renormalizes the distribution on B. In case the original distribution is defined by relative numbers, or relative areas, the same will be true of the conditional distribution given B, but with the restriction from Ω to B. To make a clear distinction, $P(A)$ or $P(AB)$ is called an *overall* or *unconditional* probability, and $P(A|B)$ a *conditional* probability.

Example 4. **Relative areas.**

Suppose a point is picked uniformly at random from the big rectangle in the diagram. Imagine that information about the position of this point is revealed to you in two stages, by the answers to the following questions:
Question 1. Is the point inside the circle B?
Question 2. Is the point inside the rectangle A?



Problem. If the answer to Question 1 is yes, what is the probability that the answer to Question 2 will be yes?

Solution. The problem is to find the probability that the point is in the rectangle A given that it is in the circle B. By inspection of the diagram, approximately half the area inside B is inside A. So the required probability is

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\text{Area}(AB)}{\text{Area}(B)} \approx 1/2$$

Remark. The formula for conditional probability in this case corresponds to the idea that given the point is in B, equal areas within B still have equal probabilities.

Tree Diagrams and the Multiplication Rule

In the above example a conditional probability was calculated from overall probabilities. But in applications there are usually many events A and B such that the conditional probability $P(A|B)$ and the overall probability $P(B)$ are more obvious than the overall probability $P(AB)$. Then $P(AB)$ is calculated using the following rearrangement of the general formula for conditional probability:

Multiplication Rule

$$P(AB) = P(A|B)P(B)$$

This rule is very intuitive in terms of the frequency interpretation. If, for example, B happens over the long run about $1/2$ the time ($P(B) = 1/2$), and about $1/3$ of the times that B happens A happens too ($P(A|B) = 1/3$), then A and B happens about $1/3$ of $1/2 = 1/3 \times 1/2 = 1/6$ of the time ($P(AB) = P(A|B)P(B) = 1/6$).

The multiplication rule is often used to set up a probability model with intuitively prescribed conditional probabilities. Typically, A will be an event determined by some overall outcome which can be thought of as occurring by stages, and B will be some event depending just on the first stage. If you think of B happening before A it is more natural to rewrite the multiplication rule, with BA instead of AB and the two factors switched:

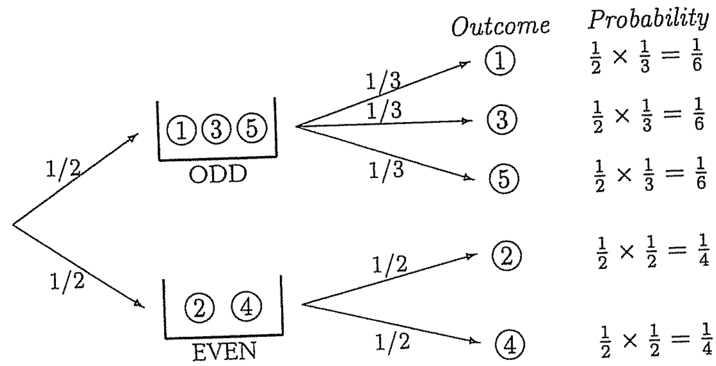
$$P(BA) = P(B)P(A|B)$$

In words, the chance of B followed by A is the chance of B times the chance of A given B.

Example 5. **Picking a box, then a ball.**

Problem. Suppose that there are two boxes, labeled odd and even. The odd box contains three balls numbered 1, 3, 5. The even box contains two balls labeled 2, 4. One of the boxes is picked at random by tossing a fair coin. Then a ball is picked at random from this box. What is the probability that the ball drawn is ball 3?

Solution. A scheme like this can be represented in a *tree diagram*. Each branch represents a possible way things might turn out. Probabilities and conditional probabilities are indicated along the branch.



Because the box is chosen by a fair coin toss,

$$P(\text{odd}) = P(\text{even}) = 1/2$$

The only way to get 3 is to first pick the odd box, then pick 3. By assumption

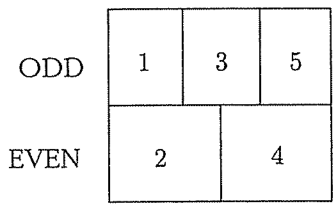
$$P(3|\text{odd}) = 1/3$$

Now by the multiplication rule,

$$P(3) = P(\text{odd and } 3) = P(\text{odd})P(3\,|\,\text{odd}) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

This is the product of the probabilities along the path representing the outcome 3. The corresponding products along the other possible branches give the distribution displayed in the tree diagram.

This is a different representation of the same problem, using a Venn diagram.



Remark 1. A naive approach to the above problem would be to assume that all outcomes were equally likely. But this would imply

$$P(\text{first box}) = P(\text{odd}) = 3/5$$

$$P(\text{second box}) = P(\text{even}) = 2/5$$

which is inconsistent with the box being chosen by a fair coin toss.

Remark 2. The problem could also be solved without conditional probabilities by a symmetry argument, assuming that

$$P(1) = P(3) = P(5) \quad \text{and} \quad P(2) = P(4)$$

$$P(1) + P(3) + P(5) = P(2) + P(4) = 1/2$$

These equations yield the same answer as above.

To summarize the method of the previous example:

Multiplication Rule in a Tree Diagram

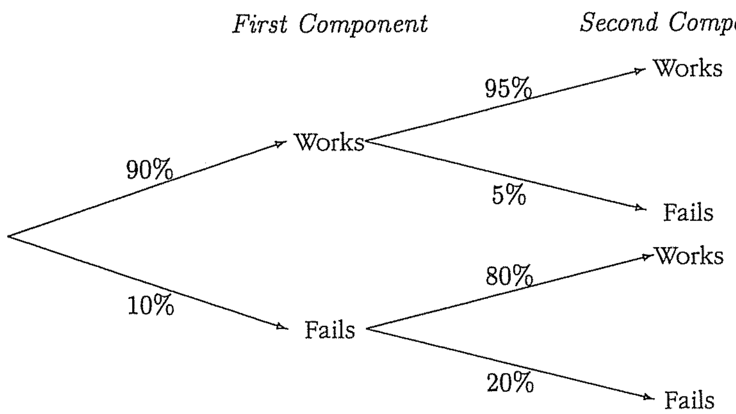
After setting up a tree diagram whose paths represent joint outcomes, the multiplication rule is used to define a distribution of probability over paths. The probability of each joint outcome represented by a path is obtained by multiplying the probability and conditional probability along the path.

Example 6. **Electrical components.**
Suppose there are two electrical components. The chance that the first component fails is 10%. If the first component fails, the chance that the second component fails is 20%. But if the first component works, the chance that the second component fails is 5%.

Problem. Calculate the probabilities of the following events:

1. at least one of the components works;
2. exactly one of the components works;
3. the second component works.

Solution. Here is the tree diagram showing all possible performances of the first and second components. Probabilities are filled in using the above data and the rule of complements.



By inspection of the diagram,

$$\begin{aligned}
P(\text{at least one works}) &= 1 - P(\text{both fail}) \\
&= 1 - 0.1 \times 0.2 = 0.98 \\
P(\text{exactly one works}) &= P(\text{first works and second fails}) \\
&\quad + P(\text{first fails and second works}) \\
&= 0.9 \times 0.05 + 0.1 \times 0.8 = 0.125 \\
P(\text{second works}) &= P(\text{first works and second works}) \\
&\quad + P(\text{first fails and second works}) \\
&= 0.9 \times 0.95 + 0.1 \times 0.8 = 0.935
\end{aligned}$$

Averaging Conditional Probabilities

The last two parts of the previous example illustrate a rule of average conditional probabilities: for any events A and B, the overall probability $P(A)$ is the average of the two conditional probabilities $P(A|B)$ and $P(A|B^c)$ with weights $P(B)$ and $P(B^c)$:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

In the example, B and B^c were (first works) and (first fails), while A was (exactly one works) in one instance, and (second works) in the other. The formula gives the probability of A as the sum of products of probabilities along paths leading to A in the tree diagram. The event B defines a partition of the whole outcome space Ω into two events B and B^c, corresponding to two initial branches in the tree. There is a similar formula for any partition $B_1, ..., B_n$ of the whole outcome space Ω, corresponding to n initial branches of a tree. For any event A the events $AB_1, ..., AB_n$ form a partition of A, so

$$P(A) = P(AB_1) + \cdots + P(AB_n)$$

by the addition rule. Applying the multiplication rule to each term gives

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)$$

This important result is summarized in the following box.

Rule of Average Conditional Probabilities

For a partition $B_1, \ldots, B_n$ of Ω,

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)$$

In words: the overall probability $P(A)$ is the weighted average of the conditional probabilities $P(A|B_i)$ with weights $P(B_i)$.

Example 7. **Sampling without replacement.**

Problem. Suppose two cards are dealt from a well-shuffled deck of 52 cards. What is the probability that the second card is black?

Solution. A common response to this question is that you can't say. It depends on whether the first card is black or not. If the first card is black, the chance that the second is black is 25/51, since no matter which black card the first one is, the second is equally likely to be any of the 51 remaining cards, and there are 25 black cards remaining. If the first card is red, the chance that the second is black is 26/51, by similar reasoning. These are the *conditional* probabilities of black on the second card given black and red, respectively, on the first card. But the question does not refer to the first card at all. The *overall* probability of black on the second card is the *average* of these conditional probabilities:

$$\begin{aligned} P(\text{second black}) &= P(\text{second black}|\text{first black})P(\text{first black}) \\ &\quad + P(\text{second black}|\text{first red})P(\text{first red}) \\ &= \frac{25}{51}\cdot\frac{1}{2} + \frac{26}{51}\cdot\frac{1}{2} = \left(\frac{25+26}{51}\right)\times\frac{1}{2} = \frac{1}{2} \end{aligned}$$

Discussion. This can also be argued by symmetry. Since there are equal numbers of black and red cards in the deck, the assumptions made at the start are symmetric with respect to black and red. This makes

$$P(\text{second black}) = P(\text{second red})$$

Since

$$P(\text{second black}) + P(\text{second red}) = 1$$

this gives the answer of 1/2. This argument shows just as well that if n cards are dealt, then $P(n\text{th card black}) = 1/2$, $P(n\text{th card an ace}) = 1/13$, and so on.

Independence

We have just seen that for any events A and B, $P(A)$ is the average of the conditional probabilities $P(A|B)$ and $P(A|B^c)$, weighted by $P(B)$ and $P(B^c)$. Suppose now that the chance of A does not depend on whether or not B occurs, and in either case equals p, say. In symbols:

$$P(A|B) = P(A|B^c) = p \tag{1}$$

Then also the unconditional probability of A is p:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = pP(B) + pP(B^c) = p$$

For example, A might be the event that a card dealt from a well-shuffled deck was an ace, B the event that a die showed a six. Such events A and B are called *independent*. Intuitively, independent events have no influence on each other. It would be reasonable to suppose that any event determined by a card dealt from a shuffled deck would be independent of any event determined by rolling a die. To be brief, the deal and the die roll would be called independent.

One more example: two draws at random from a population would be independent if done with replacement between draws, but *dependent* (i.e., not independent) if done without replacement.

Independence of events A and B can be presented mathematically in a variety of equivalent ways. For example, it was just shown that the definition (1) above (which assumes both $P(B) > 0$ and $P(B^c) > 0$), implies

$$P(A|B) = P(A) \tag{2}$$

A similar calculation shows that (2) implies (1). The formula $P(A|B) = P(AB)/P(B)$ shows (2) is equivalent to the following:

Multiplication Rule for Independent Events

$$P(AB) = P(A)P(B)$$

The multiplication rule is usually taken as the formal mathematical definition of independence, to include the case of events with probability 0 or 1. (Such an event is then, by definition, independent of every other event.)

The multiplication rule brings out the symmetry of independence. Assuming $P(A) > 0$, and using the fact that $AB = BA$ and $P(A)P(B) = P(B)P(A)$, the multiplication rule allows (2) to be turned around to

$$P(B|A) = P(B) \tag{3}$$

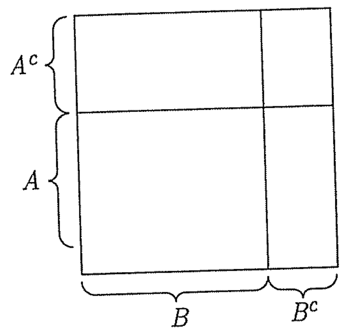
and (1) can be turned around similarly.

Assuming A and B are independent, all of these formulae hold also with either A^c substituted for A, B^c for B, or with both substitutions. This is obvious for (1), hence also true for the others. To spell out an example, since A splits into AB^c and AB,

$$\begin{aligned} P(AB^c) &= P(A) - P(AB) \\ &= P(A) - P(A)P(B) \quad \text{assuming the multiplication rule for } A \text{ and } B \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c) \quad \text{by the rule of complements.} \end{aligned}$$

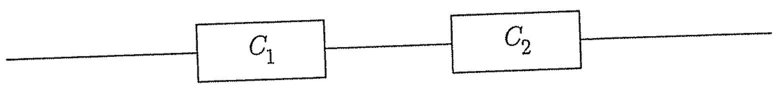
So the multiplication rule works just as well with B^c instead of B. The same goes for A^c instead of A.

Here the various probabilities determined by independent events A and B are illustrated graphically as proportions in a Venn diagram. Event A is represented by a rectangle lying horizontally, event B by a rectangle standing vertically.



Example 8. **Reliability of two components in series.**

A system consists of two components C_1 and C_2, each of which must remain operative for the overall system to function. The components C_1 and C_2 are then said to be connected in series, and represented diagrammatically as follows:



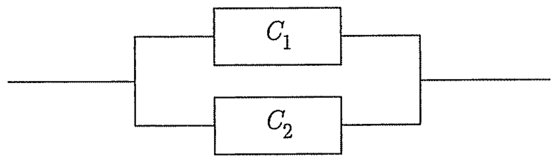
Let W_i be the event that component C_i works without failure for a given period of time, say one day. The event that the whole system operates without failure for one day is the event that both C_1 and C_2 operate without failure, that is, the event W_1W_2. The probabilities $P(W_1)$ and $P(W_2)$ are called the reliabilities of components C_1 and C_2. The probability $P(W_1W_2)$ is the reliability of the whole system. Suppose that the component reliabilities $P(W_1)$ and $P(W_2)$ are known from empirical data of past performances of similar components, say $P(W_1) = 0.9$ and $P(W_2) = 0.8$. If the particular components C_1 and C_2 have never been used together before, $P(W_1W_2)$ cannot be known empirically. But it may still be reasonable to assume that the events W_1 and W_2 are independent. Then the reliability of the whole system would be given by the formula

$$P(\text{system works}) = P(W_1W_2) = P(W_1)P(W_2) = 0.9 \times 0.8 = 0.72$$

Hopefully this number, 0.72, would give an indication of the long-run relative frequency of satisfactory performance of the system. But bear in mind that such a number is based on a theoretical assumption of independence which may or may not prove well founded in practice. The sort of thing which might prevent independence is the possibility of failures of both components due to a common cause, for example, voltage fluctuations in a power supply, the whole system being flooded, the system catching fire, etc. For the series system considered here such factors would tend to make the reliability $P(W_1W_2)$ greater than if W_1 and W_2 were independent, suggesting that the number, 0.72, would be too low an estimate of the reliability.

Example 9. **Reliability of two components in parallel.**

A method of increasing the reliability of a system is to put components in parallel, so the system will work if either of the components works. Two components C_1 and C_2 in parallel may be represented diagrammatically as follows:



Suppose, as in the last example, that the individual components C_1 and C_2 have reliabilities $P(W_1)$ and $P(W_2)$, where W_1 is the event that C_1 works. The event that the whole system functions is now the event $W_1 \cup W_2$ that either C_1 or C_2 works. The complementary event of system failure is the event F_1F_2 that both C_1 and C_2 fail, where F_i is the complement of W_i. Thus the reliability of the whole system is

$$P(\text{system works}) = P(W_1 \cup W_2) = 1 - P(F_1F_2)$$

If W_1 and W_2 are assumed independent, so are F_1 and F_2. In that case

$$P(\text{system works}) = 1 - P(F_1)P(F_2)$$

For example, if the component reliabilities are $P(W_1) = 0.9$ and $P(W_2) = 0.8$ as before, then $P(F_1) = 0.1$ and $P(F_2) = 0.2$, and the system reliability is

$$P(\text{system works}) = 1 - (0.1)(0.2) = 0.98$$

This is a considerable improvement over the reliability of the individual components. The assumption of independent failures must be viewed with particular suspicion in parallel systems, as it tends to lead to exaggerated estimates of system reliabilities. Suppose, for example, that all failures of component C_1 and half the failures of component C_2 occur due to severe voltage fluctuation in a power supply common to C_1 and C_2. Then F_1 is the event of a voltage fluctuation, and it should be assumed that $P(F_1|F_2) = 0.5$ instead of the independence assumption $P(F_1|F_2) = 0.1$. With the new assumptions,

$$P(F_1F_2) = P(F_2)P(F_1|F_2) = (0.2)(0.5) = 0.1$$

$$P(\text{system works}) = 1 - P(F_1F_2) = 0.9$$

As a general rule, failures of both components due to a common cause will tend to decrease the reliability of a parallel system below the value predicted by an independence assumption.

Exercises 1.4

1. In a particular population of men and women, 92% of women are right handed, and 88% of men are right handed. Indicate whether each of the following statements is (i) true, (ii) false, or (iii) can't be decided on the basis of the information given.
   a) The overall proportion of right handers in the population is exactly 90%.
   b) The overall proportion of right handers in the population is between 88% and 92%.
   c) If the sex ratio in the population is 1-to-1 then a) is true.
   d) If a) is true then the sex ratio in the population is 1-to-1.
   e) If there are at least three times as many women as men in the population, then the overall population of right handers is at least 91%.

2. A light bulb company has factories in two cities. The factory in city A produces two-thirds of the company's light bulbs. The remainder are produced in city B, and of these, 1% are defective. Among all bulbs manufactured by the company, what proportion are not defective and made in city B?

3. Suppose:
   $P(\text{rain today})=40\%$; $P(\text{rain tomorrow})=50\%$; $P(\text{rain today and tomorrow})=30\%$.
   Given that it rains today, what is the chance that it will rain tomorrow?

4. Two independent events have probabilities 0.1 and 0.3. What is the probability that
   a) neither of the events occurs?
   b) at least one of the events occurs?
   c) exactly one of the events occurs?

5. There are two urns. The first urn contains 2 black balls and 3 white balls. The second urn contains 4 black balls and 3 white balls. An urn is chosen at random, and a ball is chosen at random from that urn.
   a) Draw a suitable tree diagram.
   b) Assign probabilities and conditional probabilities to the branches of the tree.
   c) Calculate the probability that the ball drawn is black.

6. Suppose two cards are dealt from a deck of 52. What is the probability that the second card is a spade given that the first card is black?

7. Suppose A and B are two events with $P(A) = 0.5$, $P(A \cup B) = 0.8$.

a) For what value of $P(B)$ would A and B be mutually exclusive?

b) For what value of $P(B)$ would A and B be independent?

8. A hat contains a number of cards, with

30% white on both sides;

50% black on one side and white on the other;

20% black on both sides.

The cards are mixed up, then a single card is drawn at random and placed on the table. If the top side is black, what is the chance that the other side is white?

9. Three high schools have senior classes of size 100, 400, and 500, respectively. Here are two schemes for selecting a student from among the three senior classes:

A: Make a list of all 1000 seniors, and choose a student at random from this list.

B: Pick one school at random, then pick a student at random from the senior class in that school.

Show that these two schemes are not probabilistically equivalent. Here is a third scheme:

C: Pick school i with probability p_i ($p_1 + p_2 + p_3 = 1$), then pick a student at random from the senior class in that school.

Find the probabilities p_1, p_2, and p_3 which make scheme C equivalent to scheme A.

10. Suppose electric power is supplied from two independent sources which work with probabilities 0.4, 0.5, respectively. If both sources are providing power enough power will be available with probability 1. If exactly one of them works there will be enough power with probability 0.6. Of course, if none of them works the probability that there will be sufficient supply is 0.

a) What are the probabilities that exactly k sources work for $k = 0, 1, 2$?

b) Compute the probability that enough power will be available.

11. Assume identical twins are always of the same sex, equally likely boys or girls. Assume that for fraternal twins the firstborn is equally likely to be a boy or a girl, and so is the secondborn, independently of the first. Assume that proportion p of twins are identical, proportion $q = 1-p$ fraternal. Find formulae in terms of p for the following probabilities for twins:

a) P(both boys)

b) P(firstborn boy and secondborn girl)

c) P(secondborn girl | firstborn boy)

d) P(secondborn girl | firstborn girl).

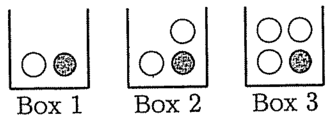
12. Give a formula for $P(F|G^c)$ in terms of $P(F)$, $P(G)$, and $P(FG)$ only.

1.5 Bayes' Rule

The rules of conditional probability, described in the last section, combine to give a general formula for updating probabilities called *Bayes' rule.* Before stating the rule in general, here is an example to illustrate the basic setup.

Example 1. **Which box?**

Suppose there are three similar boxes. Box i contains i white balls and one black ball, $i = 1, 2, 3$, as shown in the following diagram.



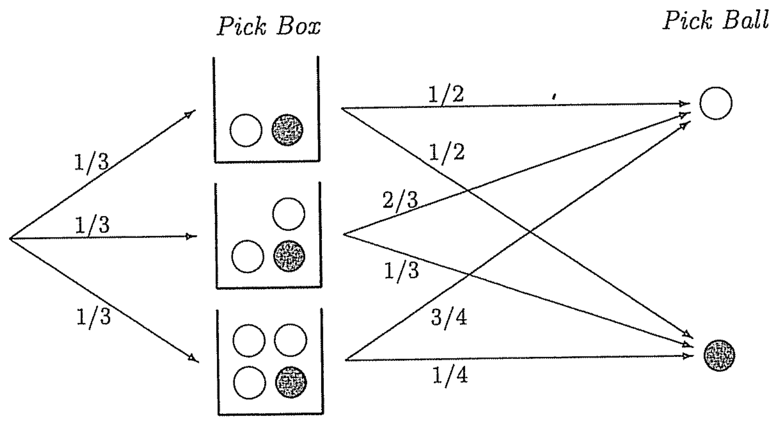
Suppose I mix up the boxes and then pick one at random. Then I pick a ball at random from the box and show you the ball. I offer you a prize if you can guess correctly what box it came from.

Problem. Which box would you guess if the ball drawn is white and what is your chance of guessing right?

Solution. An intuitively reasonable guess is Box 3, because the most likely explanation of how a white ball was drawn is that it came from a box with a large proportion of whites. To confirm this, here is a calculation of

$$P(\text{Box } i|\text{white}) = \frac{P(\text{Box } i \text{ and white})}{P(\text{white})} \qquad (i = 1, 2, 3) \qquad (*)$$

These are the chances that you would be right if you guessed Box i, given that the ball drawn is white. The following diagram shows the probabilistic assumptions:



From the diagram, the numerator in (*) is

$$P(\text{Box } i \text{ and white}) = P(\text{Box } i)P(\text{white}|\text{Box } i) = \frac{1}{3} \times \frac{i}{i+1} \quad (i = 1, 2, 3)$$

By the addition rule, the denominator in (*) is the sum of these terms over $i = 1, 2, 3$:

$$P(\text{white}) = \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times \frac{3}{4} = \frac{23}{36} \quad \text{and}$$

$$P(\text{Box } i|\text{white}) = \frac{\frac{1}{3} \times \frac{i}{i+1}}{\frac{23}{36}} = \frac{12}{23} \times \frac{i}{i+1} \qquad (i = 1, 2, 3)$$

Substituting for $i/(i+1)$ for $i = 1, 2, 3$ gives the following numerical results:

i	1	2	3
$P(\text{Box } i\|\text{white})$	6/23	8/23	9/23

This confirms the intuitive idea that Box 3 is the most likely explanation of a white ball. Given a white ball, the chance that you would be right if you guessed this box would be $9/23 \approx 39.13\%$.

Suppose, more generally, that events $B_1, \ldots, B_n$ represent n mutually exclusive possible results of the first stage of some procedure. Which one of these results has occurred is assumed unknown. Rather, the result A of some second stage has been observed, whose chances depend on which of the B_i's has occurred. In the previous example A was the event that a white ball was drawn and B_i the event that it came from a box with i white balls. The general problem is to calculate the probabilities of the events B_i given occurrence of A (called *posterior probabilities*), in terms of

(i) the unconditional probabilities $P(B_i)$ (called *prior probabilities*);

(ii) the conditional probabilities $P(A|B_i)$ (called *likelihoods*).

Here is the general calculation:

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)} \qquad \text{(multiplication rule)}$$

where, by the rule of average conditional probabilities, the denominator is

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)$$

which is the sum over $i = 1$ to n of the expression $P(A|B_i)P(B_i)$ in the numerator. The result of this calculation is called Bayes' rule.

, 3:

ite
ox

ive
has
een
ous
me
ties
of

tor.

Bayes' Rule

For a partition $B_1, \ldots, B_n$ of all possible outcomes,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)} \quad (i = 1, \ldots, n)$$

It is better not to try to memorize this formula, as it is easily derived from the basic rules of conditional probability which are easier to remember. Rather, understand the sequence of steps by which it is derived. These are the same steps used to solve the balls and boxes problem.

Example 2. **Which box? (continued).**

Consider again the same three boxes as in the previous example. Suppose I pick a box. Then I pick a ball at random from the box and show you the ball. I offer you a prize if you can guess correctly what box it came from.

Problem. Which box would you guess if the drawn ball is white, and what is your chance of guessing right?

Discussion. The wording of this problem is identical to the wording of Example 1 above, except that the sentence "Suppose I mix up the boxes and then pick one at random" has been replaced by "Suppose I pick a box". A naive approach to the new problem is to suppose it is the same as the old one, with the answer:

$$\text{guess Box 3, with probability of being right} = 9/23$$

But this makes an *implicit assumption* that I am equally likely to pick any one of the three boxes. And the problem cannot be solved without assuming some values π_i for the probabilities that I pick box $i, i = 1, 2, 3$. These probabilities π_i are called *prior probabilities* because they refer to your opinion about which box I picked, prior to learning the color of the ball drawn. Once you have assigned these prior probabilities $\pi_i, i = 1, 2, 3$, the previous calculations can be repeated. From the prior probabilities π_i and the probabilities $i/(i + 1)$ of getting the observed result, given box i (the *likelihoods*), you can obtain the *posterior probabilities* by Bayes' rule:

$$P(\text{Box } i|\text{white}) = \frac{\pi_i(\frac{i}{i+1})}{\pi_1 \times \frac{1}{2} + \pi_2 \times \frac{2}{3} + \pi_3 \times \frac{3}{4}}$$

Thus, given that a white ball was drawn, to maximize your chance of guessing correctly you should guess box i for whichever i maximizes $\pi_i \left(\frac{i}{i+1}\right)$. Which i this is depends on the π_i. The probabilities in question are now clearly a matter of your *opinion* about how I picked the box. There remains the problem of how to assign the prior probabilities π_i. This is a tricky business, as it depends on psychological factors, such as whether or not you think I am deliberately trying to make it hard for you to guess, and if so what strategy you think I'm using. For further analysis, see Exercises 1.5.7 and 1.5.8.

In principle, every application of Bayes' rule is such as the above examples of guessing the box that produced a particular color of ball. There is always the problem of deciding what the prior probabilities should be. Most often the prior probabilities will only make sense in a subjective interpretation of probability. But in problems like the next example (false positives) the prior probabilities may be known as population proportions. This example is like a scheme with two boxes D and D^c:

Box D containing 95% balls labeled + and 5% labeled −

Box D^c containing 2% balls labeled + and 98% labeled −

If box D has prior probability 1%, and a draw from the box yields a +, what is the chance that the + came from box D? As the solution shows, such extremely skewed priors and likelihoods may lead to surprising conclusions.

Example 3. False positives.

Problem. Suppose that a laboratory test on a blood sample yields one of two results, positive or negative. It is found that 95% of people with a particular disease produce a positive result. But 2% of people without the disease will also produce a positive result (a *false positive*). Suppose that 1% of the population actually has the disease. What is the probability that a person chosen at random from the population will have the disease, given that the person's blood yields a positive result?

Solution. Let $P(F)$ denote the proportion of people in the population with characteristic F. Then $P(F|G)$ is the proportion of those in the population with characteristic G who also have characteristic F. The desired probability is $P(D|+)$ where D indicates the disease, and + indicates a positive test result. The data in the problem indicate that

$$P(+|D) = 0.95, \;\; P(+|D^c) = 0.02, \;\; P(D) = 0.01, \;\; P(D^c) = 0.99.$$

Applying Bayes' rule with $A = +$, $B_1 = D$, $B_2 = D^c$, gives

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{(.95)(.01)}{(.95)(.01) + (.02)(.99)} \\ &= \frac{95}{293} \approx 32\% \end{aligned}$$

Discussion. Thus only 32% of those persons who produce a positive test result actually have the disease. At first this result seems surprisingly low. The point is that because the disease is so rare, the number of true positives coming from the few people with the disease is comparable to the number of false positives coming from the many without the disease.

Interpretation of conditional probabilities. In applications of Bayes' rule it is important to keep in mind the interpretation of the various probabilities involved. Typically, the likelihoods $P(A|B_i)$ will admit a long-run frequency interpretation. If the prior probabilities $P(B_i)$ also have a long-run frequency interpretation, then so too will the conditional probability $P(B_i|A)$ given by Bayes' formula. In Example 3 there were two hypotheses $B_1 = D$ that a person was diseased and $B_2 = D^c$ that a person was not. The observed event was the event $A = +$ of a positive laboratory test. There the conditional probability $P(D|+)$ admitted an empirical interpretation, as that proportion of individuals in the population in question showing a positive test who actually had the disease. This conditional probability also admits a long-run frequency interpretation in terms of repeated sampling of that population, or some other population with the same characteristics assumed in the calculations. Among persons who produce a positive laboratory test, the long-run proportion with the disease will most likely be close to $P(D|+) \approx 32\%$.

There are many situations, however, where it is impossible to give a long-run frequency interpretation to the prior probabilities $P(B_i)$. The same must then be said of the posterior probabilities $P(B_i|A)$ which are calculated in terms of them, even if the likelihoods $P(A|B_i)$ have long-run frequency interpretations.

Calculations by Bayes' rule can often be simplified by noting that it is only the ratios $P(B_i)$ to $P(B_j)$ (the *prior odds ratios*) and the ratios $P(A|B_i)$ to $P(A|B_j)$ (the *likelihood ratios*) which matter. As you can check as an exercise, if the prior odds ratios are written as, say, R_i to R_j, and the likelihood ratios as, say, L_i to L_j, meaning that

$$P(B_i) = cR_i \qquad \text{for some constant } c$$

and

$$P(A|B_i) = dL_i \qquad \text{for some constant } d$$

then the *posterior odds ratios* $P(B_i|A)$ to $P(B_j|A)$ are simply R_iL_i to R_jL_j, and

$$P(B_i|A) = \frac{R_iL_i}{R_1L_1 + \cdots + R_nL_n}$$

This is summarized by the following:

Bayes' Rule for Odds

posterior odds = prior odds × likelihoods.

Bayes' rule for odds shows clearly how the prior odds are just as important a factor as the likelihood ratio in computing the posterior odds. If the prior odds don't make sense in terms of long-run frequencies, neither will the posterior odds.

But even if the probabilities don't admit a long-run frequency interpretation, you might find it useful to regard the probabilities in Bayes' rule as subjective probabilities. Bayes' rule then dictates how opinions should be revised in the light of new information, to be consistent with the rules of probability. Here is a typical example.

Example 4. Diagnosis of a particular patient.

Problem. Suppose a doctor is examining a patient from the population in Example 3. This patient was not chosen at random. He walked into the doctor's office because he was feeling sick. After examining the patient, but not seeing the result of the blood test, the doctor's opinion is that there is a 30% chance that the patient has the disease. How should the doctor revise her opinion after seeing a positive blood test?

Solution. To be consistent with the rules of probability, the doctor should use Bayes' rule. Now the prior probabilities are

$$P(D) = 30\%, \quad P(D^c) = 70\%$$

while it might be reasonable to suppose that the likelihoods

$$P(+|D) = 95\%, \quad P(+|D^c) = 2\%$$

are the same as before. The posterior probability can be calculated as before, using Bayes' rule, but with the new prior probabilities. In terms of odds, the prior odds in favor of the disease are 3 to 7, the likelihood ratio in favor of the disease is 95 to 2, so the posterior odds in favor are 3×95 to 7×2, or 285 to 14. So given the positive blood test result, the doctor should revise her opinion and say that the patient has the disease with probability

$$\frac{285}{285+14} = \frac{285}{299} = 0.95317$$

Discussion. Notice how working with prior odds of 30 to 70 instead of 1 to 99 has a drastic effect on the conclusion. Provided the prior odds are not heavily against the disease, the evidence of the blood test carries a lot of weight. The likelihood ratio of 95 to 2 overwhelms the doctor's prior odds of 3 to 7, so there should be little doubt left in the doctor's mind after seeing the positive blood test. The puzzling question in this kind of application is how does the doctor come up with the odds of 3 to 7 after the medical examination? To come up with such odds, the doctor must make an intuitive judgment based on the whole complex of evidence gained from an examination of the patient. It seems impossible to adequately formalize this process mathematically. The theory does not help the doctor come up with a prior opinion, or explain how the doctor should revise an opinion in the light of complex information such as is gained from a medical examination. All the theory can do in this context is to suggest how an opinion should be revised in the light of a single additional piece of information, such as the result of a blood test.

Notice how the terms prior and posterior are relative terms, like today and tomorrow. The posterior distribution after today's test will be the prior distribution for tomorrow's test. So an opinion can be revised repeatedly using Bayes' rule. At each stage in this process, all probabilities should be computed conditionally on everything that has gone before.

Exercises 1.5

1. There are two boxes, the odd box containing 1 black marble and 3 white marbles, and the even box containing 2 black marbles and 4 white marbles. A box is selected at random, and a marble is drawn at random from the selected box.

a) What is the probability that the marble is black?

b) Given the marble is white, what is the probability that it came from the even box?

2. **Polya's urn scheme.** An urn contains 4 white balls and 6 black balls. A ball is chosen at random, and its color noted. The ball is then replaced, along with 3 more balls of the same color (so that there are now 13 balls in the urn). Then another ball is drawn at random from the urn.

a) Find the chance that the second ball drawn is white. (Draw an appropriate tree diagram.)

b) Given that the second ball drawn is white, what is the probability that the first ball drawn is black?

c) Suppose the original contents of the urn are w white and b black balls, and that after a ball is drawn from the urn, it is replaced along with d more balls of the same color. In part a), w was 4, b was 6, and d was 3. Show that the chance that the second ball drawn is white is $\frac{w}{w+b}$. [Note that the probability above does not depend on the value of d.]

3. A manufacturing process produces integrated circuit chips. Over the long run the fraction of bad chips produced by the process is around 20%. Thoroughly testing a chip to determine whether it is good or bad is rather expensive, so a cheap test is tried. All good chips will pass the cheap test, but so will 10% of the bad chips.

a) Given a chip passes the cheap test, what is the probability that it is a good chip?

b) If a company using this manufacturing process sells all chips which pass the cheap test, over the long run what percentage of chips sold will be bad?

4. A digital communications system consists of a transmitter and a receiver. During each short transmission interval the transmitter sends a signal which is to be interpreted as a zero, or it sends a different signal which is to be interpreted as a one. At the end of each interval, the receiver makes its best guess at what was transmitted. Consider the events:

$$T_0 = \{\text{Transmitter sends } 0\}, \quad R_0 = \{\text{Receiver concludes that a 0 was sent}\},$$
$$T_1 = \{\text{Transmitter sends } 1\}, \quad R_1 = \{\text{Receiver concludes that a 1 was sent}\}.$$

Assume that $P(R_0|T_0) = 0.99$, $P(R_1|T_1) = 0.98$, and $P(T_1) = 0.5$. Find:

a) the probability of a transmission error given R_1;

b) the overall probability of a transmission error.

c) Repeat a) and b) assuming $P(T_1) = 0.8$ instead of 0.5.

5. **False diagnosis.** The fraction of persons in a population who have a certain disease is 0.01. A diagnostic test is available to test for the disease. But for a healthy person the chance of being falsely diagnosed as having the disease is 0.05, while for someone with the disease the chance of being falsely diagnosed as healthy is 0.2. Suppose the test is performed on a person selected at random from the population.

   a) What is the probability that the test shows a positive result (meaning the person is diagnosed as diseased, perhaps correctly, perhaps not)?

   b) What is the probability that the person selected at random is one who has the disease but is diagnosed healthy?

   c) What is the probability that the person is correctly diagnosed and is healthy?

   d) Suppose the test shows a positive result. What is the probability that the person tested actually has the disease?

   e) Do the above probabilities admit a long-run frequency interpretation? Explain.

6. An experimenter observes the occurrence of an event A as the result of a particular experiment. There are three different hypotheses, H_1, H_2, and H_3, which the experimenter regards as the only possible explanations of the occurrence of A. Under hypothesis H_1, the experiment should produce the result A about 10% of the time over the long run, under H_2 about 1% of the time, and under H_3 about 39% of the time. Having observed A, the experimenter decides that H_3 is the most likely explanation, and that the probability that H_3 is true is

$$\frac{39\%}{10\% + 1\% + 39\%} = 78\%.$$

   a) What assumption is the experimenter implicitly making?

   b) Does the probability 78% admit a long-run frequency interpretation?

   c) Suppose the experiment is a laboratory test on a blood sample from an individual chosen at random from a particular population. The hypothesis H_i is that the individual's blood is of some particular type i. Over the whole population it is known that the proportion of individuals with blood of type 1 is 50%, the proportion with type 2 blood is 45%, and the remaining proportion is type 3. Revise the experimenter's calculation of the probability of H_3 given A, so that it admits a long-run frequency interpretation. Is H_3 still the most likely hypothesis given A?

7. **Guessing what box.** Consider a game as in Examples 1 and 2, where I pick one of the three boxes, then you guess which box I picked after seeing the color of a ball drawn at random from the box. Then you learn whether your guess was right or wrong. Suppose we play the game over and over, replacing the ball drawn and mixing up the balls between plays. Your objective is to guess the box correctly as often as possible.

a) Suppose you know that I pick a box each time at random (probability $1/3$ for each box). And suppose you adopt the strategy of guessing the box with highest posterior probability given the observed color, as described in Example 1, in case the observed color is white. About what proportion of the time do you expect to be right over the long run?

b) Could you do any better by another guessing strategy? Explain.

c) Suppose you use guessing strategy found in a), but I was in fact randomizing the choice of the box each time, with probabilities $(1/2, 1/4, 1/4)$ instead of $(1/3, 1/3, 1/3)$. Now how would your strategy perform over the long run?

d) Suppose you knew I was either randomizing with probabilities $(1/3, 1/3, 1/3)$, or with probabilities $(1/2, 1/4, 1/4)$. How could you learn which I was doing? How should you respond, and how would your response perform over the long run?

8. Optimal strategies for guessing what box. (Continuation of Exercise 7, due to David Blackwell.) The question now arises: What randomizing strategy should I use to make it as hard as possible for you to guess correctly? Consider what happens if I use the $(\frac{6}{23}, \frac{9}{23}, \frac{8}{23})$ strategy, and answer the following questions:

a) What box should you guess if you see a black ball?

b) What box should you guess if you see a white ball?

c) What is your overall chance of winning?

You should conclude that with this strategy, your chance of winning is at most $\frac{9}{23}$, no matter what you do. Moreover, you have a strategy which guarantees you this chance of winning, no matter what randomization I use. It is the following:
If black, guess 1 with probability $\frac{18}{23}$, 2 with probability $\frac{5}{23}$, and 3 with probability 0.
If white, guess 1 with probability 0, 2 with probability $\frac{11}{23}$, and 3 with probability $\frac{12}{23}$.

d) Check that using this strategy, you win with probability $\frac{9}{23}$, no matter what box I pick.

According to the above analysis, I can limit your chance of winning to $\frac{9}{23}$ by a good choice of strategy, and you can guarantee that chance of winning by a good choice of strategy. The fraction $\frac{9}{23}$ is called the *value* of the above game, where it is understood that the payoff to you is 1 for guessing correctly, 0 otherwise. Optimal strategies of the type discussed above and a resulting value can be defined for a large class of games between two players called zero-sum games. For further discussion consult books on game theory.

9. A box contains three "shapes", as described in Example 1.3.3. One of the shapes is a fair die, and lands flat with probability $1/3$. The other two shapes land flat with probabilities $1/2$ and $2/3$, respectively.

a) One of the three shapes will be chosen at random, and rolled. What is the chance that the number rolled is 6?

b) Given that the number rolled is 6, what is the chance that the fair die was chosen?