CHAPTER 1

THINKING ABOUT CHANCE

In the introduction to this first part of the text, we learned that chance is used to select samples from the population that are, in the long run, representative of the population from which they came (Figure 1.1). Before we can appreciate how chance influences the composition of those samples, however, we need to understand some things about chance itself. In this chapter, we will look at the basic properties of chance and see how the chances of individual events can be combined to address health issues.

1.1 PROPERTIES OF PROBABILITY

To begin with, we should point out that there are two terms that can be used interchangeably: **chance** and **probability**. In everyday language, probability (or chance) tells us how many times something happens relative to the number of times it could happen. For example, we might think of the probability that a patient presenting

Introduction to Biostatistical Applications in Health Research with Microsoft® Office Excel®, First Edition. Robert P. Hirsch.

Companion Website: www.wiley.com/go/hirsch/healthresearch

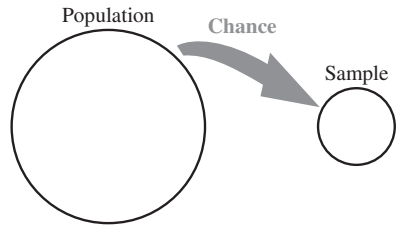

FIGURE 1.1 Chance determines which data values in the population end up in the sample.

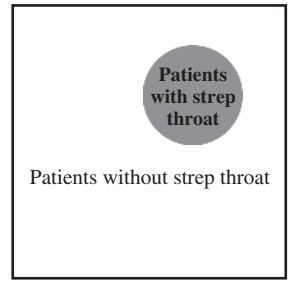

FIGURE 1.2 An example of a Venn diagram. The rectangular area represents all observations. The circular area represents the observations in which the event occurs. The area within the rectangle but outside of the circle represents those observations in which the event did not occur.

with a sore throat has streptococcal pharyngitis. If we can expect 1 patient to actually have streptococcal pharyngitis out of every 10 patients seen with a sore throat, then the probability of having streptococcal pharyngitis is 0.10. Or equivalently, there is a 10% chance that a person selected at random from among persons with sore throats would have strep throat.

In statistical terminology, the number of times something happens is called its **frequency** and that "something" is called an **event**. The opportunities for an event to occur are called **observations**.[1] When using the concept of probability, we need to understand that there are two possible results for each observation: either the event occurs or the event does not occur. In the previous example, the event was streptococcal pharyngitis and the patients seen with a sore throat were the observations.

Everyday language is often cumbersome when discussing issues in statistics. An alternative approach is to examine events and observations graphically. We do this by constructing a **Venn diagram**. In a Venn diagram, we use a rectangle to symbolize all of the observations and a circle to symbolize those observations in which the event occurs. Figure 1.2 is a Venn diagram we could use to think about the probability that a patient with a sore throat has streptococcal pharyngitis.

[1] Statisticians also refer to the opportunity for an event to occur as a **trial**. Since the term *trial* refers to a clinical experiment in health research, we will exclusively use the term *observation* to refer to the opportunity for an event to occur.

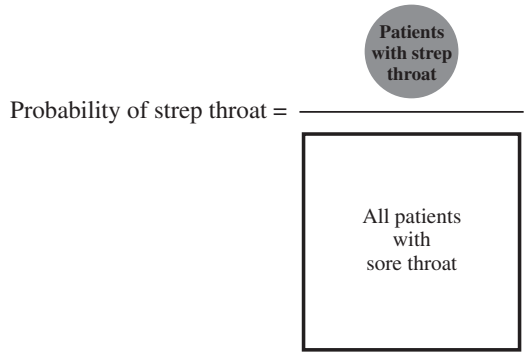

FIGURE 1.3 A Venn equation illustrating the probability a patient with sore throat has streptococcal pharyngitis.

There are some aspects of observations and events that are evident in a Venn diagram. For instance, we can see that the entire rectangle outside of the circle corresponds to observations in which the event does not occur. When an event does not occur, we say that the **complement** of the event occurs. In this case, the event is having strep throat and its complement is not having strep throat.

The way a Venn diagram tells us about the magnitude of the probability is by the area of the circle representing the event relative to the area of the entire rectangle. A way we can compare these areas is by creating a **Venn equation**. A Venn equation uses the parts of a Venn diagram in a mathematical equation that show how the probability of an event is calculated. For the probability that a patient with sore throat has streptococcal pharyngitis, the Venn equation would look like Figure 1.3.

A Venn equation helps us see another property of probabilities, that probabilities have a distinct range of possible values. Since an event cannot exist without an observation, the circle can only be as big as the rectangle. In other words, the numerator must be a subset of the denominator. The result of this property is to make the largest possible value for a probability equal to one (or 100%). The value of one occurs when every observation in the denominator is also an event in the numerator. When the probability of an observation being an event has a value of one, it is **certain** that the event will occur.

The numerator of a probability contains the number of events. The largest value possible is equal to the number of observations. The smallest value possible is zero. If the numerator of a probability is equal to zero, this implies that none of the observations are events and, therefore, the probability is equal to zero as well. A probability of zero indicates that it is **impossible** for an event to occur. A probability can be no smaller than zero and no larger than one.[2]

When we want to calculate a probability, it is easier to use some mathematical shorthand. To symbolize a probability, we use a lowercase p followed by a set of

[2] This range of possible values between zero and one means that a probability is also a proportion.

parentheses. Within those parentheses, we identify the event addressed by the probability. Then, the equation looks like this:

$$p(\text{event}) = \frac{\text{number of events}}{\text{number of observations}} \tag{1.1}$$

Next, let us take a look at an example that illustrates calculation of a probability and its interpretation.

■ Example 1.1

Suppose we did throat cultures for 100 patients who complained of a sore throat and 10 of those cultures were positive for streptococcus. What is the probability a person picked at random would have a positive strep culture?

In this question, a positive strep test is the event and someone with sore throat is an observation. To calculate the probability of a person having a positive strep culture, we can use Equation (1.1).

$$p(\text{event}) = \frac{\text{number of events}}{\text{number of observations}} = \frac{10}{100} = 0.1$$

Thus, there is a probability of 0.1 (or a 10% chance) that a person selected from the group of patients with a sore throat would be positive for streptococcus. ■

A part of the shorthand we use to show how probabilities are calculated concerns the complement of an event (i.e., an observation in which the event does not occur). Rather than inserting the description of the complement of the event within the parentheses, we more often put a bar over the description of the event. So, $p(\overline{\text{event}})$ stands for the probability of the complement of the event occurring (i.e., the probability of the event *not* occurring). For the complement of having strep throat, we could use $p(\overline{\text{strep}})$. There are two properties of a collection of events that an event and its complement always demonstrate. The first is **mutual exclusion**. A collection of events is said to be mutually exclusive if it is impossible for two or more events to occur in a single observation. In this case, it is certainly impossible for a person both to have strep throat and to not have strep throat.

The second property of an event and its complement is that they are **collectively exhaustive**. A collection of events is said to be collectively exhaustive if every observation is certain to consist of at least one of the events. Here, this implies that every person with a sore throat either has or does not have strep throat. Clearly, this is true.

For events that are both mutually exclusive and collectively exhaustive (like an event and its complement), there is a special relationship among the events: The sum of their probabilities is equal to one. In mathematical language, the relationship between the probability of an event occurring and the probability of the complement of the event occurring is shown in Equation (1.2):

$$p(\text{event}) + p(\overline{\text{event}}) = 1 \tag{1.2}$$

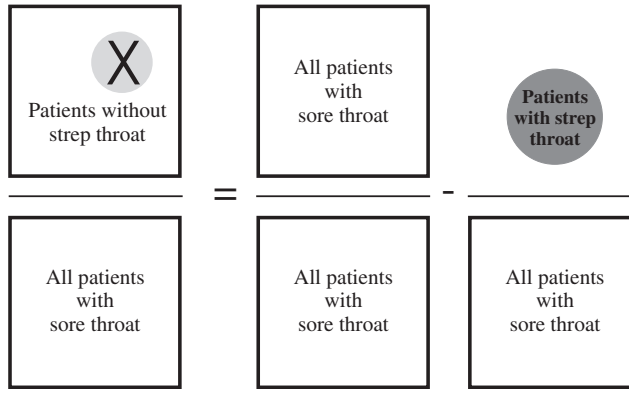

FIGURE 1.4 A Venn equation illustrating the relationship between the probability of the complement of the event (e.g., not having strep throat) and the probability of the event (e.g., having strep throat).

A little bit of algebra shows us that we can calculate the probability of the complement of an event by subtracting the probability of the event from one. This is shown in Equation (1.3):

$$p(\overline{\text{event}}) = 1 - p(\text{event}) \tag{1.3}$$

This relationship can also be described in graphic language as in the Venn equation in Figure 1.4.

So far, we have seen how we can think about probabilities using everyday language, graphic language, and mathematical language. Each one of these ways of examining statistical issues has its own advantages. The sort of things we have learned about probability includes the fact that probabilities have a discrete range of possible values ranging from zero (indicating that the event cannot occur) to one (indicating that the event always occurs). Also, we have examined the relationship between an event and its complement. This relationship has two important properties of a collection of events. These properties are mutually exclusive and collectively exhaustive. A collection of events is mutually exclusive if only one of the events can occur in a single observation. To be collectively exhaustive, the collection of events needs to encompass every possibility so that at least one of the events occurs in every observation. Next, we will take a look at other kinds of collections of events.

1.2 COMBINATIONS OF EVENTS

There are two ways we might be interested in how two or more events relate to each other. One way is that the events occur together in the same observation. We call this the **intersection** of events. Another way is that at least one event occurs in an observation. We call this the **union** of events.

1.2.1 Intersections

In health research and practice, we are often interested in situations in which more than one event occurs in a single observation. For instance, we might be interested in the relationship between a high-fat diet and development of atherosclerosis. The sorts of people in whom we would be most interested are those who have both of those events, since they are the ones for whom a high-fat diet could have contributed to the risk of disease.

In statistical terminology, we refer to the occurrence of two or more events in a single observation as the intersection of the events. Figure 1.5 illustrates the probabilities of a high-fat diet and atherosclerosis and the intersection of those two events. Their intersection is where the two circles overlap. These are the observations in which a person has both a high-fat diet and atherosclerosis.

The probability of an observation including both events (i.e., the intersection of those events) considers the size of the overlap relative to all the observations. Figure 1.6 shows a Venn equation representing the probability of the intersection of high-fat diet and atherosclerosis.

If we want to calculate the probability of an intersection of events, we use what is called the **multiplication rule**. To see how the multiplication rule works, let us begin with a Venn equation (Figure 1.7).

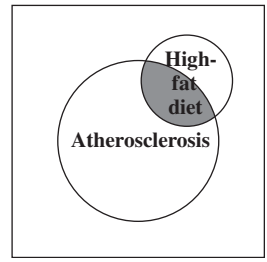

FIGURE 1.5 Venn diagram illustrating the relationship between a high-fat diet and development of atherosclerosis. The area in which the circles overlap represents those persons who have both a high-fat diet and atherosclerosis.

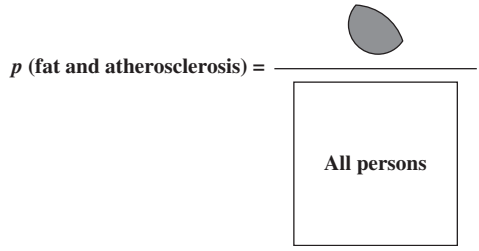

FIGURE 1.6 Venn equation for the probability that a person has both a high-fat diet (fat) and has atherosclerosis. In the numerator is the area of overlap (intersection) of the two circles in the Venn diagram (Figure 1.5). The denominator represents everyone whether or not they have a high-fat diet or atherosclerosis (i.e., the entire rectangle).

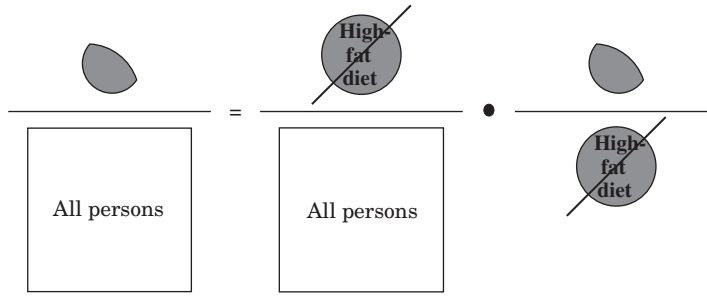

FIGURE 1.7 Venn equation of the multiplication rule used to calculate the intersection of high-fat diet and atherosclerosis.

To the left of the equals sign in the Venn equation in Figure 1.7 is the probability of the intersection of having a high-fat diet and developing atherosclerosis as shown in Figure 1.6. In the numerator of that probability are the persons who had both events. In the denominator are all persons regardless of diet or disease. Immediately to the right of the equals sign is the probability that someone has a high-fat diet. In the numerator of that probability are the persons with a high-fat diet and in the denominator are, as before, all persons regardless of diet or disease.

The second fraction to the right of the equals sign also is a probability[3], but it looks different from any probability we have encountered so far. Specifically, it does not include all the observations (represented by the rectangle in a Venn diagram) in its denominator. Rather, it includes only those persons with a high-fat diet in its denominator. This is an example of a very important kind of probability, called a **conditional probability**. A conditional probability tells us the probability of an event occurring given that another event has occurred. In this case, the conditional probability tells us the probability of a person having atherosclerosis given that the person has a high-fat diet.

In mathematical notation, a conditional probability also looks different from other probabilities we have encountered. Equation (1.4) illustrates the mathematical notation for the Venn equation in Figure 1.7.

$$p(A \text{ and } B) = p(A) \cdot p(B|A) \tag{1.4}$$

where

$p(A \text{ and } B)$ = probability that an observation will include both event A and event B (i.e., the probability of the intersection of A and B)[4]

$p(A)$ = probability that an observation includes event A (i.e., the unconditional probability of event A)

$p(B|A)$ = probability that an observation will include event B given that it includes event A (i.e., a conditional probability of event B)

[3] Recall that, to be a probability, a fraction's numerator must be a subset of its denominator. This is the case here, because those persons with both a high-cholesterol diet and atherosclerosis (the numerator) are all included in the circle representing persons with atherosclerosis (the denominator).

[4] In set notation, this is $p(A \cap B)$.

Or, in terms of a high-fat diet and atherosclerosis,

$$p(\text{fat and atherosclerosis}) = p(\text{fat}) \cdot p(\text{atherosclerosis} \mid \text{fat}) \tag{1.5}$$

From a statistical point of view, it does not matter which event is addressed by the conditional probability.[5] Thus, the probability of the intersection of high-fat diet and atherosclerosis could also be calculated as

$$p(\text{fat and atherosclerosis}) = p(\text{atherosclerosis}) \cdot p(\text{fat} \mid \text{atherosclerosis}) \tag{1.6}$$

In Equations (1.5) and (1.6), we can see that a vertical line is used to separate the two events in the parentheses of a conditional probability. The event to the left of the vertical line is called the **conditional event**. It is the conditional event that the probability addresses. In Equation (1.5), the conditional event is having atherosclerosis, so this conditional probability tells us about the chance that someone has atherosclerosis. The event to the right of the vertical line is called the **conditioning event**. The conditioning event defines the circumstance in which we are interested in the probability of the conditional event. Here, having a high-fat diet is the conditioning event. Thus, Equation (1.5) tells us that we are interested in the probability of having atherosclerosis given (i.e., under the condition) that someone has a high-fat diet.

The reason that conditional probabilities are so important in health research is the fact that they tell us about an important aspect of the relationship between events. Namely, conditional probabilities can be used to see if the occurrence of one event changes the probability of the occurrence of another event. If, for example, we are interested in whether there is this sort of relationship between a high-fat diet and having atherosclerosis, we could compare the conditional probability in Equation (1.5) with the probability that someone has atherosclerosis given that they do not have a high-fat diet ($p(\text{atherosclerosis} \mid \overline{\text{fat}})$). If those two conditional probabilities have the same value, then we can conclude that a high-fat diet does not influence the chance of having atherosclerosis. In that case, the three probabilities in Equation (1.7) are all equal to the same value.

$$p(\text{atherosclerosis} \mid \text{fat}) = p(\text{atherosclerosis} \mid \overline{\text{fat}}) = p(\text{atherosclerosis}) \tag{1.7}$$

Or, in more general terms,

$$p(B \mid A) = p(B \mid \overline{A}) = p(B) \tag{1.8}$$

[5] The way the probability of the intersection is calculated depends only on which probabilities are obtained as part of a particular health research study. If our information about the relationship between the high-fat diet and atherosclerosis comes from a cohort study (a study in which the probability of disease is compared between exposed and unexposed persons), for example, the conditional probability we would measure is the probability of the disease given exposure status. In a case–control study (a study in which the odds of being exposed is compared between persons who have and do not have the disease), however, the conditional probability we measure is the probability of the exposure given the disease status.

where

$p(B|\overline{A})$ = probability that an observation will include event B given that it does not include event A (i.e., another conditional probability of event B)

In statistical terminology, we say two events are **statistically independent** when the probability of one of the events is not affected by occurrence of the other event.[6] In biologic terms, events that are statistically independent cannot have a causal relationship (or any other type of relationship).

To determine if events are statistically independent, we need to compare only two of the three probabilities in Equation (1.8). If those two probabilities are equal to the same value, then all three probabilities are the same and the conditional and conditioning events are statistically independent. We will take a look at an example of this relationship shortly, but first let us see how conditional probabilities are calculated.

To calculate a conditional probability, we use Equation (1.4) algebraically rearranged as in Equation (1.9).

$$p(B|A) = \frac{p(A \text{ and } B)}{p(A)} \tag{1.9}$$

Or, in terms of a high-fat diet and having atherosclerosis,

$$p(\text{atherosclerosis}|\text{fat}) = \frac{p(\text{fat and atherosclerosis})}{p(\text{fat})} \tag{1.10}$$

This process of identifying statistical independence is illustrated in Example 1.2.

■ Example 1.2

Suppose that, in a particular valley of the Mojave Desert, there are 2500 residents. Of those 2500 residents, 625 work for ACME Borax, Inc., a company that recovers chemicals from the brine under a salt flat that covers most of the valley floor. Of the 2500 residents of the valley, 500 have been diagnosed with leukemia. Of the 500 diagnosed with leukemia, 125 are persons who work for ACME Borax, Inc. Given that information, is working for ACME statistically independent of being diagnosed with leukemia?

First, let us consider the relationship between working for ACME and having leukemia. We are told that 625 persons work for ACME and, of those, 125 have leukemia. From that information, we can calculate the probability of having leukemia

[6] The term "statistically independent" as statisticians use it can be confusing when we consider the everyday meaning of "independence." If we were to say, for example, that two persons are independent, we are likely to infer that there is no connection between them. This is not what the statistician is implying. Rather, the statistician is saying that you do not need to consider whether or not one event has occurred when addressing the probability of another event. When a statistician implies that there is no overlap between events, the statistician says that they are "mutually exclusive" rather than "statistically independent."

under the condition that a person works for ACME using Equation (1.9):

$$p(\text{leukemia} \mid \text{ACME}) = \frac{p(\text{leukemia and ACME})}{p(\text{ACME})} = \frac{125/2500}{625/2500} = 0.2$$

To determine if working for ACME and having leukemia are statistically independent events, we need to compare that conditional probability with either the probability of having leukemia given that a person does not work for ACME or with the overall (i.e., unconditional) probability of having leukemia. The latter probability is

$$p(\text{leukemia}) = \frac{\text{number with leukemia in valley}}{\text{total number in valley}} = \frac{500}{2500} = 0.2$$

Since these two probabilities are equal to the same value, we can conclude that working for ACME and having leukemia are statistically independent events. In other words, working for ACME does not change the probability of having leukemia. ■

So far, we have seen that we can use the multiplication rule to calculate the probability of two events occurring in a single observation (i.e., the intersection of those events). To calculate the probability of the intersection of more than two events, we simply include each additional event in the multiplication of conditional probabilities. For each additional event, we include the conditional probability of the event with the conditioning events being all of the events listed previously in the equation. For example, we can calculate the intersection of three events as shown in Equation (1.11).

$$p(A \text{ and } B \text{ and } C) = p(A) \cdot p(B \mid A) \cdot p(C \mid A \text{ and } B) \tag{1.11}$$

where

$p(C \mid A \text{ and } B)$ = probability that an observation will include event C given it includes events A and B (i.e., a conditional probability of event C)

If the events are statistically independent, we can use a simplified version of the multiplication rule. This simplification is to multiply the unconditional probabilities of the events. Equation (1.12) shows the simplified version for the intersection of three events examined in Equation (1.11).

$$p(A \text{ and } B \text{ and } C) = p(A) \cdot p(B) \cdot p(C) \tag{1.12}$$

The reason we can use this simplified version of the multiplication rule is that, by definition, the conditional and unconditional probabilities are the same for statistically independent events (as shown in Equation (1.8)). If the three events are not statistically independent, however, we need to use Equation (1.11) to calculate the intersection of events.

1.2.2 Unions

When our interest is in the probability of any (i.e., one or more) of a collection of events occurring in the same observation, we say we are interested in the **union** of those events. Suppose, for example, we are considering two risk factors for atherosclerosis: high-fat diet and smoking. In that case, we might be interested in calculating the probability a person has at least one of those risk factors (i.e., either high-fat diet or smoking or both high-fat diet and smoking). To illustrate this, let us add smoking to the Venn diagram in Figure 1.5. Then the Venn diagram of all three events will look something like the one in Figure 1.8.

The union of the two risk factors is satisfied if a person either has a high-fat diet or smokes (or both). Thus, the numerator of the probability of the union of those two events includes the part of the Venn diagram covered by either circle.

Figure 1.9 shows the Venn equation for the union of smoking and high-fat diet.

To calculate the probability of the union of two events, we use the **addition rule**. As the name implies, in the addition rule the probabilities of each of the events are added together.

Since adding the probabilities together includes the intersection of those events twice, the probability of the intersection of the events must be subtracted from the

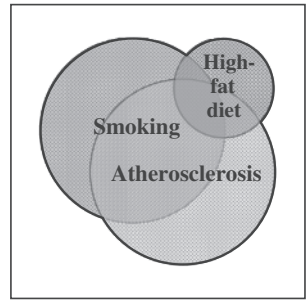

FIGURE 1.8 Venn diagram showing the relationship between high-fat diet, smoking, and atherosclerosis.

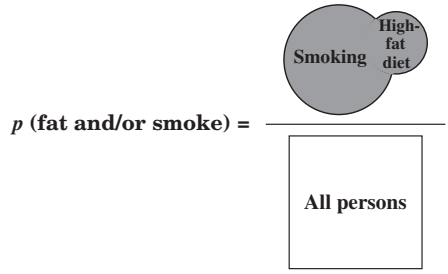

FIGURE 1.9 Venn equation showing the probability of the union of smoking and/or high-fat diet.

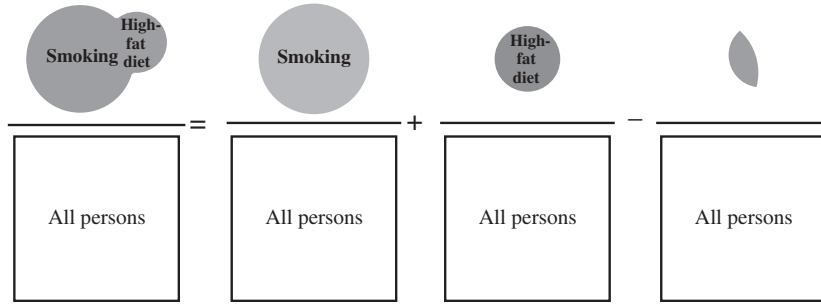

FIGURE 1.10 Venn equation showing calculation of the union of smoking and a high-fat diet using the addition rule.

sum. This calculation for the union of smoking and a high-fat diet is illustrated by the Venn equation in Figure 1.10.

In mathematical terms, the calculation of the union of two events is performed as shown in Equation (1.13).

$$p(A \text{ and/or } B) = p(A) + p(B) - P(A \text{ and } B) \tag{1.13}$$

where

$p(A \text{ and/or } B)$ = probability an observation will include event A and/or event B (i.e., the probability of the union of events A and B)[7]

Now, let us take a look at an example addressing the union of two events.

■ Example 1.3

Suppose we are planning a clinical trial of a new live vaccine. In this study, we want to exclude persons who are either pregnant or immunocompromised. Suppose we estimate that the population from which we are planning to take our sample includes 20% of the total number of persons who are pregnant and 10% of the total number of persons who are immunocompromised. If being pregnant and being immunocompromised are statistically independent events, what proportion of our sample will be excluded due to either of these characteristics?

To calculate the probability of the union of two events, we use Equation (1.13). For this application, Equation (1.13) looks like the following:

$$p(\text{preg and/or comp}) = p(\text{preg}) + p(\text{comp}) - p(\text{preg and comp})$$

We know the probability that a person selected at random from the population will be pregnant ($p(\text{Preg}) = 0.2$) and the probability that a person selected at random from the population will be immunocompromised ($p(\text{comp}) = 0.1$). We are not given the

[7] In set notation, this is $p(A \cup B)$.

probability of the intersection of these two events (i.e., the probability a person will be both pregnant and immunocompromised). We are told, however, that these two events are statistically independent. This tells us that we can use the simplified version of the multiplication rule illustrated for three statistically independent events in Equation (1.12). For the two events of being pregnant and being immunocompromised, their intersection can be calculated as follows:

$$p(\text{preg and comp}) = p(\text{preg}) \cdot p(\text{comp}) = 0.2 \cdot 0.1 = 0.02$$

Now that we have the probability of the intersection of the two events, we are ready to calculate their union.

$$\begin{aligned} p(\text{preg and/or comp}) &= p(\text{preg}) + p(\text{comp}) - p(\text{preg and comp}) \\ &= 0.2 + 0.1 - 0.02 = 0.28 \end{aligned}$$

Thus, we can expect that 28% of the persons we select from the population will be excluded from the study because they are either pregnant or immunocompromised (or both). ■

As with the multiplication rule we used to calculate the probability of the intersection of events, the addition rule for calculation of the probability of the union of events has a simplified version that can be used under a special condition. For the addition rule, the condition is that the events are mutually exclusive. If so, the probability of the union of events can be calculated by simply adding together the probabilities of the events. The intersections of the events do not need to be subtracted from that sum because, by definition, the probability of the intersection of two mutually exclusive events is equal to zero.

1.3 BAYES' THEOREM

Earlier, we learned there are two types of events in a conditional probability: the conditional event(s) and the conditioning event(s). We also learned that these types of events have very different roles in a conditional probability. The conditional event is the event for which the probability is calculated (i.e., conditional probabilities tell us the chance of the conditional event occurring). All of the characteristics of unconditional probabilities (those discussed at the beginning of this chapter) apply to the conditional event. For instance, the probability of the complement of the conditional event is found by subtracting the conditional probability from one (see Equation (1.3)). Equation (1.14) shows that relationship for conditional probabilities:

$$p(\overline{A} \mid B) = 1 - p(A \mid B) \tag{1.14}$$

The conditioning event defines the condition under which we are interested in the probability of the conditional event. None of the characteristics of unconditional

probabilities discussed at the beginning of this chapter apply to the conditioning event. For example, we cannot find the probability of the conditional event given that the conditioning event does not occur by subtracting from one the conditional probability given that the conditioning event occurs. Equation (1.15) shows this inequality in mathematical notation:

$$p(A \mid \overline{B}) \neq 1 - p(A \mid B) \tag{1.15}$$

So, there are important differences in the way conditional and conditioning events affect interpretation of conditional probabilities. Under most circumstances, we need to only keep these differences in mind. Under some circumstances, however, the conditional probabilities we know something about have the conditional and conditioning events reversed relative to our interest. Examples of such "backward" conditional probabilities are the sensitivity and specificity of a diagnostic test. Sensitivity tells us the probability that a person with a particular disease (D) will have a positive test result (T). Specificity tells us the probability that a person without that disease ($\overline{D}$) will have a negative test result ($\overline{T}$). In mathematic notation, Equations (1.16) and (1.17) describe the sensitivity and specificity of a diagnostic test.

$$\text{Sensitivity} = p(T \mid D) \tag{1.16}$$

$$\text{Specificity} = p(\overline{T} \mid \overline{D}) \tag{1.17}$$

Since the conditioning event is having the disease, sensitivity can only be interpreted for those persons known to have the disease. Likewise, specificity can only be interpreted for those persons known not to have the disease. When a diagnostic test is used, however, it is not known whether or not the person has the disease. What is known is whether the test has a positive or negative result. To interpret a diagnostic test, we need to interchange the conditional and conditioning events in sensitivity and specificity.[8] The way we do this is by using Bayes' theorem. Equation (1.18) shows Bayes' theorem in general terms.

$$P(B \mid A) = \frac{p(B) \cdot p(A \mid B)}{[p(B) \cdot p(A \mid B)] + [p(\overline{B}) \cdot p(A \mid \overline{B})]} \tag{1.18}$$

where

$p(B \mid A)$ = probability of event B occurring given event A has occurred
$p(B)$ = unconditional probability of event B occurring
$p(A \mid B)$ = probability of event A occurring given event B has occurred
$p(\overline{B})$ = unconditional probability of event B not occurring
$p(A \mid \overline{B})$ = probability of event A occurring given event B has not occurred

[8] Since sensitivity and specificity are "backward" conditional probabilities, you might wonder why we use them to address the performance of a diagnostic test. The reason is the way studies examine diagnostic tests. In one study, the test is used on persons with the disease. That study estimates sensitivity. In another study, the test is used on persons without the disease. That study estimates specificity.

Example 1.4 applies Bayes' theorem to sensitivity and specificity.

■ **Example 1.4**

Suppose we are screening a particular population for cervical cancer. In that population, one out of 1000 women has cervical cancer. The diagnostic test we use for screening has a sensitivity of 0.9 and a specificity of 0.7. What is the probability a person with a positive test result really has cervical cancer?

To begin with, let us take a look at the information we have. Knowing that one out of 1000 women in the population have cervical cancer tells us that the probability that any particular woman has cervical cancer is 0.001. A sensitivity of 0.9 implies that a person with cervical cancer has a 90% chance of having a positive test result and a specificity of 0.7 implies that a person without cervical cancer has a 70% chance of having a negative test result. In mathematic notation, we know

$$\begin{aligned} p(D) &= 0.001 \\ p(T\,|\,D) &= 0.9 \\ p(\overline{T}\,|\,\overline{D}) &= 0.7 \end{aligned}$$

Our interest is in the probability of a person with a positive test result has cervical cancer. At first this sounds like the sensitivity, but it is not the same thing. In sensitivity, our interest is confined to people who have the disease. To interpret a positive test result however, we need to confine our interest to persons with a positive result. Thus, our interest is in the same events that make up sensitivity, but with the conditional and conditioning events transposed. We can transpose conditional and conditioning events by using Bayes' theorem (from Equation (1.18)).

$$\begin{aligned} p(D|T) &= \frac{p(D)\cdot p(T\,|\,D)}{[\,p(D)\cdot p(T\,|\,D)] + [\,p(\overline{D})\cdot p(T\,|\,\overline{D})]} \\ &= \frac{0.001\cdot 0.9}{[0.001\times 0.9] + [\{1-0.001\}\times\{1-0.7\}]} = 0.003 \end{aligned}$$

So, the probability that a woman with a positive test result has cervical cancer is 0.003. You might be surprised that this probability is so low. The principal reason for this is the fact that cervical cancer occurs in only one out of 1000 women. Among women with positive test result, this changes to three out of 1000. So, the chance of cervical cancer is three times as great among women with a positive test result, but it is still a low probability (i.e., 0.003). Bayes' theorem has helped us to appreciate the meaning of a positive test result when the diagnostic test is used for screening this population. ■

Now that we have an understanding of probabilities, we are ready to apply what we have learned to the process of taking samples from populations. In the next chapter, we will begin doing this by focusing on populations.