Chapter 5 Probability
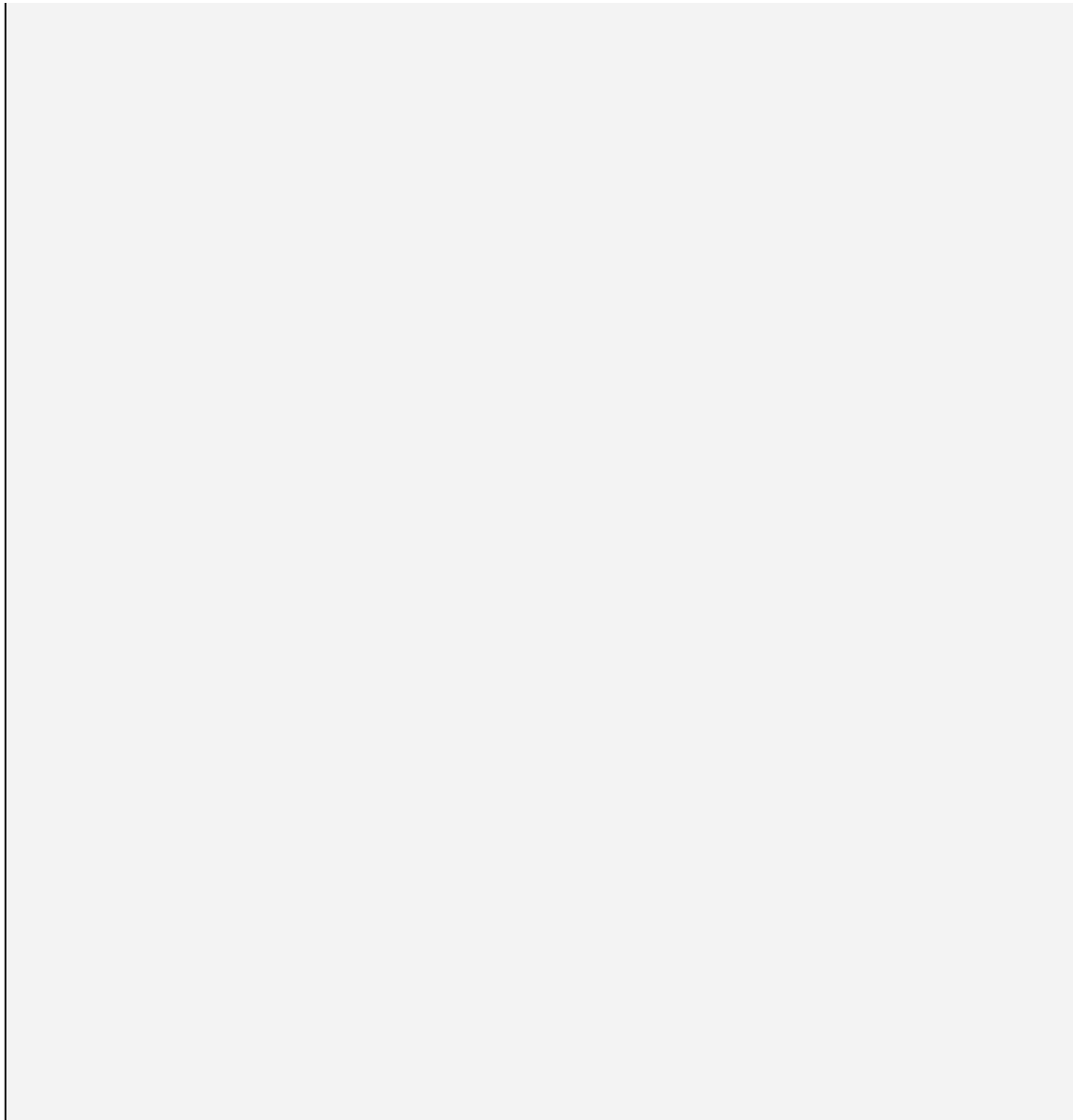
White-breasted nuthatch

Description

-

The concept of probability is important in almost every field of science, including biology. Probability is the backbone of data analysis. In Chapter 4, we made statements about probability to quantify the uncertainty of parameter estimates, and we will see even more uses of probability throughout the book.

Probability is essential to biology because we almost always look at the natural world by way of a sample, and as we have seen, chance can play a major role in the properties of samples. In this chapter, we will discuss the

basic principles of probability and basic probability calculations. In [Chapter 6](), we will begin to apply these concepts to data analysis.

5.1 The probability of an event

Imagine that you have 1000 songs on your phone, each of them recorded exactly once. When you push the "shuffle" button, the phone plays a song at random from the list of 1000. The probability that the first song played is your single favorite song is $1/1000$, or 0.001. The probability that the first song played is not your favorite song is $999/1000$, or 0.999. What exactly do these numbers mean?

The concept of probability rests on the notion of a **random trial**. A random trial is a process or experiment that has two or more possible outcomes whose occurrence cannot be predicted with certainty. Only one outcome is observed from each repetition of a random trial. In the phone songs example, a random trial consists of pushing the shuffle button once. The specified outcome is "your favorite song is played," which is one of 1000 possible outcomes. Other examples of random trials include

- Flipping a coin to see if heads or tails comes up,
- Rolling a pair of dice to see what the sum of their numbers is, and
- Rolling a die 10 times to measure the proportion of times a "6" comes up.

What is probability? To answer this, we need to define the **event** of interest and the list of all possible **outcomes** of a random trial. An event is any potential subset of all the possible outcomes. For example, there are six possible outcomes if we roll a six-sided die—the numbers one through six. The event of interest could be "the result is an even number," "the result is a number greater than three," or even the simple event "the result is four." As the last example shows, an outcome is also an event. However, events can be more complicated subsets of outcomes, so we will define principles of probability mainly in terms of events.

The **probability** of an event is the *proportion* of all random trials in which the specified event occurs when the same random process is repeated over and

over again independently and under the same conditions.[1] In an infinite number of random trials carried out in exactly the same way, the probability of an event is the fraction of the trials in which the event occurs.

> The *probability* of an event is the proportion of times the event would occur if we repeated a random trial over and over again under the same conditions. Probability ranges between zero and one.

A useful shorthand is the following:

$\Pr[A]$ means "the probability of event A."

Thus, if we want to state the probability of "rolling a four" with a six-sided die, then we can write

$\Pr[rolling\ a\ four] = 1/6.$

Because probabilities are proportions, they must always fall between zero and one, inclusive. An event has probability zero if it *never* happens, and an event has probability one if it *always* happens.

Flipping coins and rolling dice are not biological processes, but they are simple and familiar and the probabilities are known. Their relevance to biology is nevertheless high because they mimic the process of sampling. Randomly sampling 10 new babies and counting the number that are female is mimicked by flipping a coin 10 times and counting the number of heads (assuming both have probability $1/2$). Randomly sampling 100 individuals from a human population and counting the proportion that are left-handed is mimicked by rolling a six-sided die 100 times and counting the proportion of sixes (assuming that a "six" and a "left-handed person" both have probability $1/6$).

In other words, randomly sampling a population represents a random trial just like rolling a die or flipping a coin. The value of a variable measured on a

randomly sampled individual is an outcome of a random trial. The following are therefore also random trials:

- Randomly sampling an individual from a population of sockeye salmon to see what its weight is, and
- Randomly sampling 1000 fetuses from clinics in a large North American city to measure the proportion that have Down syndrome.

5.2 Venn diagrams

One useful way to think about the probabilities of events is with a graphical tool called a **Venn diagram**. <u>The area of the diagram represents all possible outcomes of a random trial</u>, and we can represent various events as areas within the diagram. The probability of an event is proportional to the area it occupies in the diagram.

Figure 5.2-1 shows a Venn diagram for one roll of a fair six-sided die. The six possible outcomes fill the diagram, indicating that these are all possible results. The area of the box for each outcome is the same, showing that these outcomes are equally probable in this particular example. They each contain $1/6$ of the area of the Venn diagram.
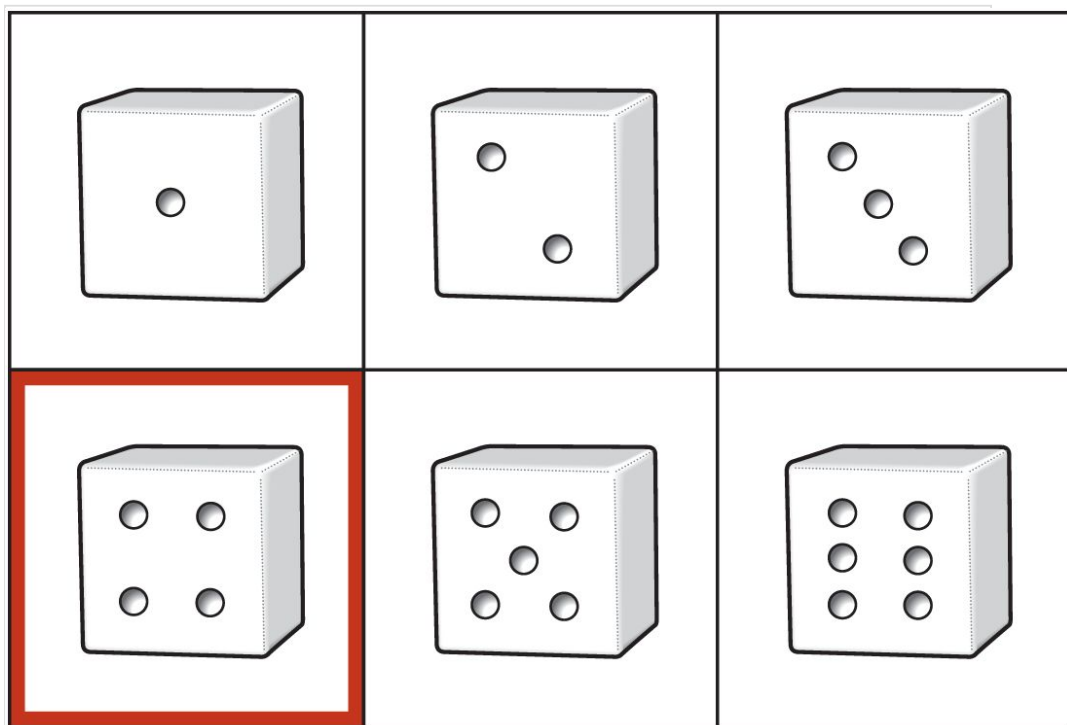
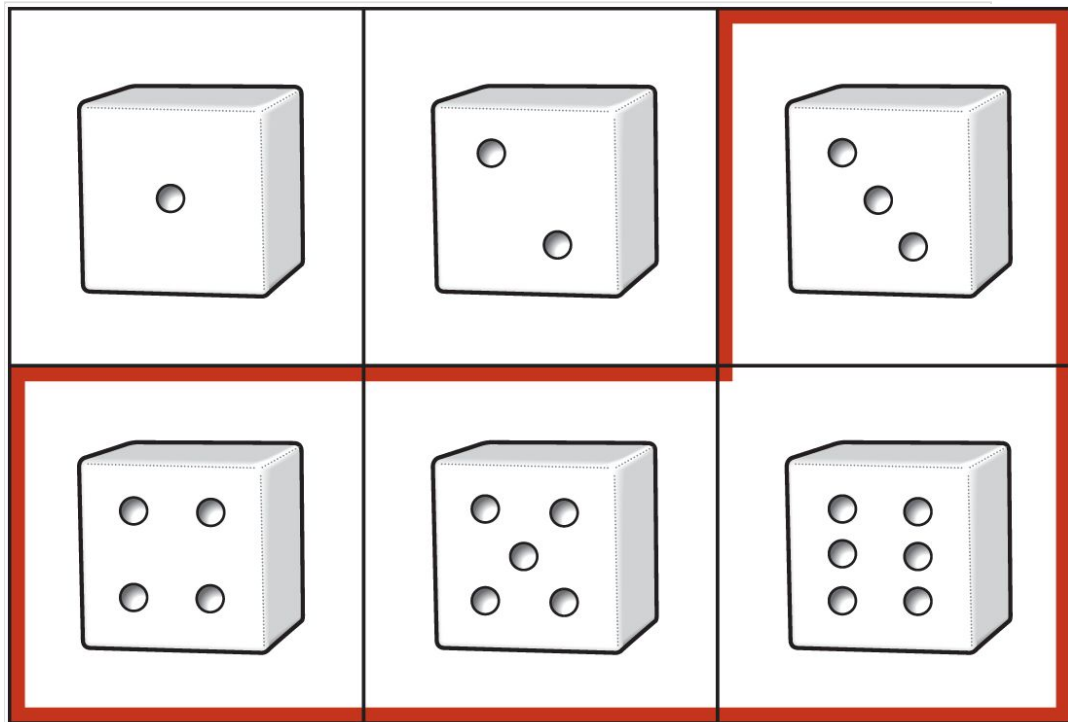Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.2-1

A Venn diagram for the possible outcomes of a roll of a six-sided die. The area corresponding to the event "the result is a four" is highlighted in red.

Description

-

We can use Venn diagrams to show more complicated events as well. In Figure 5.2-2, for example, the event "the result is greater than two" is shown.
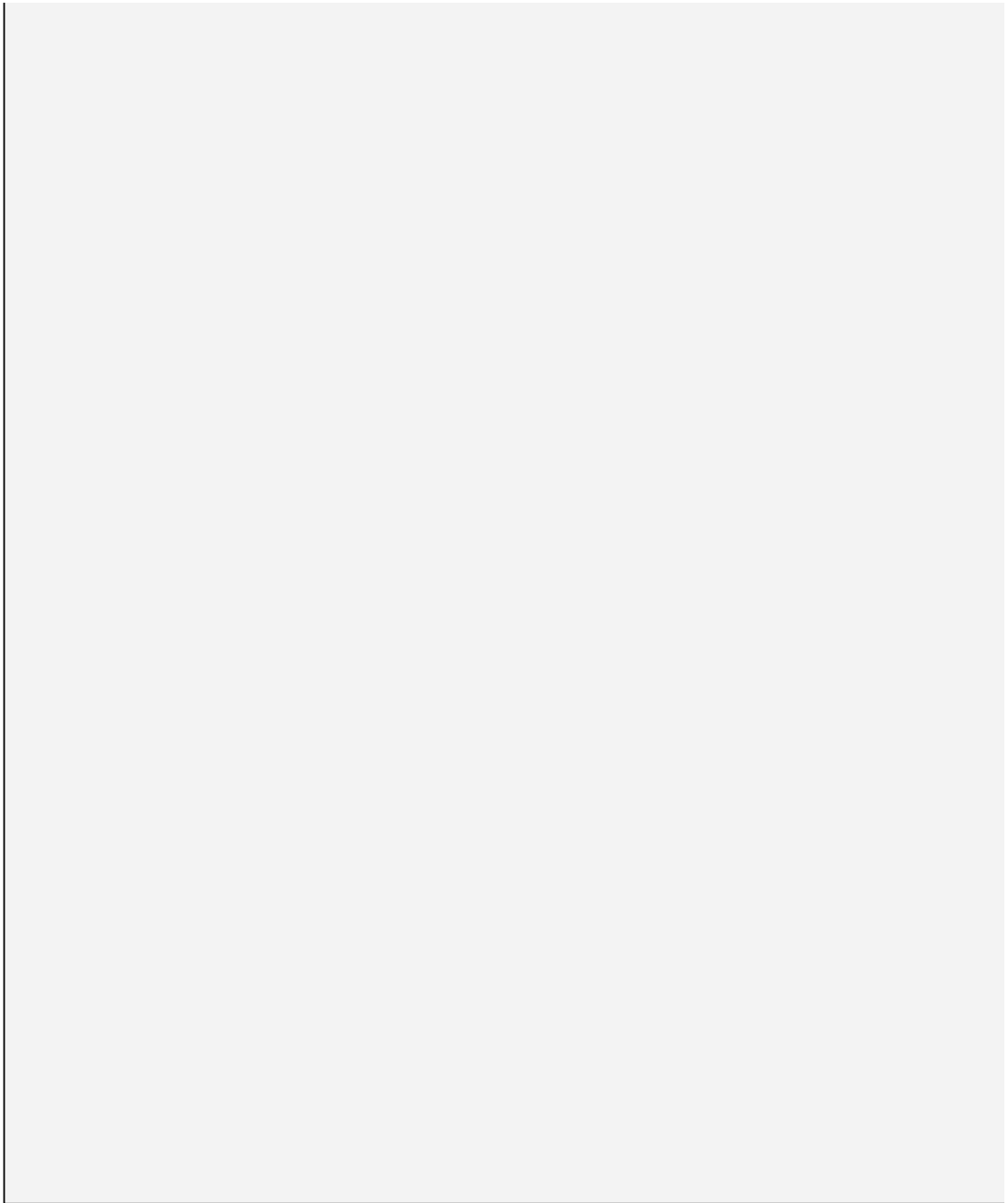

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.2-2

A Venn diagram showing the event "the result is greater than two" highlighted in red. The probability of this event is 4/6=2/3 $4/6 = 2/3$, equal to the area of the red region.

Description

-

5.3 Mutually exclusive events

When two events cannot both occur at the same time, we say that they are **mutually exclusive**. For example, a single die rolled once cannot yield both a one and a six. The events "one" and "six" are mutually exclusive events for this random trial.

> Two events are *mutually exclusive* if they cannot both occur at the same time.

Sometimes physical constraints explain why certain events are mutually exclusive. It is impossible, for example, for more than one number to result from a single roll of a die. Sometimes events are mutually exclusive because they never occur simultaneously in nature. For example, "has teeth" and "has feathers" are mutually exclusive events when we randomly sample a single living animal species from a list of all existing animal species, because no living animals have both teeth and feathers. If we sample a living animal species at random, the probability that it has both teeth and feathers is zero, although plenty of animals have teeth *or* feathers.

In mathematical terms, two events A and B are mutually exclusive if

$$\Pr[A \text{ and } B] = 0.$$

Here, $\Pr[A \text{ and } B]$ means the probability that both A and B occur.
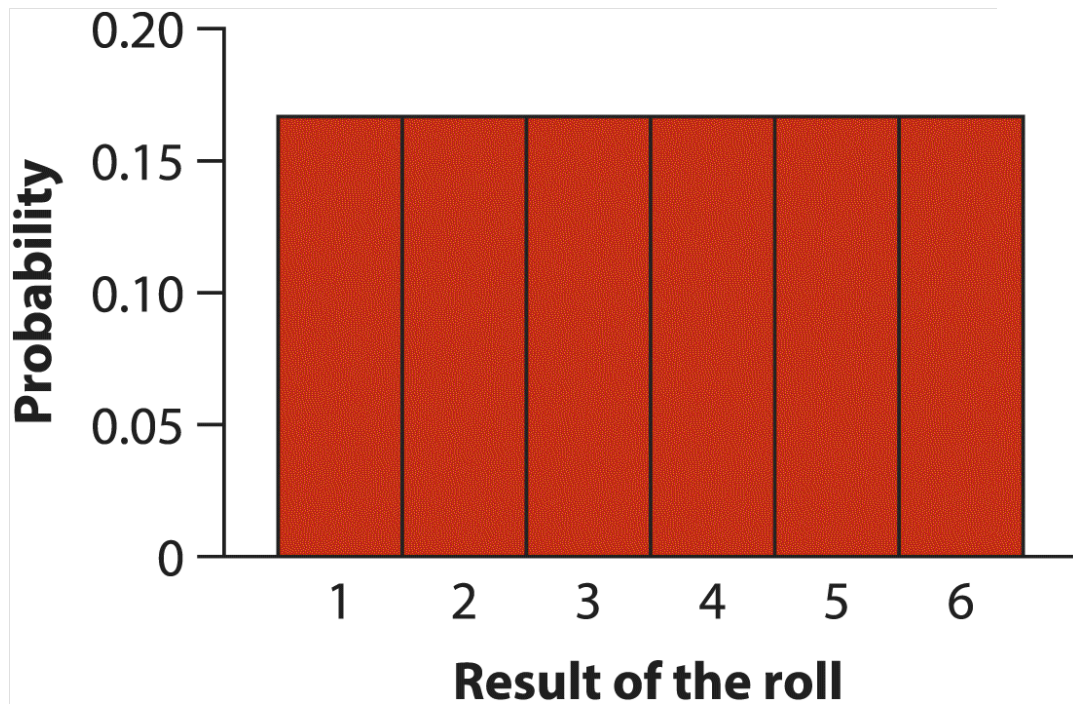
5.4 Probability distributions

A **probability distribution** describes the probabilities of each of the possible outcomes of a random trial. Some probability distributions can be described mathematically, while others are just a list of the possible outcomes and their probabilities. The precise meaning of a probability distribution depends on whether the variable is discrete or continuous.

> A *probability distribution* is a list of the probabilities of all mutually exclusive outcomes of a random trial.

Discrete probability distributions

A discrete numerical variable is measured in indivisible units. Categorical variables are discrete, as are many numerical variables. A discrete probability distribution gives the probability of each possible value of a discrete variable. Categorical and discrete numerical variables have discrete probability distributions. For example, the probability distribution of outcomes for the single roll of a fair die is given in Figure 5.4-1. In this case, all integers between one and six are equally probable outcomes (probability =1/6=0.167). $(\text{probability} = 1/6 = 0.167)$. The histogram in Figure 5.4-2 shows the probability distribution for the sum of the two numbers resulting from a roll of two dice. Here the different outcomes are *not* equally probable.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.4-1

The probability distribution of outcomes resulting from the roll of a single six-sided fair die. The probability of each possible outcome is 1/6=0.167.
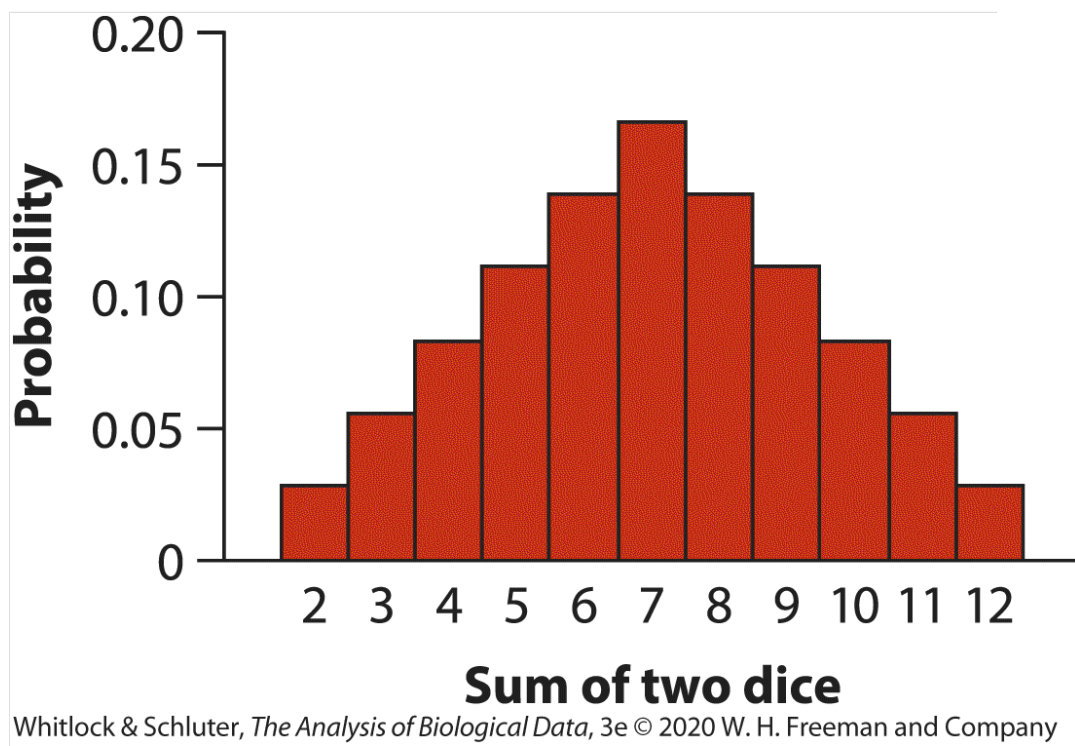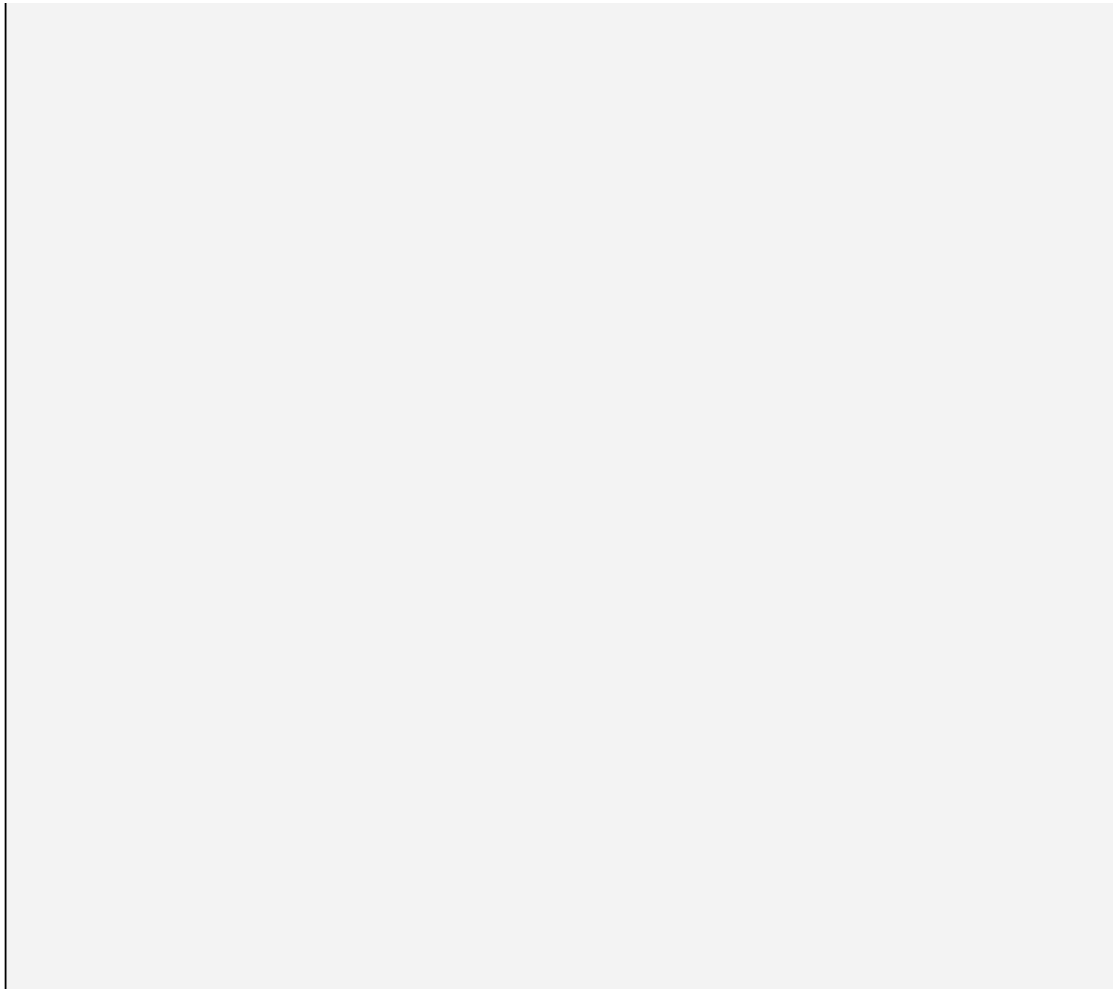$1/6 = 0.167.$

Description

-

Sum of two dice

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Description

The horizontal axis is labeled Sum of two dice, ranging from 2 to 12 with increments of 2. The vertical axis is labeled as Probability, ranging from 0 to 0 point 2 with increments of 0 point zero five. The approximate data are as follows. 2, zero point zero three; 3, zero point zero six; 4, zero point zero eight; 5, zero point one two; 6, zero point one four; 7, zero point one seven; 8, zero point one four; 9, zero point one two; 10, zero point zero eight; 11, zero point zero six; 12, zero point zero three.

Because all possible outcomes are taken into account, the sum of all probabilities in a probability distribution must add to one. This is because a probability distribution has to describe all possible outcomes, and the probability that *some* outcome occurs from a random trial is one.

Continuous probability distributions

Unlike discrete variables, continuous numerical variables can take on any real number value within some range. Between any two values of a continuous variable, an infinite number of other values are possible. We describe a continuous probability distribution with a curve whose height is **probability density**. A probability density allows us to describe the probability of any range of values for the variable.

The normal distribution, first introduced in Section 1.4, is a continuous probability distribution. It is bell-shaped like the curve shown in Figure 5.4-3. We'll see much more of this distribution in Chapter 10 and later.
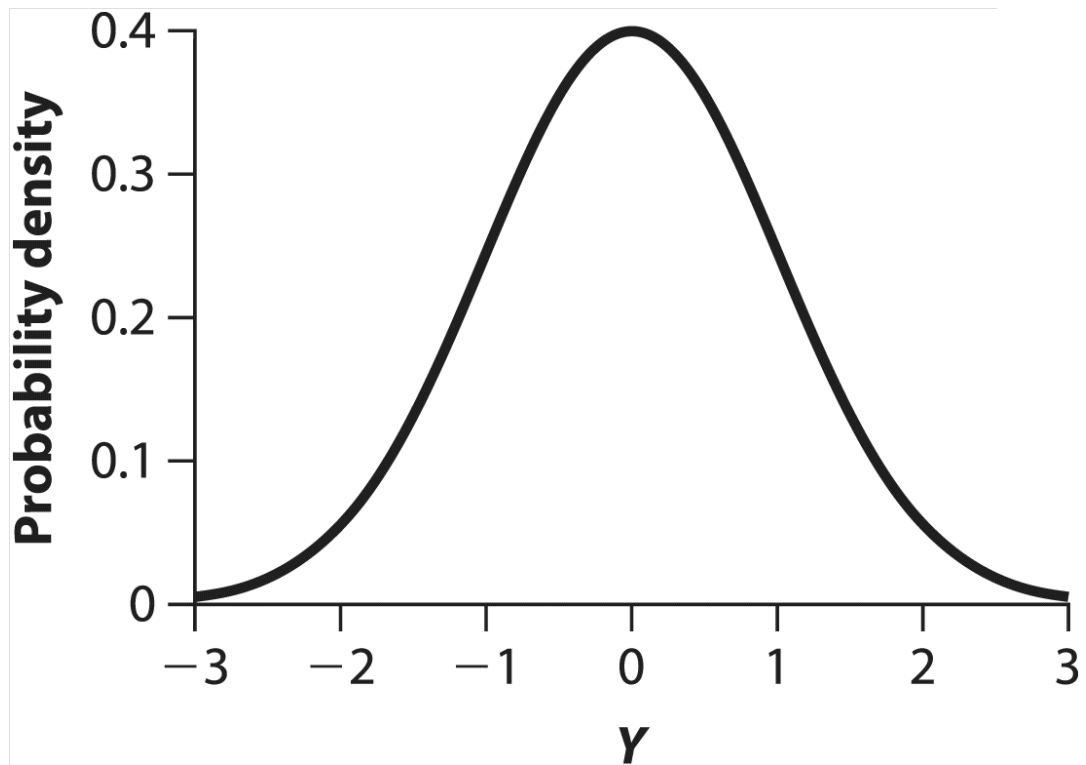
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.4-3

A normal distribution.

Description

The horizontal axis is marked from negative 3 to 3 with increment of 1. The vertical axis is labeled Probability density, marked from 0 to zero point four with increments of zero point one. A normal distribution curve starts from (minus 3, 0), peaks at (0, 0 point 4), slides down, and ends at (3, 0).
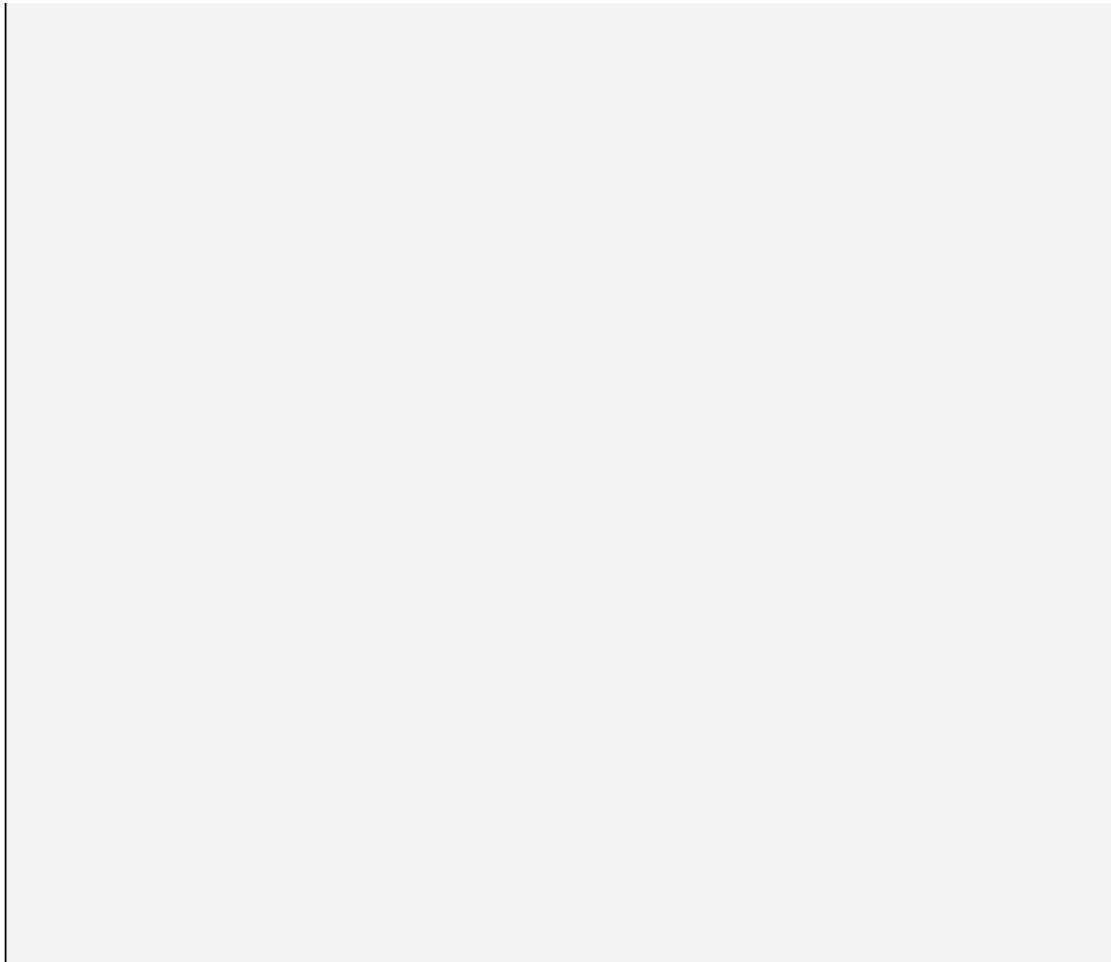
Imagine that we sample a random number from this distribution—let's call the number Y. Unlike discrete probability distributions, the height of a continuous probability curve at the value of $Y = 2.4$ does not give the probability of obtaining $Y = 2.4$. Because a continuous probability distribution covers an infinite number of possible outcomes, the probability of obtaining any specific outcome is infinitesimally small and therefore zero.

With continuous probability distributions, such as the normal curve, it makes more sense to talk about the probability of obtaining a value of Y that falls within some range. The probability of obtaining a value of Y within some range is indicated by the area under the curve. For example, the probability that a single randomly chosen individual has a

measurement lying between the two numbers aa and bb equals the area under the curve between aa and bb ([Figure 5.4-4](#)).
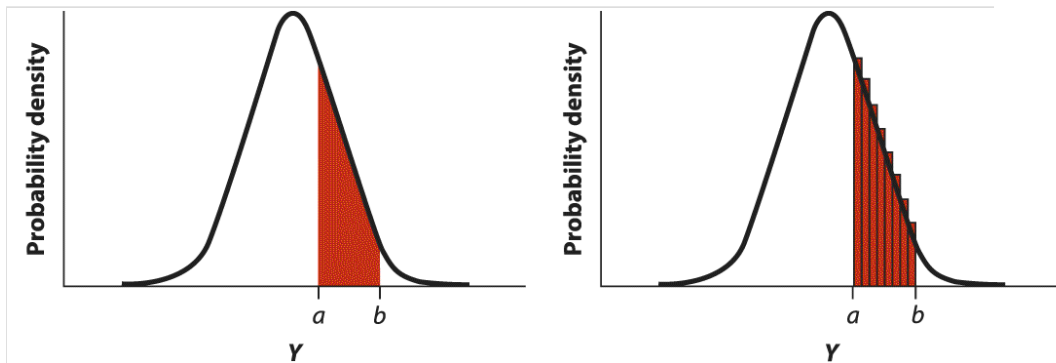
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.4-4

The probability that a randomly chosen Y-measurement$Y$-measurement lies between aa and bb is the area under the probability density curve between aa and bb (*left panel*). In the right panel we approximate the same area using discrete bars.

Description

-
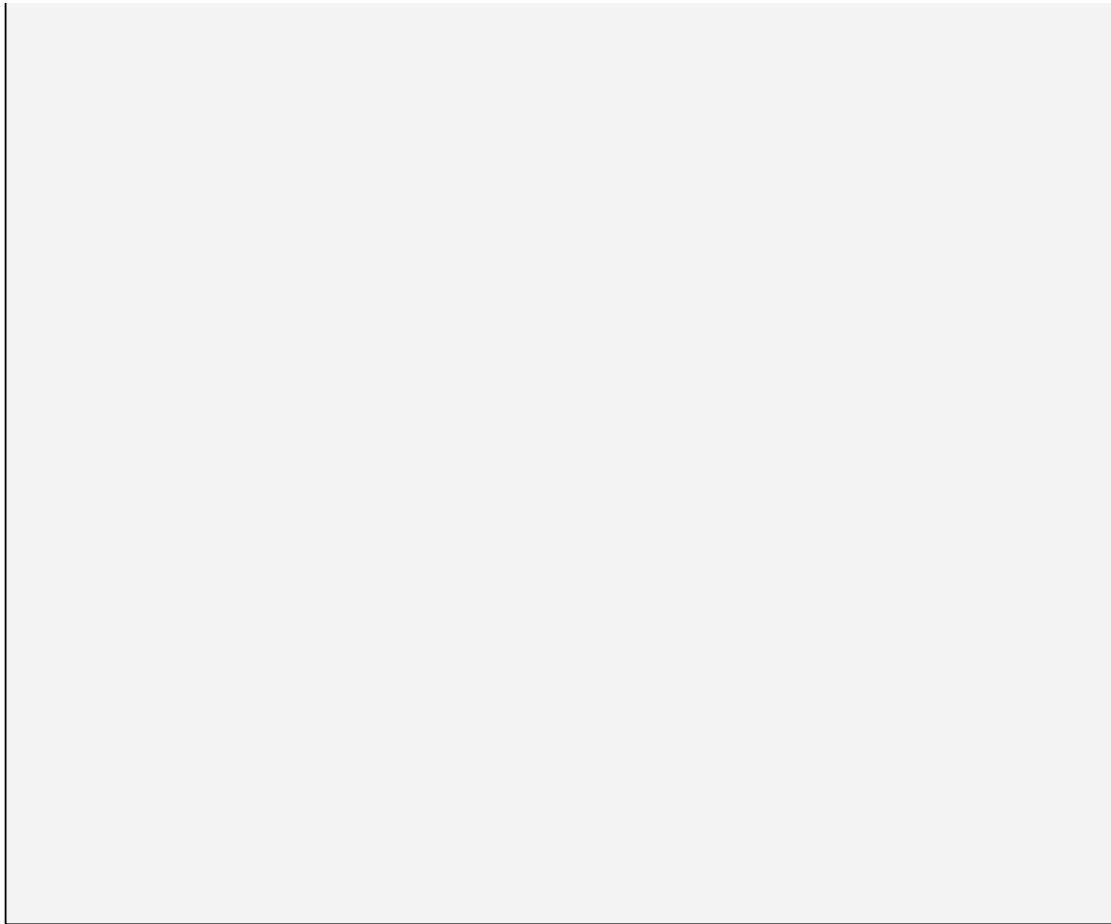
The area under the curve between aa and bb is calculated by integrating[2] the probability density function between the values aa and b.$b.$ Integration is the continuous analog of summation, <u>so integrating the probability density function from aa to bb is analogous to adding together the areas of very many narrow rectangles under the curve between aa and bb</u> (see the right panel in [Figure 5.4-4](#)).

For any continuous probability distribution, the area under the entire curve of its probability density function is always equal to one. Finally, because the probability of any individual Y-value$Y\text{-value}$ is infinitesimally small under a continuous probability density distribution, Pr[a≤Y≤b] $\Pr[a \leq Y \leq b]$ is the same as Pr[a<Y<b].$\Pr[a < Y < b].$

5.5 Either this or that: adding probabilities

Very often, we want to know the probability that we get *either* one event *or* another. For example, the probability that a randomly chosen North American has a particular ABO blood type and Rh factor ($+$ or $-$) is shown in Table 5.5-1 (Stanford Blood Center 2012). What is the probability that an American has blood type O? A person is blood type O if she is either $O+$ or $O-$. We can use the addition rule to calculate this probability.

TABLE 5.5-1 Probability that a randomly chosen American will have a given blood type. A, B, and O refer to ABO blood type, and "+" and "−" refer to Rh factor.

Blood type	Probability
$O+$	0.374
$O-$	0.066
$A+$	0.357
$A-$	0.063
$B+$	0.085
$B-$	0.015
$AB+$	0.034
$AB-$	0.006

The addition rule

If two events are mutually exclusive, then calculating the probability of one or the other event occurring is both intuitive and easy. <u>The probability of getting either of two mutually exclusive events is simply the sum of the probabilities of each of those events separately.</u> Having blood type $O-$ and having blood type $O+$ are mutually exclusive events. Therefore, the chance of a person being either $O-$ or $O+$ is the chance of being $O-$ plus the chance of being $O+$:

$$\Pr[O- \text{ or } O+] = \Pr[O-] + \Pr[O+] = 0.374 + 0.066 = 0.440.$$

Figure 5.5-1 illustrates the probability of $O-$ or $O+$ with a Venn diagram.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.5-1

The probability of O–$^{O-}$ or O+$^{O+}$ is equal to the probability of O–$^{O-}$ plus the probability of O+$^{O+}$, because the two events are mutually exclusive.

Description

First row reads: O positive, A positive, B positive, and A B positive. Second row reads: O negative, A negative, B negative, and AB negative. The cells O positive and O negative are highlighted in orange.
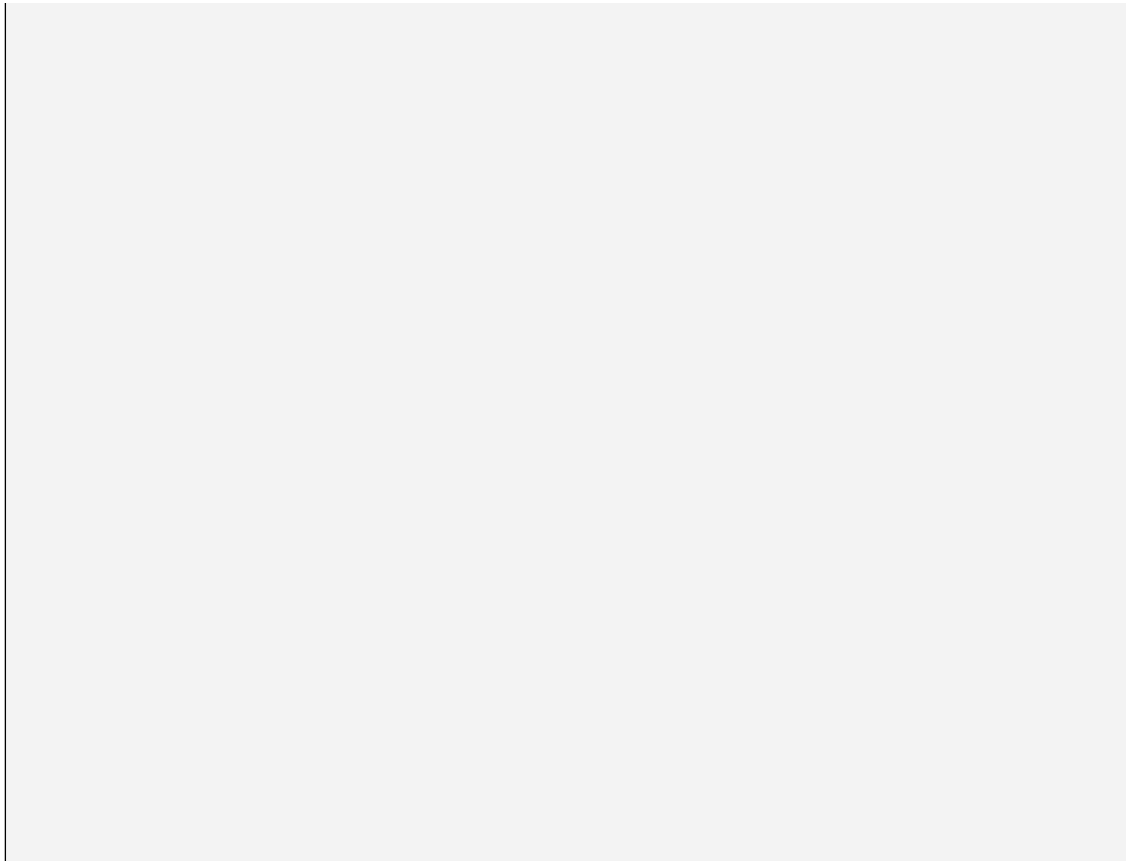
This additive property of the probabilities of mutually exclusive events is called the **addition rule**.

> The *addition rule*: If two events AA and BB are mutually exclusive, then
> Pr[A or B]=Pr[A]+Pr[B]. $\Pr[A \text{ or } B] = \Pr[A] + \Pr[B].$

The addition rule extends to more than two events as long as they are all mutually exclusive. Let's say that your blood type is B−$B-$, and we want to know the probability that you could safely donate blood to a randomly sampled American in the event of emergency. B−$B-$ blood can be safely donated to anyone who is B+$B+$, B−$B-$, AB+$AB+$, or AB−.$AB -.$ These four possibilities are mutually exclusive, because a randomly sampled American cannot have more than one of these blood types at the same time. Thus, the probability that an American is able to receive your B−$B-$ blood safely can be calculated as follows using the addition rule:

$\Pr[$

$\Pr[\text{B+or B−or AB+or AB−}]=\Pr[\text{B+}]+\Pr[\text{B−}]+\Pr[\text{AB+}]+\Pr[\text{AB−}]=0.085+0.015+0.034+0.006= 0.140.$

The addition rule is about "or" statements. If two events are mutually exclusive and we want to know the probability of being *either* one *or* the other, we can use the addition rule. This property is vital to analyzing data because it allows us to calculate the probabilities of different outcomes of random sampling when they are mutually exclusive.

The probabilities of all possible mutually exclusive outcomes add to one

The probabilities of all possible mutually exclusive outcomes of a random trial must add to one. With blood type, for example, there are eight possible outcomes. Therefore, the sum of the probabilities of all outcomes is

$$\text{Pr[O+ or O− or A+ or A− or B+ or B− or AB+ or AB−]} = \text{Pr[O+]} + \text{Pr[O−]} + \text{Pr[A+]} + \text{Pr[A−]} + \text{Pr[B+]} +$$

This means that the probability of an outcome or event *not* occurring is simply one minus the probability that it occurs. For example, the probability that you do *not* get $O+$ when you type the blood of a randomly sampled American is

$$\text{Pr[not O+]} = 1 - \text{Pr[O+]} = 1 - 0.374 = 0.626.$$

This calculation is much easier than summing the probabilities of all outcomes other than $O+$.

> The probability of an event not occurring is one minus the probability that it occurs.
>
> $$\text{Pr}[not\ A] = 1 - \text{Pr}[A].$$

The general addition rule

Not all events, though, are mutually exclusive. It is possible, for example, for the ABO blood type of a randomly sampled American to be O and his or her Rh factor to be positive $(+)$. If the two events are not mutually exclusive, how do we calculate the probability of either one or the other event occurring?

In mathematical notation, a **general addition rule** can be written as

$$\text{Pr}[A\ or\ B] = \text{Pr}[A] + \text{Pr}[B] - \text{Pr}[A\ and\ B].$$

This calculates the probability that either A or B (or both) occur.

When events A and B are mutually exclusive, $\text{Pr}[A\ and\ B] = 0$, so the generalized addition rule reduces to the addition rule for mutually exclusive events introduced previously. The reason we have to subtract the probability of both A and B occurring is illustrated in Figure 5.5-2. If we do *not* subtract the probability of both A and B occurring, then we will double-count those outcomes where both A and B occur.

$$\text{Pr}[A \text{ or } B] \quad = \quad \text{Pr}[A] \quad + \quad \text{Pr}[B] \quad - \quad \text{Pr}[A \text{ and } B]$$

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.5-2

The general addition rule. Pr[A and B] $\text{Pr}[A \text{ and } B]$ is subtracted from Pr[A]+Pr[B] $\text{Pr}[A] + \text{Pr}[B]$ so that the outcomes where both A A and B B occur (the tan shaded areas) are not counted twice.
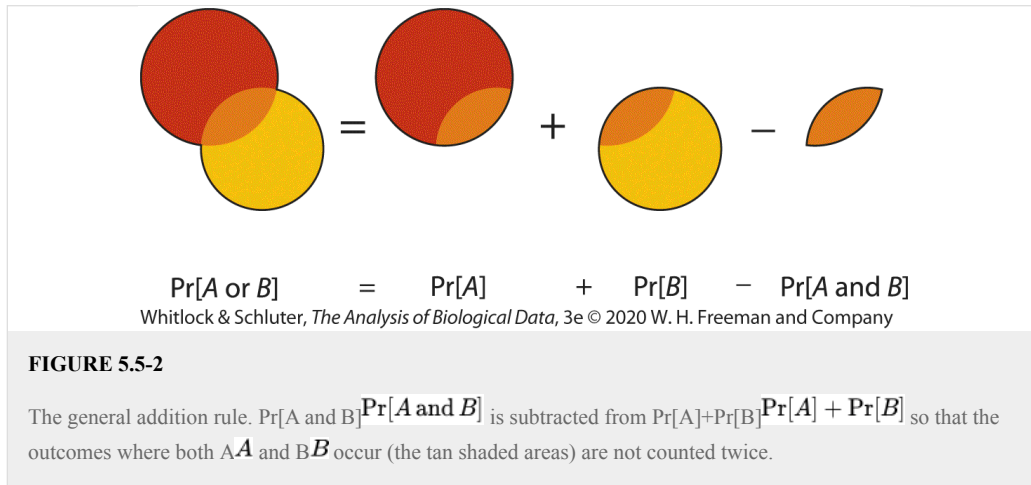
Description

A figure shows a couple of circles overlapping each other equals a circle plus another circle minus overlapping region between both these circles. The text below the figure reads, Probability of A or B equals Probability of A plus probability of B minus probability of A and B.
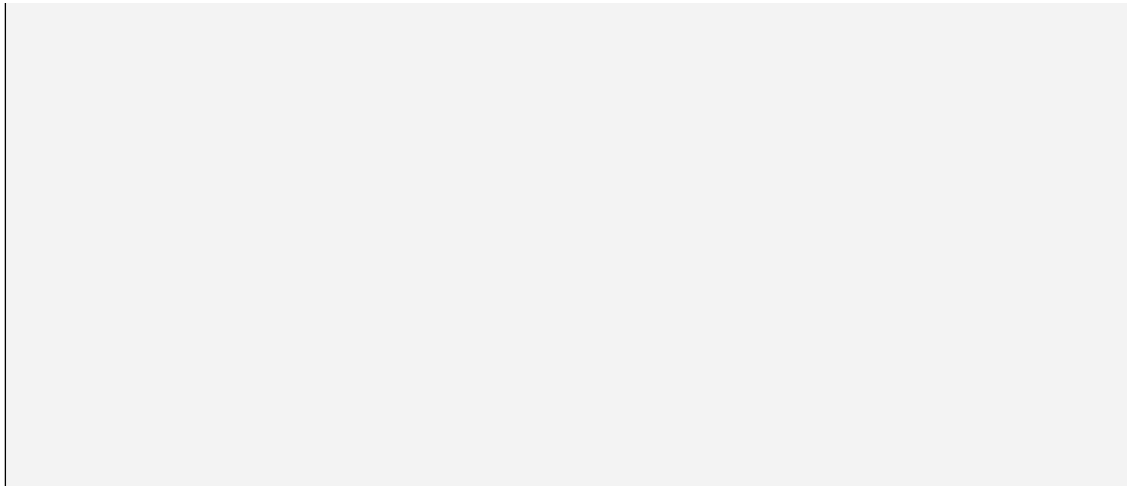
So, for example, the probability that a randomly chosen American has either the most common ABO type (O) or the most common Rh factor (+)$^{(+)}$ is

Pr[O]+Pr[+]−Pr[O and +]=0.440+0.850−0.374=0.916.$\text{Pr}[O] + \text{Pr}[+] - \text{Pr}[O \text{ and } +] = 0.440 + 0.850 - 0.374$

5.6 Independence and the multiplication rule

Science is the study of patterns, and patterns are generated by relationships between events. Men are more likely to be tall, more likely to have a beard, more likely to die young, and more likely to go to prison than women. In other words, height, beardedness, age at death, and criminal conduct are not independent of sex in the human population.

Sometimes, though, the chance of one event occurring does *not* depend on another event. If we roll two dice, for example, the number on one die does not affect the number on the other die. If knowing one event gives us no information about another event, then these two events are called independent.

Two events are **independent** if the occurrence of one does not in any way inform us about the probability that the other will also occur. When rolling the same fair die twice in a row, for example, the probability that the first roll gives a three is $1/6$, as we saw previously:

$$\Pr[\text{first roll is three}] = 1/6.$$

What is the probability that the next roll will also be a three? The probability of rolling a three on the second roll is still , regardless of whether the first roll was a three or not. Because the outcome of the first roll does not give any information about the probability of rolling a three on the second roll, we can say that the two events are independent (Figure 5.6-1).

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Probability of rolling a 3 on the first roll is 1/6.

Probability of rolling a 3 on the second roll is 1/6.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
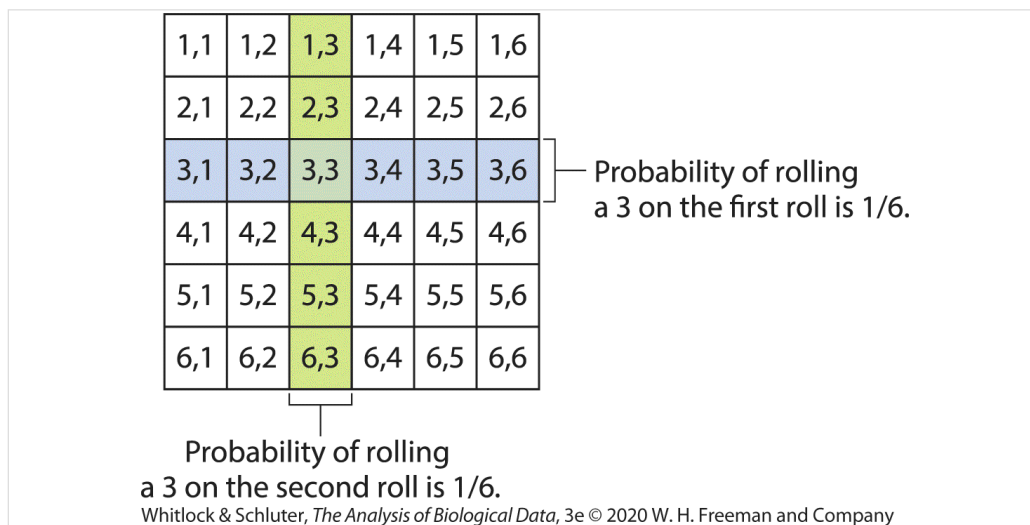
FIGURE 5.6-1

A Venn diagram for all the possible outcomes of rolling two 6-sided dice. The first digit of each pair shows the result of the roll of the first die, and the second number shows the result of the roll of the second die. Rolling a three on the first roll is shown in the blue row. The probability of rolling a three on the second roll is shown in the green column and is the same (1/6) regardless of the result of the first roll.

Description

A table of 6 rows and 6 columns shows all the possible outcomes of rolling two six-sided dice. First row reads: 1, 1; 1, 2; 1, 3; 1, 4; 1, 5; 1, 6; 2, 1; 2, 2; 2, 3; 2, 4; 2, 5; 2, 6; 3, 1; 3, 2; 3, 3; 3, 4; 3, 5; 3, 6; 4, 1; 4, 2; 4,

3; 4, 4; 4, 5; 4, 6; 5, 1; 5, 2; 5, 3; 5, 4; 5, 5; 5, 6; 6, 1; 6, 2; 6, 3; 6, 4; 6, 5; 6, 6. Third row is highlighted with the text reading, probability of rolling a 3 on the first roll is 1 over 6. The third column is highlighted with the text reading, probability of rolling a 3 on the second roll is 1 over 6.

Two events are *independent* if the occurrence of one does not inform us about the probability that the second will occur.

When the occurrence of one event provides at least some information about the results of another event, then the two events are **dependent**.

Multiplication rule

When two events are independent, then the probability that they both occur is the probability of the first event multiplied by the probability of the second event. This is called the **multiplication rule**. When we analyze data, we use this multiplication rule to determine what to expect when two variables are independent.

> The *multiplication rule*: If two events A A and B B are independent, then
> Pr[A and B]=Pr[A]× Pr[B]. $\Pr[A \text{ and } B] = \Pr[A] \times \Pr[B].$

We can see the basis of the multiplication rule in Figure 5.6-1. The area of the Venn diagram that corresponds to "rolling a three on the first die" and "rolling a three on the second die" is the region of overlap between the blue and green areas. Because the two events are independent, the area of this overlap zone is just the probability of being in the blue times the probability of being in the green:

Pr[(first roll is a three) and (second roll is a three)]=Pr[first roll is a three] ×Pr[second roll is a three]=1/6

The multiplication rule pertains to combinations with "and"—that is, that *both* events occur. If we want to know the probability of *this* and *that* occurring, and if the two events are independent, we can multiply the probabilities of each to get the probability of both occurring. Example 5.6A applies the multiplication rule to a study about smoking and high blood pressure.

EXAMPLE 5.6A: Smoking and high blood pressure

Both smoking and high blood pressure are risk factors for strokes and other vascular diseases. In the United States, approximately 17% of adults smoke and about 22% have high blood pressure. Research has shown that high blood pressure is not associated with smoking; that is, they seem to be independent of each other (Liang et al. 2001). What is the probability that a randomly chosen American adult has both of these risk factors?

Because these two events are independent, the probability of a randomly sampled individual both "smoking" and "having high blood pressure" is the probability of smoking times the probability of high blood pressure:

$\Pr[smokin$

Pr[smoking and high blood pressure]=Pr[smoking] ×Pr[high blood pressure]=0.17×0.22=0.037.

Therefore, 3.7% of adult Americans will have both of these risk factors for strokes. This calculation is shown geometrically in the Venn diagram in Figure 5.6-2.
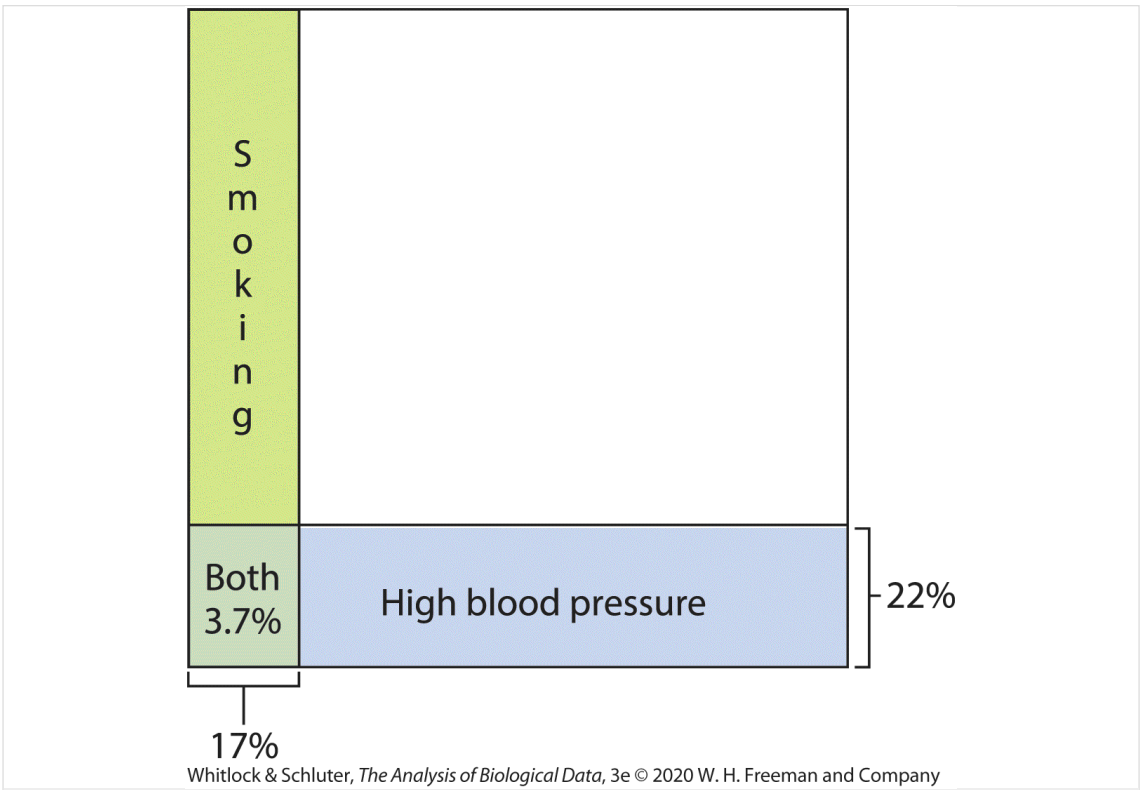
FIGURE 5.6-2

Venn diagram for the two independent factors smoking and high blood pressure. The probability of having both risk factors is proportional to the area of the rectangle in the bottom left corner.

Description

The first row reads: smoking and blank. The second row reads: Both 3 point 7 percent, High blood pressure. The first column is highlighted with the text 17 percent. The second row is highlighted with the text 22 percentage.

"And" versus "or"

Probability statements involving "and" or "or" statements are common enough, and confusing enough, that it is worth summarizing them together:

- The probability of A A or B B involves addition. That is, Pr[A or B] =Pr[A] +Pr[B]

  $\Pr[A \text{ or } B] = \Pr[A] + \Pr[B]$ if the two events A A and B B are mutually exclusive.

- The probability of A A and B B involves multiplication. That is, Pr[A and B] = Pr[A] ×Pr[B]

  $\Pr[A \text{ and } B] = \Pr[A] \times \Pr[B]$ if A A and B B are independent.

What may be confusing is that the statement involving "and" requires multiplication, not addition.

Independence of more than two events

The multiplication rule also applies to more than two events, as Example 5.6B demonstrates. If several events are all independent, then the probability of all events occurring is the product of the probabilities that each one occurs.

EXAMPLE 5.6B: Mendel's peas

Like blue eyes in humans, yellow pods in peas is a recessive trait. That is, pea pods are yellow only if both copies of the gene code for yellow. A plant having only one yellow copy and one green copy (a "heterozygote") has green pods just like the pods of plants having two green copies of the gene (a green "homozygote"). Gregor Mendel devised a method to determine whether a green plant was a heterozygote or a homozygote. He crossed the test plant to itself and assessed the pod color of 10 randomly chosen offspring. If all 10 were green, he inferred the plant was a homozygote, but if even one offspring was yellow, the test plant was classified as a heterozygote. However, he might have missed some heterozygotes, if by chance not a single yellow offspring was chosen. What is the chance of missing a heterozygote by Mendel's method? If the test plant is a homozygote, every offspring is green. If the test plant is a heterozygote, on the other hand, the chance of an offspring being green is $3/4$ and the chance of it being yellow is only $1/4$. What is the chance that all 10 offspring from a heterozygote test plant are green?

Mendel didn't carry out these calculations, but we can use our rules of probability to figure out the reliability of his approach. The chance that any one of a heterozygote's offspring is green is $3/4$. Because the genotype of each offspring is independent of the genotypes of other offspring, the probability that all 10 are green can be calculated using the multiplication rule.

$$\Pr[\text{all 10 green}] = \Pr[\text{first is green}] \times \Pr[\text{second is green}] \times \Pr[\text{third is green}] \times \cdots = 3/4 \times 3/4 \times 3/4 \times \cdots = (3/4)^{10} = 0.$$

Thus, Mendel likely misidentified about 5.6% of heterozygous individuals. On the other hand, his method correctly identified heterozygotes with probability $(1 - 0.056) = 0.944$.


5.7 Probability trees

A **probability tree** is a diagram that can be used to calculate the probabilities of combinations of events resulting from multiple random trials. We show how to use probability trees with Example 5.7.

EXAMPLE 5.7: Sex and birth order

Some couples planning a new family would prefer to have at least one child of each sex. The probability that a couple's first child is a boy[3] is 0.512. In the absence of technological intervention, the probability that their second child is a boy is independent of the sex of their first child, and so remains 0.512. Imagine that you are helping a new couple with their planning. If the couple plans to have only two children, what is the probability of getting one child of each sex?

This question requires that we know the probabilities of all mutually exclusive values of two separate variables. The first variable is "the sex of the first child." The second variable is "the sex of the second child." We can start building a probability tree by considering the two variables in sequence. Let's start with the sex of the first child. Two mutually exclusive outcomes are possible—namely, "boy" and "girl"—which we list vertically, one below the other (see figure at right). We then draw arrows from a single point on the left to both possible outcomes. Along each arrow we write the probability of occurrence of each outcome (0.512 for "boy" and 0.488 for "girl").
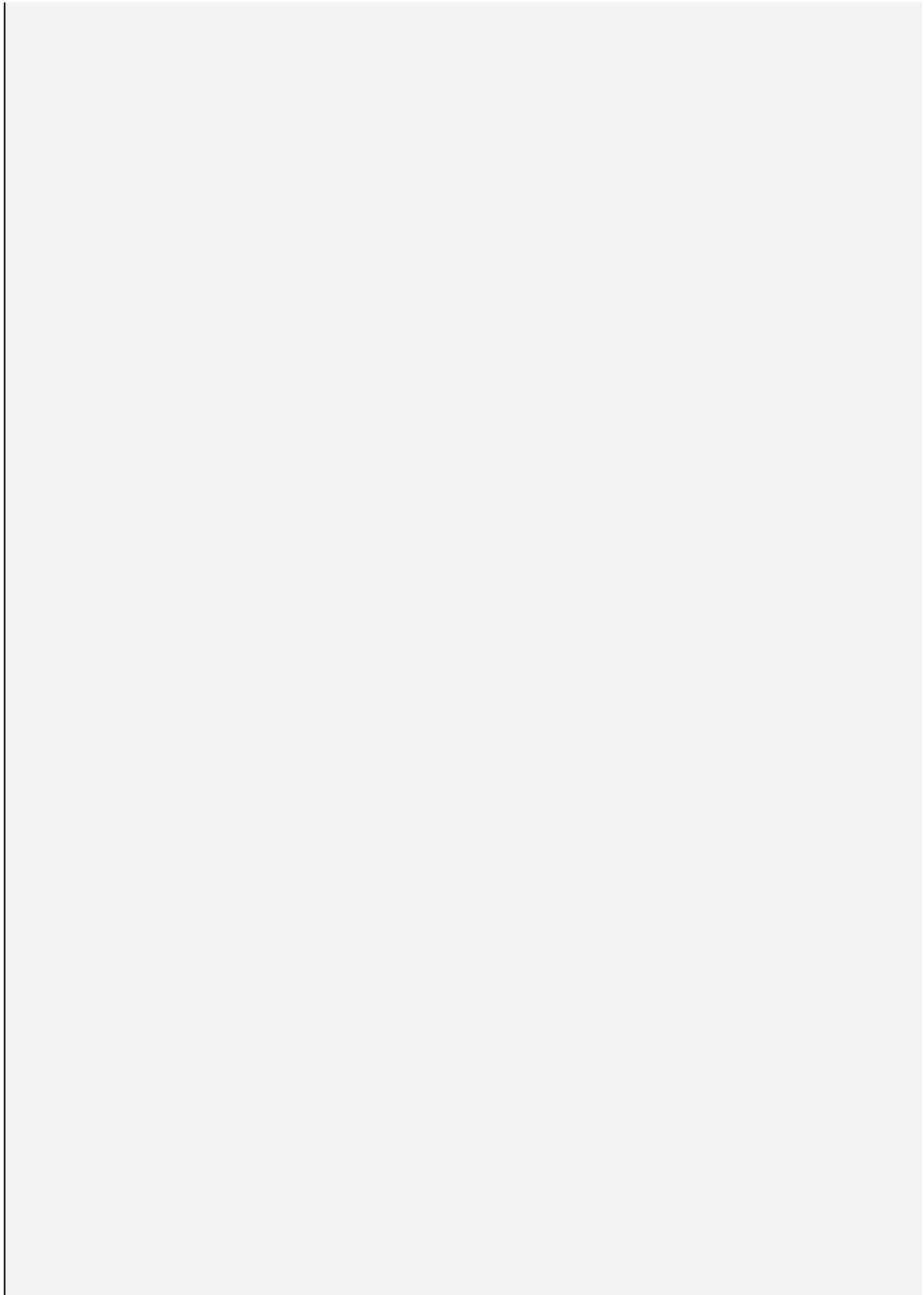
Sex of first child
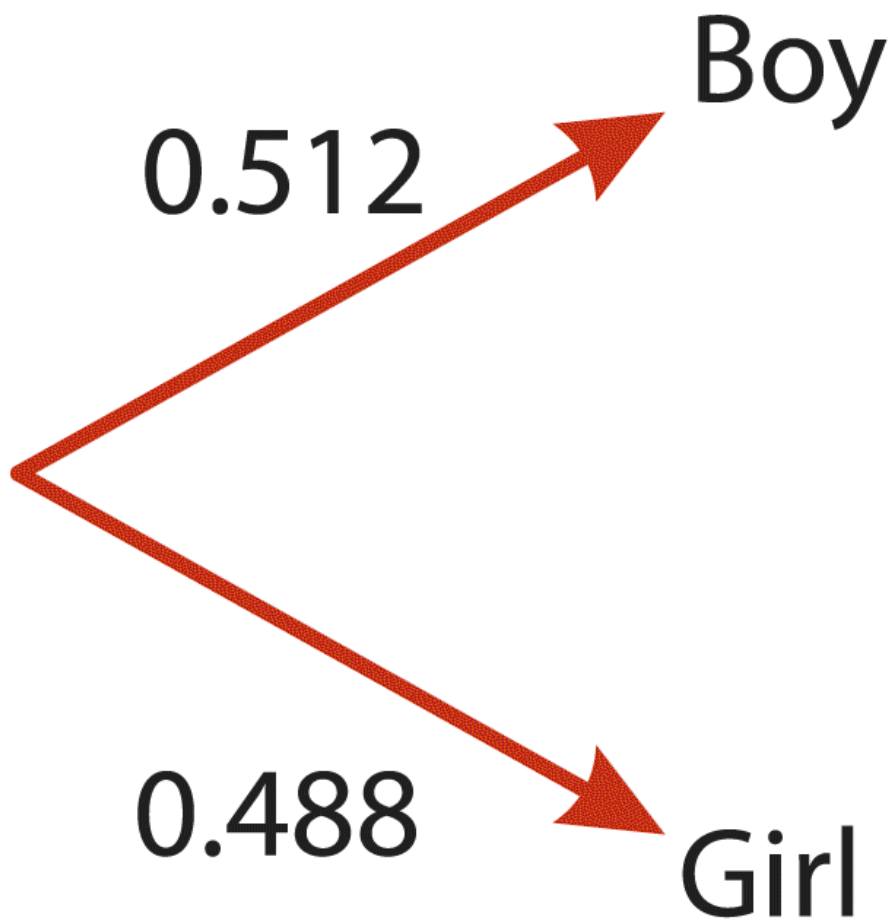
0.512 → Boy

0.488 → Girl

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Description

-

Now, we list all possible values for the second variable, but we do so separately for each possible value of the first variable. For example, for

the value "boy" for the first child, we list both possible values (i.e., "boy" and "girl") for the sex of the second child. Next, we draw arrows originating from the value "boy" for the first variable to both possible values for the second variable. Then we write the probability of each value for the second variable along each arrow. We repeat this process for the case when "girl" is the value of the first variable. The resulting probability tree is shown in Figure 5.7-1.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
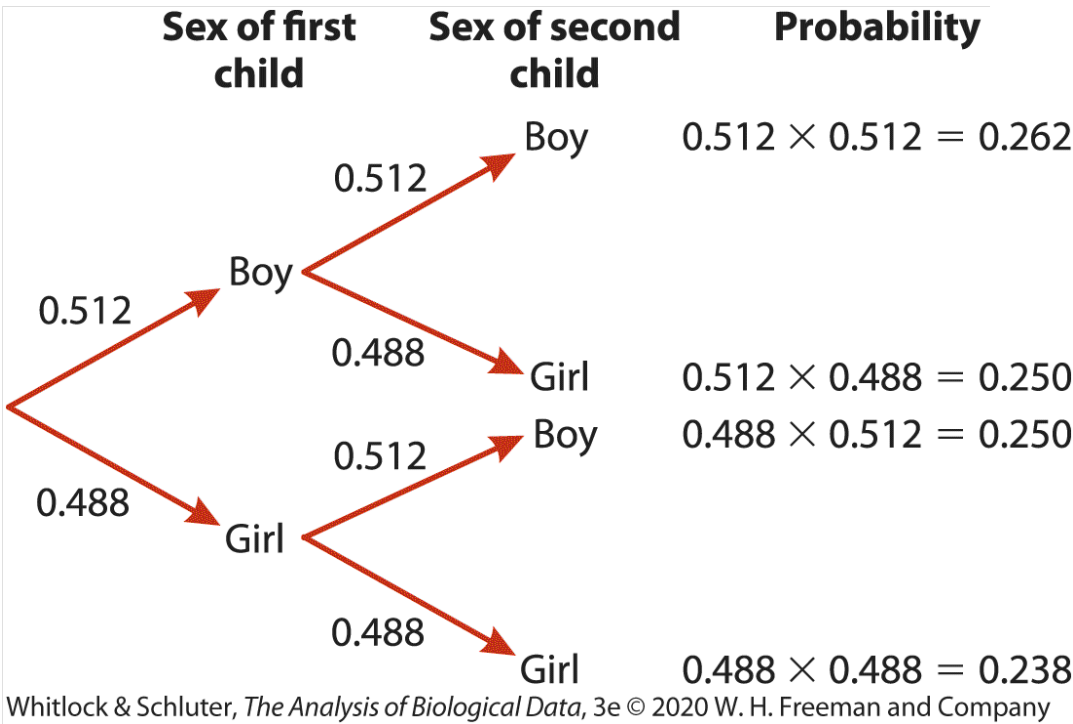
FIGURE 5.7-1

A probability tree for all possible values of a two-child family.

Description

In the first set, one ray, labeled zero point five one two, points at Sex of first child – Boy and second ray, labeled zero point four eight eight, points at Sex of first child – Girl. In the second set, one ray, labeled zero point five one two, points at Sex of second child – Boy, and second ray, labeled zero point four eight eight, points at Sex of second child – Girl. In the third set, one ray,

labeled zero point five one two, points out Sex of second child – Boy and second ray, labeled zero point four eight eight, points out Sex of second child – Girl. The text on the right side reads: Probability, zero point five one two times zero point five one two equals zero point two six two, zero point five one two times zero point four eight eight equals zero point two five zero, zero point four eight eight times zero point five one two equals zero point two five zero, zero point four eight eight times zero point four eight eight equals zero point two three eight.

At this point, we should check that our probabilities are written down correctly. For instance, the probabilities along all arrows originating from a single point must sum to one (within rounding error) because they represent all the mutually exclusive possibilities. If they don't sum to one, we've forgotten to include some possibilities or we've written down the probabilities incorrectly.

With a probability tree, we can calculate the probability of every possible sequence of values of the two variables. A sequence of values is represented by a path along the arrows of the tree that begins at the root at the far left and ends at one of the branch tips on the right. The probability of a given sequence is calculated by multiplying all of the probabilities along the path taken from the root to the tip. For example, the sequence "boy then girl" in Figure 5.7-1 has a probability of 0.512×0.488=0.250. $0.512 \times 0.488 = 0.250.$ On our probability tree, we usually list the probabilities of each sequence of values in a column to the right of the tree tips, as shown in Figure 5.7-1.

Each tree tip defines a unique and mutually exclusive sequence of events. Check Figure 5.7-1 (or any probability tree) to make sure that the probabilities of all possible sequences add to one. If they don't add to one (within rounding error), then something has gone wrong in the construction of the tree.

What is the probability of having one child of each sex in a family of two children? According to the probability tree, two of the four possible sequences result in the birth of one boy and one girl. In the first sequence, the boy is born first, followed by the girl, whereas in the second sequence, the girl is born first and the boy is born second. These two different sequences are mutually exclusive, and we are looking for the probability of either the first *or* the second sequence. By the addition rule, therefore, the probability of getting exactly one boy and one girl when having two children is the sum of the probabilities of the two alternative sequences leading to this event: 0.250+0.250=0.500. $0.250 + 0.250 = 0.500.$

We could also use the probability tree in Figure 5.7-1 to calculate probabilities of the following events:

- The probability that at least one girl is born,
- The probability that at least one boy is born, and
- The probability that both children are the same sex.

Calculate these probabilities yourself to test your understanding.[4]

It is not essential to use probability trees when calculating the probabilities of sequences of events, but they are a helpful tool for making sure that you have accounted for all of the possibilities.

5.8 Dependent events

Independent events are mathematically convenient, but when the probability of one event depends on another, things get interesting. Much of science involves identifying variables that are associated.

Sex determination is more exotic in many insects than in humans. In many species, the mother can alter the relative numbers of male and female offspring depending on the local environment. In this case, "sex of offspring" and "environment" are dependent events, as Example 5.8 demonstrates.
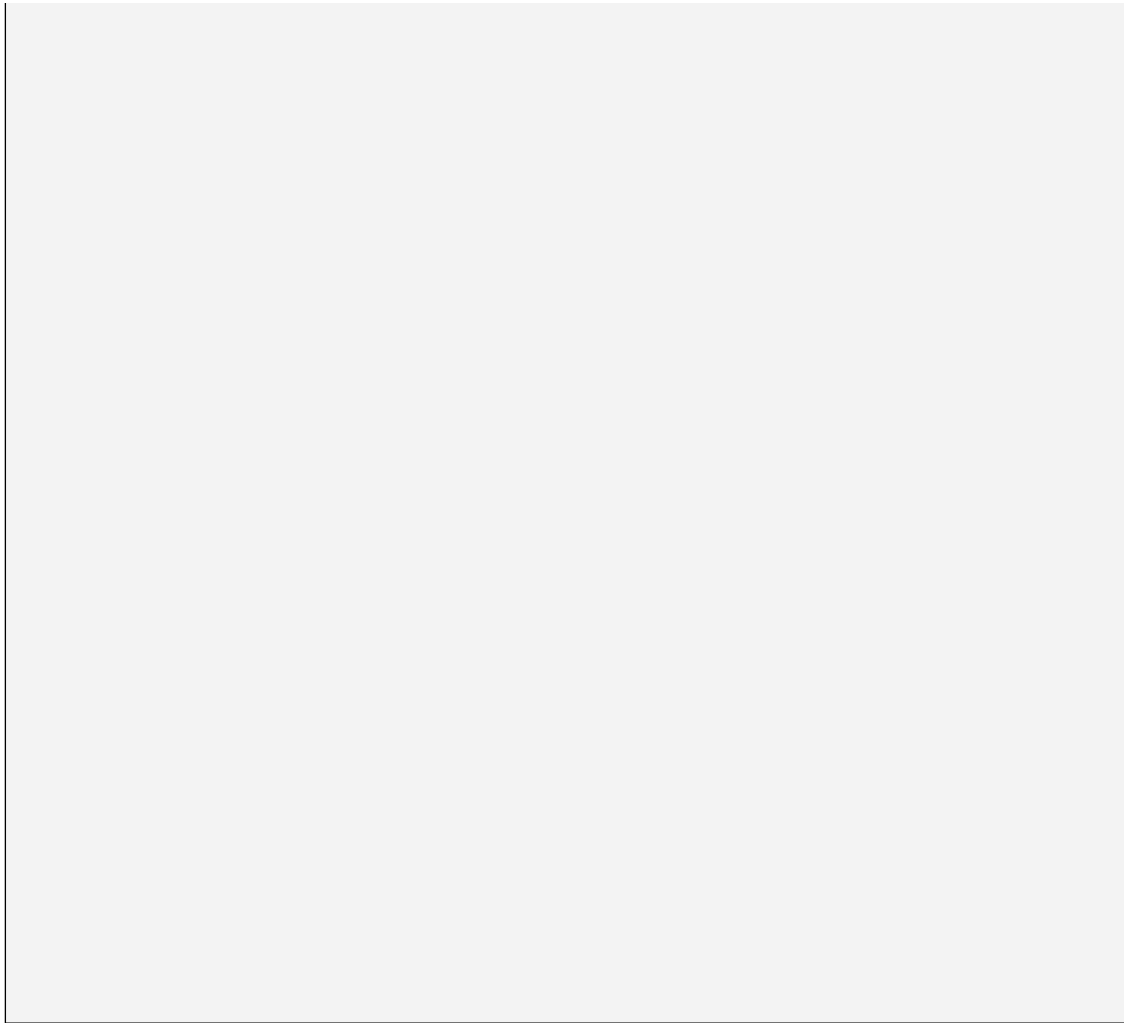

EXAMPLE 5.8: Is this meat taken?

M.E.Clark

Description
-

419

The jewel wasp, *Nasonia vitripennis*, is a parasite which lays its eggs on the pupae of flies. The larval *Nasonia* hatch inside the pupal case, feed on the live host, and grow until they emerge as adults from the now dead, emaciated host. Emerging males and females, possibly brother and sister, mate on the spot. *Nasonia* females have a remarkable ability to manipulate the sex of the eggs that they lay.[5] When a female finds a fresh host that has not been previously parasitized, she lays mainly female eggs, producing only the few sons needed to fertilize all her daughters. But if the host has already been parasitized by a previous female, the next female responds by producing a higher proportion of sons.[6] Thus, the state of the host encountered by a female and the sex of an egg laid are dependent variables (Werren 1980).

Suppose that, when a given *Nasonia* female finds a host, there is a probability of 0.20 that the host already has eggs, laid by a previous female wasp. Presume that the female can detect previous infections without error. If the host is unparasitized, the female lays a male egg with probability 0.05 and a female egg with probability 0.95. If the host already has eggs, then the female lays a male egg with probability 0.90 and a female egg with probability 0.10. Figure 5.8-1 shows a Venn diagram of these probabilities (in the form of a mosaic plot).

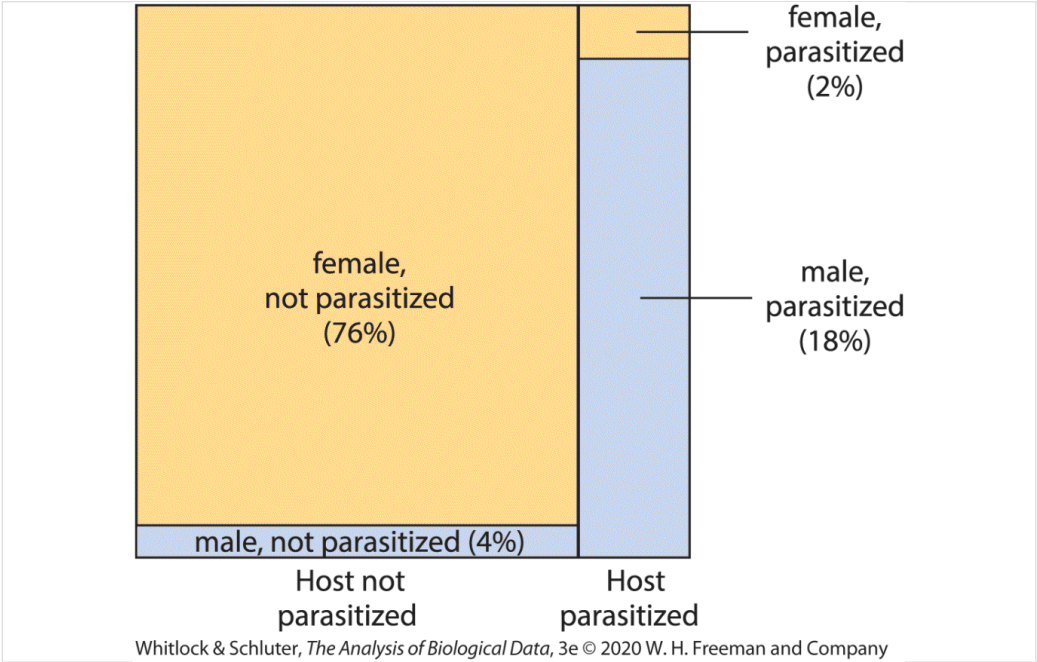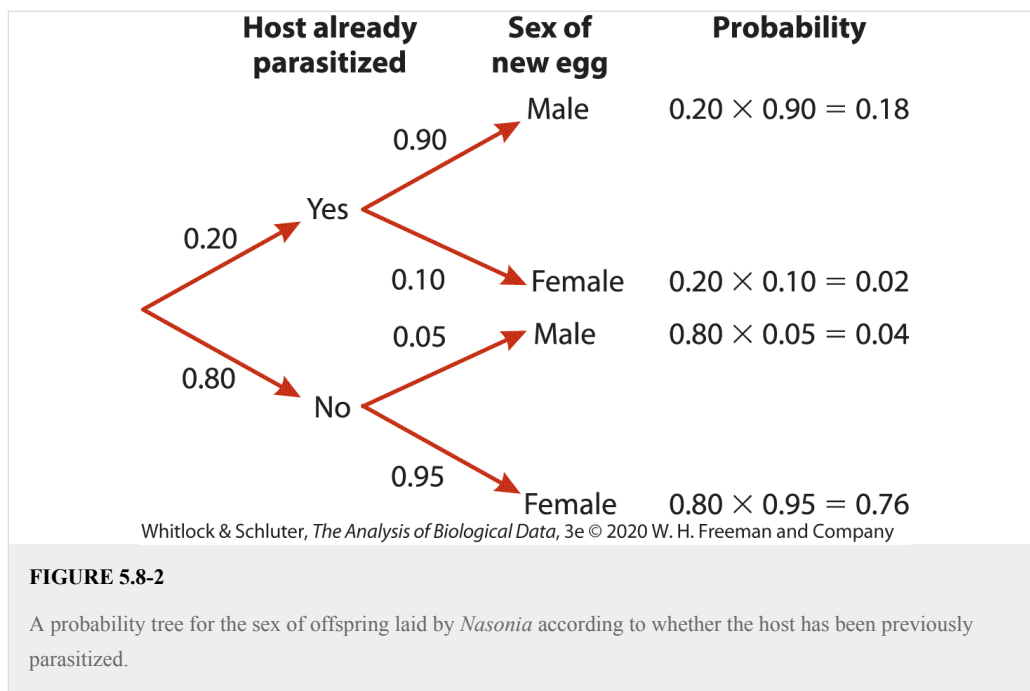Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.8-1

A Venn diagram showing that the sex of eggs laid by *Nasonia* females depends on the state of the host.

Description

-

Based on Figure 5.8-1, the events "host is previously parasitized" and "producing a male egg" are dependent. The probability of laying a male egg changes depending on whether the host has been previously parasitized. Suppose we want to know the probability that a new, randomly chosen egg is male. We can approach this question using a probability tree like the one shown in Figure 5.8-2.

Host already parasitized	Sex of new egg	Probability
	Male	$0.20 \times 0.90 = 0.18$
Yes	Female	$0.20 \times 0.10 = 0.02$
No	Male	$0.80 \times 0.05 = 0.04$
	Female	$0.80 \times 0.95 = 0.76$

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.8-2

A probability tree for the sex of offspring laid by *Nasonia* according to whether the host has been previously parasitized.

Description

In the first set, one ray, labeled zero point two, points out host already parasitized-Yes and second ray, labeled zero point eight, points out host already parasitized-No. In the second set, one ray, labeled zero point nine, points out Sex of new egg-Male and second ray, labeled zero point one, points out Sex of new egg -Female. In the third set, one ray, labeled zero point zero five, points out Sex of new egg –Male and second ray, labeled zero point nine five, points out Sex of new egg –Female. The text on the right side reads: Probability, zero point two times zero point nine equals zero point one eight, zero point two times zero point one equals zero point zero two, zero point eight times zero point zero five equals zero point zero four, zero point eight times zero point nine five equals zero point seven six.

According to the probability tree, there are exactly two paths that yield a male egg. In the first, the host is already parasitized and the mother lays a male egg. This path has probability

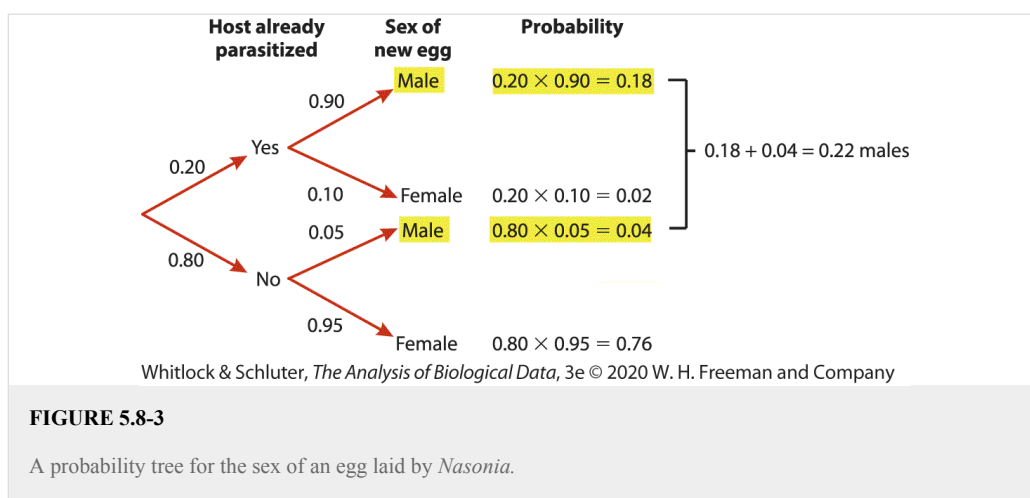Pr[host already parasitized and sex of new egg is male]=0.20×0.90=0.18. $\Pr[host\ already\ parasitized\ and\ sex$

In the second path, the host is not previously parasitized and the female lays a male egg. This second path has probability

Pr[host not already parasitized and sex of new egg is male] = 0.80×0.05 = 0.04. $\Pr[host\ not\ already\ parasitize$

The probability of a new egg being male is the sum of the probabilities of these two mutually exclusive paths:

Pr[male]=0.18+0.04=0.22. $\Pr[male] = 0.18 + 0.04 = 0.22.$

The probability of an egg being male in this population is 0.22. See Figure 5.8-3.

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

FIGURE 5.8-3

A probability tree for the sex of an egg laid by *Nasonia.*

Description

In the first set, one ray, labeled zero point two, points out host already parasitized-Yes and second ray, labeled zero point eight, points out host already parasitized-No. In the second set, one ray, labeled zero point nine, points out Sex of new egg-Male which is highlighted and second ray, labeled zero point one, points out Sex of new egg -Female. In the third set, one ray, labeled zero point zero five, points out Sex of new egg –Male which is highlighted and second ray, labeled zero point nine five, points out Sex of new egg –Female. The text on the right side reads: Probability, zero point two times zero point nine equals zero point one eight which is highlighted, zero point two times zero point one equals zero point zero two, zero point eight times zero point zero five equals zero point zero four which is highlighted, zero point eight times zero point nine five equals zero point seven six. "zero point one eight plus zero point zero four equals zero point two two males" is shown.
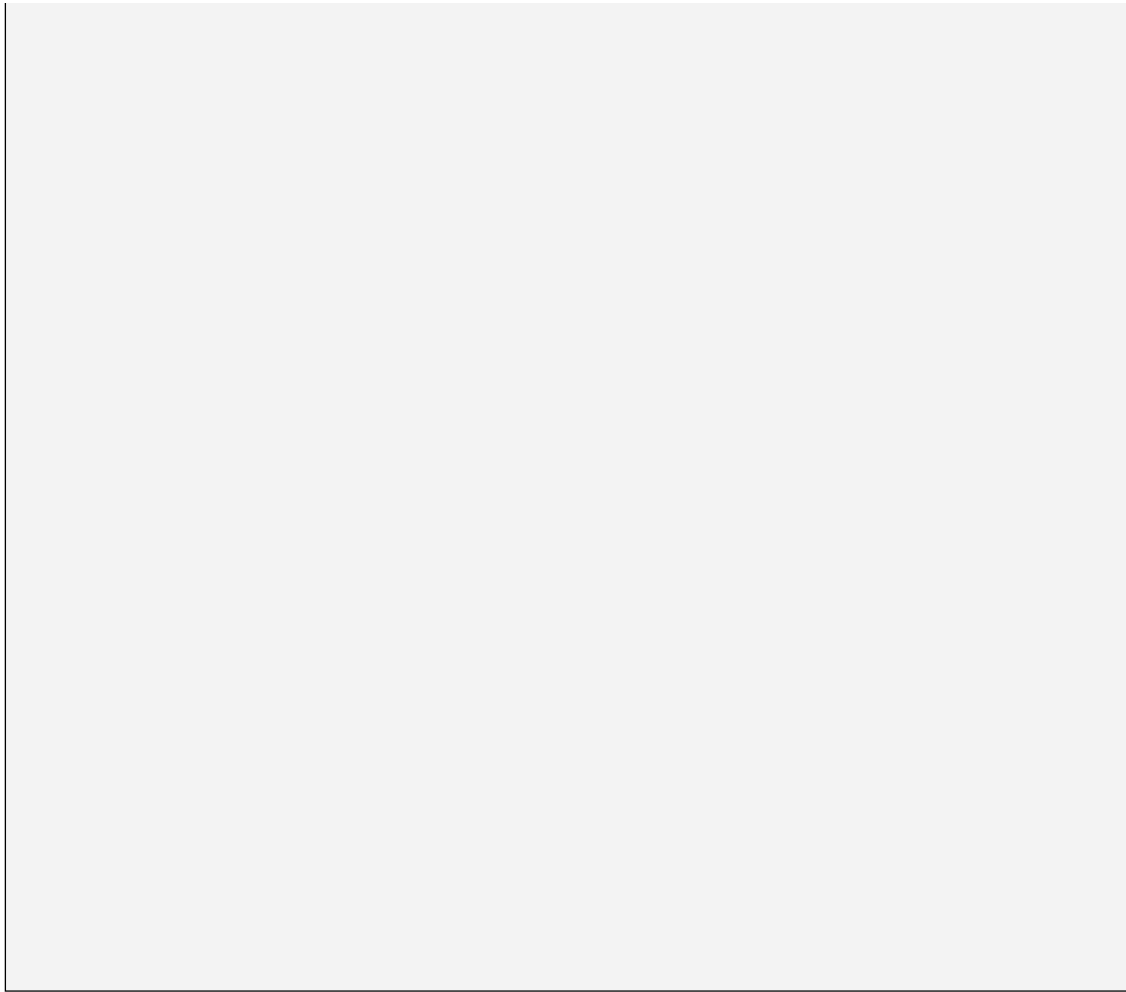
The probability tree shows that the event "sex of new egg is male" depends on whether the host encountered by a mother has been previously parasitized. Does this mean that the events "host already parasitized" and "sex of new egg is male" are not independent? One way to confirm this is via the multiplication rule, which applies only to independent events. The probability that "the host already had been parasitized and the sex of the new egg is male" is 0.18. This is *not* what we would have expected assuming independence, though. If we multiply the probability that the new egg is a male (0.22, as we just calculated) and the probability that a host is already parasitized (0.20), we get 0.22×0.20=0.044 $0.22 \times 0.20 = 0.044$, which is different from the actual probability of these two events (0.18). Based on the definition of "independence," then, these two events are *not* independent.

5.9 Conditional probability and Bayes' theorem

If we want to know the chance of an event, we need to take account of all existing information that might affect its outcome. If we want to know the probability that we will see an elephant on our afternoon stroll, for example, we would get a different answer depending on whether our walk was in the Serengeti or downtown Manhattan. The algebra of conditional probability lets us hone our statements about the chances of random events in the context of extra information.

Conditional probability

Conditional probability is the probability of an event given that another event occurs.

> The conditional probability of an event is the probability of that event occurring *given that* a condition is met.

In Example 5.8, the conditional probability that a jewel wasp will lay a male egg is 0.90 *given that* the host that she is laying on already has wasp eggs (i.e., has already been parasitized). Confirm this for yourself by looking at Figure 5.8-2 again. We write conditional probability in the following way:
$$\Pr[\text{new egg is male}|\text{host is previously parasitized}] = 0.90.$$
More generally, $\Pr[\text{event}\,|\,\text{condition}]$ represents the probability that the event will happen given that the condition is met. The vertical bar in the middle of this expression is a symbol that means "given that" or "when the following condition is met." (Be careful not to confuse it with a division sign.)

The plot in Figure 5.8-1 illustrates the meaning of this conditional probability. Ninety percent of the area corresponding to "host parasitized" represents the cases when the offspring is male, with the remaining 10% being females. The probability of a male is different under the condition "host parasitized" than under the condition "host not parasitized."

Conditional probability has many important applications. If we want to know the overall probability of a particular event, we sum its probability across every possible condition, weighted by the probability of that condition. This is known as the **law of total probability**.

> According to the *law of total probability,* the probability of an event, A, is
> $$\Pr[A] = \sum_{\text{All values of } B} \Pr[B]\,\Pr[A|B],$$
> where B represents all possible mutually exclusive values of the conditions.

One way of thinking about this formula is that it gives the weighted average probability of A over all possible mutually exclusive conditions.

The Venn diagram in Figure 5.8-1 makes it possible to visualize this, too. The probability of being male is obtained by adding the two blue areas, one for the condition when the host is already parasitized and the other for when the host is not already parasitized. The width of these boxes is proportional to the probability of the condition; the height is proportional to Pr[male | host condition]. $\Pr[male \mid host\ condition].$ By multiplying the width by the height of each box we find its area (its probability), and by adding all such boxes together we find the total probability of males.

To calculate the probability that a new egg is a male, we must consider two possible conditions: (1) the host is already parasitized and (2) the host is not parasitized. Thus, we'll have two terms on the right side of our equation:

Pr[egg is male]=Pr[host already parasitized] Pr[egg is male | host already parasitized]   +Pr[host not This is the same answer that we got from the probability tree, but now we can see how it can be derived from statements of conditional probability.

The general multiplication rule

With conditional probability statements, we can find the probability of a combination of two events even if they are not independent. When two events are not independent, the probability that both occur can be found by multiplying the probability of one event by the conditional probability of the second event, given that the first has occurred. This is the **general multiplication rule.**

> The *general multiplication rule* finds the probability that both of two events occur, even if the two are dependent:
> Pr[A and B]=Pr[A] Pr[B|A]. $\Pr[A\ and\ B] = \Pr[A]\ \Pr[B|A].$

This rule makes sense, if we think it through. For two events (A and B)$(A\ and\ B)$ to occur, event AA must occur. By definition, this happens with probability Pr[A].$\Pr[A].$ Now that we know AA has occurred, the probability that BB also occurred is Pr[B | A].$\Pr[B \mid A].$ Multiplying these together gives us the probability of both AA and BB occurring.

It doesn't matter which event we label AA and which we label B.$B.$ The reverse is also true; that is, Pr[A and B]=Pr[B]Pr[A|B]. $\Pr[A\ and\ B] = \Pr[B]\ \Pr[A|B].$
With the jewel wasps, for example, if we wanted to know the probability that a host had already been parasitized and that the mother wasp laid a male egg, we would multiply the probability that it had been parasitized (0.2) times the probability of a male egg *given that* the egg was already parasitized (0.9), to get 0.18. We can see the same probability by following the appropriate path (the top one) through the probability tree in Figure 5.8-2.

If AA and BB are independent, then having information about AA gives no information about BB, and therefore Pr[B|A]=Pr[B].$\Pr[B|A] = \Pr[B].$ That is, the general multiplication rule reduces to the

multiplication rule when the events are independent.

Sampling without replacement

One common use of conditional probability is **sampling without replacement**. This process occurs when the specific outcome of one random trial eliminates or depletes that outcome from the possibilities and so changes the probability distribution of values for subsequent random trials.

As a simple example, consider drawing cards randomly from a fair card deck in which the 52 ordinary cards have been shuffled and so are in random order. What is the probability of drawing three cards in the precise sequence "ace-2-3," ignoring card suit, without returning the cards to the deck? The probability that the first card drawn is an ace is $4/52$, because there are four aces out of the 52 cards. The key novelty is that the outcome of the first draw changes the probabilities of outcomes for later draws if the card is not returned to the deck. For example, if the first card is an ace, then the probability of a 2 in the next draw is changed because there are now only 51 cards in the deck. The chance that the second card is a 2 is now $4/51$. And if we have already taken an ace and a 2 from the deck, the probability that the third card is a 3 is $4/51$, because there are 50 cards left and four of them are 3's. So the probability that the first three draws are in the sequence ace-2-3 is $(4/52) \times (4/51) \times (4/50)$.

In contrast, when sampling with replacement, the sampled individual is not removed from the population after sampling. In this case, the frequencies of possible outcomes in the population are not changed by successive samples.

When sampling populations for biological study, we usually choose populations that are large enough that the sampling of each individual doesn't change the probability distribution of possible values in the individuals that remain. We assume that the effects of depletion are so slight that they don't matter. This will not always be the case, however.

Bayes' theorem

One powerful mathematical relationship about conditional probability is **Bayes' theorem.**[7]

According to Bayes' theorem, for two events A and B,

$$\Pr[A|B] = \frac{\Pr[B|A]\ \Pr[A]}{\Pr[B]}.$$

Bayes' theorem may seem rather complicated, but it can be derived from the general multiplication rule. Because

$$\Pr[A \text{ and } B] = \Pr[B]\ \Pr[A|B]$$

and

$$\Pr[A \text{ and } B] = \Pr[A]\ \Pr[B|A],$$

it is also true that

$$\Pr[B]\ \Pr[A|B] = \Pr[A]\ \Pr[B|A].$$

Dividing both sides by $\Pr[B]$ gives Bayes' theorem. <u>Example 5.9</u> applies Bayes' theorem to the detection of Down syndrome.

EXAMPLE 5.9: Detection of Down syndrome

Down syndrome (DS) is a chromosomal condition that occurs in about one in 1000 pregnancies. The most accurate test for DS in wide use requires amniocentesis, which unfortunately carries a risk of miscarriage (about one in 200). It would be better to have an accurate test of DS without the risks. One such test in common use is called the triple test, which screens for levels of three hormones in maternal blood at around 16 weeks of pregnancy.

The triple test is not perfect, however. It does not always correctly identify a fetus with DS (an error called a false negative), and sometimes it incorrectly identifies a fetus with a normal set of chromosomes as DS (an error called a false positive). Under normal conditions, the sensitivity of the triple test (i.e., the probability that a fetus with DS will be correctly scored as having DS) is 0.60. The false-positive rate (i.e., the probability that a test would say incorrectly that a normal fetus had DS) is 0.05 (<u>Newberger 2000</u>).

Most people's intuition is that these numbers are acceptable. Based on the probabilities given, the triple test would seem to be right most of the time. But, if the test on a randomly chosen fetus gives a positive result (i.e., it indicates that the fetus has DS), what is the probability that this fetus actually has DS? Make a guess at the answer before we work it through.

To address this question, we need Bayes' theorem. We want to know a conditional probability—the probability that a fetus has DS given that its triple test showed a positive result. In other words, we want to know $\Pr[DS|positive\ result]$. Using Bayes' theorem,

$$\Pr[DS|positive\ result] = \frac{\Pr[positive\ re}{\Pr[posit}$$

We've been given $\Pr[Positive\ result\,|\,DS]$ and $\Pr[DS]$, the two factors in the numerator, but we haven't been given $\Pr[Positive\ result]$, the term in the denominator. We can figure out the probability of a positive result, though, by using the law of total probability introduced earlier in this section. That is, we can sum over all the possibilities to find the probability of a positive result.

$$\Pr[positive\ result] = (\Pr[positive\ result|DS]\ \Pr[DS]) \quad +(\Pr[positive\ result|no\ DS]\ \Pr[no\ DS]) = (0.60 \times 0.001$$

The probability of something *not* occurring is equal to one minus the probability of it occurring, so the probability that a randomly chosen fetus does *not* have DS is one minus the probability that it has DS.

According to [Example 5.9](#), Pr[DS]=0.001 $\Pr[DS] = 0.001$, so Pr[no DS]=1−0.001=0.999 $\Pr[no\,DS] = 1 - 0.001 = 0.999$ in the preceding equation.

Now, returning to Bayes' theorem, we can find the answer to our question.

Pr[DS|positive result]=0.60×0.0010.05055=0.012. $\Pr[DS|positive\ result] = \frac{0.60 \times 0.001}{0.05055} = 0.012.$

There is a very low probability (i.e., 1.2%) that a fetus with a positive score on the triple test actually has DS!

Many people find it more intuitive to think in terms of numbers rather than probabilities for these kinds of calculations. For every million fetuses tested, 1000 will have DS, and 999,000 will not. Of those 1000, 60% or 600 will test positive. Of the 999,000, 5% or 49,950 will test false-positive. Out of a million tests, therefore, there are 600+49,950=50,550 $600 + 49{,}950 = 50{,}550$ positive results, only 600 of which are true positives. The 600 true positives divided by the 50,550 total positives is 1.2%, the same answer as we got before. DS babies have a high probability of being detected, but they are a very small fraction of all babies. Thus, the true positive results get swamped by the false positives.

This high false-positive ratio is not unusual. Many diagnostic tools have high proportions of false positives among the positive cases. In this case, erring on the side of caution is appropriate because, when the triple test returns a positive result, it can be checked by amniocentesis.

Did you think that the probability of DS with a positive result would be higher? If so, you're not alone. A survey of practicing physicians found that their grasp of conditional probability with false positives was extremely poor ([Elstein 1988](#)). In a question about false-positive rates, where the correct answer was that 7.5% of patients with a positive test result had breast cancer, 95% of the doctors guessed that the answer was 75%! If these doctors had a better understanding of probability theory, they could avoid overstating the risks of serious disease to their patients, thus reducing unnecessary stress.

5.10 Summary

- Probability is an important concept in biology. One reason is that randomly sampling a population represents a random trial whose outcomes are governed by the rules of probability.
- A random trial is a process or experiment that has two or more possible outcomes whose occurrence cannot be predicted with certainty.
- The probability of an event is the proportion of times the event occurs if we repeat a random trial over and over again under the same conditions.
- A probability distribution describes the probabilities of all possible outcomes of a random trial.
- Two events (A and B) are mutually exclusive if they cannot both occur (i.e., $\Pr[A \text{ and } B] = 0$). If A and B are mutually exclusive, then the probability of A or B occurring is the sum of the probability of A occurring and the probability of B occurring (i.e., $\Pr[A \text{ or } B] = \Pr[A] + \Pr[B]$). This is the addition rule.
- The general addition rule gives the probability of either of two events occurring when the events are not mutually exclusive:
  $$\Pr[A \text{ or } B] = \Pr[A] + \Pr[B] - \Pr[A \text{ and } B].$$

  The general addition rule reduces to the addition rule when A and B are mutually exclusive, because then $\Pr[A \text{ and } B] = 0.$
- Two events are independent if knowing one outcome gives no information about the other outcome. More formally, A and B are independent if $\Pr[A \text{ and } B] = \Pr[A]\Pr[B].$ This is the multiplication rule.
- Probability trees are useful devices for calculating the probabilities of complicated series of events.
- If events are not independent, then they are said to be dependent. The probability of two dependent events both occurring is given by the general multiplication rule: $\Pr[A \text{ and } B] = \Pr[A]\Pr[B|A].$

431

- The conditional probability of an event is the probability of that event occurring given some condition.
- Probability trees and Bayes' theorem are important tools for calculations involving conditional probabilities.
- The law of total probability, $\Pr[A] = \sum_{\text{All values of } B} \Pr[B]\, \Pr[A|B]$, makes it possible to calculate the probability of an event (A) from all of the conditional probabilities of that event. The law multiplies, for all possible conditions (B), the probability of that condition $(\Pr[B])$ times the conditional probability of the event assuming that condition $(\Pr[A|B])$.

Online resources

Learning resources associated with this chapter are online at
https://whitlockschluter3e.zoology.ubc.ca/chapter05.html