4

Probability and Life Tables

Chapter Outline

As was mentioned in Chapter 3, often we want to do more than simply analyze or summarize the data in graphs or statistics. For example, we may want to determine whether two drugs or treatments are equally effective and safe or whether the age-adjusted death rates for two areas are the same. To answer these questions, we require knowledge of *probability*, the topic of this chapter.

4.1 A Definition of Probability

We have all encountered the use of probability — in the weather forecast, for example. The forecast usually involves an estimate of the probability of rain, as in the statement that "the probability of rain tomorrow is 20 percent." As its use in the weather forecast demonstrates, *probability* is a numerical assessment of the likelihood of the occurrence of an outcome of a random variable. In the weather forecast, weather is the random variable and rain is one of its possible outcomes.

Before considering the numerical assessment of likelihood, we should consider random variables. There are both discrete and continuous random variables. A *discrete* (nominal, categorical or ordinal) *random variable* is a quantity that reflects an attribute or characteristic that takes on different values with specified probabilities. A *continuous* (interval or ratio) *random variable* is a quantity that reflects an attribute or characteristic that falls within an interval with specified probabilities.

Hypertension status is a discrete random variable when the values or levels of this variable are defined as its presence (can be defined as systolic blood pressure greater than 140 mmHg, diastolic blood pressure greater than 90 mmHg, or taking antihypertensive medication) or absence. Other examples of discrete random variables include racial status, the number of children in a family, and type of health insurance. Examples of continuous random variables include height, blood pressure, and the amount of lead emissions as they are usually measured.

Table 4.1 Percent of population in selected racial groups: United States, 2000.

Race	Number	Percent
Total	281,421,906	100.0
White	211,460,626	75.1
Black or African American	34,658,190	12.3
Asian and Pacific Islander	10,641,833	3.8
American Indian and Alaskan Native	2,475,956	0.9
Some other races	15,359,073	5.5
Two or more races	6,826,228	2.4

Source: U.S. Bureau of the Census, 2000

We shall define probability of the occurrence of an outcome or interval of a random variable as its relative frequency in an infinite number of trials or in a population. A probability is a population *parameter*. An observed proportion (relative frequency) from a sample is a *statistic* that can be used to estimate a probability. We shall use the data in Table 4.1 to demonstrate the calculation of the probability of different racial categories in the United States in 2000. As shown in Table 4.1, there are four major racial groups used in the U.S. Population Census and a fifth category that combines all other races. Those who claimed two or more races are in the sixth category.

The probability of a person selected at random being white was 0.751 (= 211460626 /281421906), or 75.1 percent. The corresponding probabilities of being black, Asian and Pacific Islander, American Indian, and some other races were 0.123, 0.038, 0.009, and 0.055, respectively. Finally, the probability of a person claiming two or more races was 0.024. These six probabilities sum to 1.000 or 100.0 percent, as shown in Table 4.1.

Since a probability is the number of occurrences of an outcome divided by the total number of occurrences of all possible outcomes of the variable under study, this means that a probability cannot be larger than 1.00 or 100 percent in value. By the same reasoning, a probability cannot be smaller than 0.00 or 0 percent in value. Therefore, the only valid values for probabilities range from 0 to 1 or 0 to 100 percent. Additionally, use of the relative frequency definition means that the sum of the probabilities of all the possible outcomes of a random variable must be 1.00 or 100 percent. If a probability falls outside the 0 to 1 range, or if the sum of the probabilities of all the possible outcomes of a variable do not sum to 1 (with allowance for rounding), a mistake has been made.

For many variables in the health field, the probability of an outcome is estimated from a large number of observations and may change over time. For example, the probabilities of the different racial groups in the United States in 2020 will be different from the 2000 probabilities. As an additional example of changing probabilities, the estimates of the age-adjusted probabilities of obese persons (body mass index greater than or equal to 30) among U.S. adults (ages 20–74 years) increased from 0.151 in 1976–1980 to 0.233 in 1988–1994 and to 0.311 in 1999–2002 (NCHS 2004). This change in the values of a probability contrasts with the lack of change in the probabilities associated with physical phenomena, such as tossing a coin or a pair of dice. For example, when a fair coin is tossed, the probability of a head is assumed to be 0.5 or 50 percent, and it does not change.

The listing of the probabilities of all possible outcomes of a discrete variable is its *probability distribution*. For example, the probability distribution of the racial composition of the U.S. population in 2000 is shown in the last column of Table 4.1. More will be said about probability distributions and their use in the next chapter.

4.2 Rules for Calculating Probabilities

A few basic rules govern the calculation of probabilities of compound outcomes or events, and we will use the data in Table 4.2 to explain them. Entries in Table 4.2 are the number of live births by birth weight and the trimester in which prenatal care was begun. For example, the entry in the third row and the second column, 3271, is the number of live births to women who had begun their prenatal care during their second trimester and whose babies' birth weights were greater than 7.7 lb.

4.2.1 Addition Rule for Probabilities

The data in Table 4.2 can be used to determine whether or not there is a relation between the timing of the beginning of prenatal care and birth weight. However, before examining this issue, let us calculate a few additional probabilities. For example, the probability of a woman in Harris County, Texas, in 1986 having a low birth weight baby (less than or equal to 5.5 lb) was 0.069 (= 3541/51473). This value is very close to the 1986 value of 0.068 for the United States (NCHS 1992). Let us consider a slightly more complex example. The probability of late prenatal (third trimester) or no prenatal care is simply the sum of their individual probabilities, that is, 2337/51473 + 1695/51473 which is 0.078 (= 4032/51473). This value is slightly greater than the corresponding 1986 U.S. value of 0.060 (NCHS 1992). In these calculations of probabilities, we are considering births in Harris County, Texas, in 1986 as our population. If the intended population were Texas or the United States, then the preceding values would be sample estimates — that is, observed proportions — of the probabilities. However, a sample consisting of births in Harris County should not be used to draw inferences about births in Texas or the United States because the Harris County births are likely not to be representative of either of these two larger units.

So far, these probabilities have focused on row or column totals (marginal totals), not on the numbers in the interior of the table (cell entries). Entries in the interior of the table deal with the intersection of outcomes or events. For example, the outcome of a woman having a live birth of less than or equal to 5.5 lb and having begun her prenatal care during the first trimester is the intersection of those two individual outcomes. The

Table 4.2 Number of live births by birth weight and trimester of first prenatal care: Harris County, Texas, 1986 (excluding 1,180 births with unknown birth weight or trimester of first prenatal care).

	Trimester Prenatal Care Began				
Birth Weight	**1st**	**2nd**	**3rd**	**No Care**	**Total**
≤5.5 lb; ~2,500 g	2,412	754	141	234	3,541
5.6–7.7 lb; ~2,500–3,500 g	20,274	5,480	1,458	1,014	28,226
>7.7 lb; ~3,500 g	15,250	3,271	738	447	19,706
Total	37,936	9,505	2,337	1,695	51,473

Source: Harris County Health Department, 1990, Table 1.S probability of this intersection — that is, of these two outcomes occurring together — is easily found to be 0.047 (= 2412/51473).

We just found the probability of a baby weighing less than or equal to 5.5 lb by using the row total of 3541 and dividing it by the grand total of 51,473. Note that we can also express this probability in terms of the probability of the intersection of a birth weight of less than or equal to 5.5 lb with each of the prenatal care levels — that is,

$$\begin{aligned}\Pr\{\leq 5.5\,\text{lb}\} &= \Pr\{\leq 5.5\,\text{lb \& 1st trim.}\} + \Pr\{\leq 5.5\,\text{lb \& 2nd trim.}\}\\ &\quad + \Pr\{\leq 5.5\,\text{lb \& 3rd trim.}\} + \Pr\{\leq 5.5\,\text{lb \& no care}\}\\ &= \frac{2412}{51473} + \frac{754}{51473} + \frac{141}{51473} + \frac{234}{51473} = \frac{3541}{51473}.\end{aligned}$$

This can be expressed in symbols. Let A represent the outcome of a birth weight less than or equal to 5.5 lb and B_i, $i = 1$ to 4, represent the four prenatal care levels. The symbol $\cap$ is used to indicate the intersection (to be read as "and") of two individual outcomes. Then we have

$$\Pr\{A\} = \Pr\{A \cap B_1\} + \Pr\{A \cap B_2\} + \Pr\{A \cap B_3\} + \Pr\{A \cap B_4\}$$

which, using the summation symbol, is

$$\Pr\{A\} = \sum_i \Pr\{A \cap B_i\}. \tag{4.1}$$

Suppose now that we want to find for a woman who had a live birth the probability that either the birth weight was 5.5 lb or less or the woman began her prenatal care during the first trimester. It is tempting to add the two individual probabilities — of a birth weight less than or equal to 5.5 lb and of prenatal care beginning during the first trimester — as we had done previously. However, if we added the entries in the first row (birth weights less than or equal to 5.5 lb) to those in the first column (prenatal care begun during the first trimester), the entry in the intersection of the first row and column would be included twice. Therefore, we have to subtract this intersection from the sum of the two individual probabilities to obtain the correct answer. The calculation is

$$\begin{aligned}\Pr\{\leq 5.5\,\text{lb or 1st trim.}\} &= \Pr\{\leq 5.5\} + \Pr\{\text{1st trim.}\} - \Pr\{\leq 5.5\,\text{lb and 1st trim.}\}\\ &= \frac{3541 + 37936 - 2412}{51473} = 0.759.\end{aligned}$$

This can be succinctly stated in symbols. Let A represent the outcome of live births of 5.5 lb or less and B represent the outcome of the initiation of prenatal care during the first trimester. An additional symbol $\cup$ is used to indicate the union (to be read as "or") of two individual outcomes. The intersection of these two outcomes is represented by $A \cap B$. In symbols, the rule is

$$\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}. \tag{4.2}$$

This rule also was used in the earlier example of late or no prenatal care, but, in that case, the outcomes were disjointed — that is, there was no overlap or intersection. Hence, the probability of the intersection was zero.

As the sum of the probabilities of all possible outcomes is one, if there are only two possible outcomes — say, A and not A (represented by $\bar{A}$) — we also have the following relationship:

$$\Pr\{A\} = 1 - \Pr\{\bar{A}\}. \tag{4.3}$$

4.2.2 Conditional Probabilities

Suppose we change the wording slightly in the preceding example. Based on the data in Table 4.2, we now want to find the probability of a woman having a live birth of less than or equal to 5.5 lb (event A) *conditional on or given* that her prenatal care was begun during the first trimester (event B). The word *conditional* limits our view in that we now focus on the 37,936 women who began their prenatal care during the first trimester. Thus, the probability of a woman having a live birth weighing less than or equal to 5.5 lb, given that she began her prenatal care during the first trimester, is 0.064 (= 2412 /37936). Dividing both the numerator and denominator of this calculation by 51473 (the total number of women) does not change the value of 0.064, but it allows us to define this *conditional probability* (the probability of A conditional on the occurrence of B) in terms of other probabilities. The numerator divided by the total number of women (2412 /51473) is the probability of the intersection of A and B, and the denominator divided by the total number of women (37936/51473) is the probability of B. In symbols, this is expressed as

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} \tag{4.4}$$

where $\Pr\{A \mid B\}$ represents the probability of A given that B has occurred.

Conditional probabilities often are of greater interest than the unconditional probabilities we have been dealing with as will be shown following. Before doing that, note that we can use the conditional probability formula to find the probability of the intersection — that is,

$$\Pr\{A \cap B\} = \Pr\{A \mid B\} \cdot \Pr\{B\}. \tag{4.5}$$

Thus, if we know the probability of A conditional on the occurrence of B, and we also know the probability of B, we can find the probability of the intersection of A and B. Note that we can also express the probability of the intersection as

$$\Pr\{A \cap B\} = \Pr\{B \mid A\} \cdot \Pr\{A\}. \tag{4.6}$$

Table 4.3 repeats the data from Table 4.2 along with three different sets of probabilities. The first set of probabilities (row R) is conditional on the birth weight; that is, it uses the row totals as the denominators in the calculations. The second set (row C) is conditional on the trimester that prenatal care was begun; that is, it uses the column totals in the denominator. The third set of probabilities (row U) is the unconditional set — that is, those based on the total of 51,473 live births. The probabilities in the Total column are the probabilities of the different birth weight categories; that is, the probability distribution of the birth weight variable and those beneath the Total row are the probabilities of the different trimester categories — that is, the probability distribution of the prenatal care variable. As just mentioned, these probabilities are based on the population of births in Harris County, Texas, in 1986.

Table 4.3 Number and probabilities of live births by trimester of first prenatal care and birth weight: Harris County, Texas, 1986.

		Trimester Prenatal Care Began				
Birth Weight		**1st**	**2nd**	**3rd**	**No Care**	**Total**
≤5.5 lb; ~2,500 g		2,412	754	141	234	3,541
	R[a]	0.681	0.213	0.040	0.066	
	C	**0.064**	**0.079**	**0.060**	**0.138**	**0.069**
	U	0.047	0.015	0.003	0.005	
5.6–7.7 lb; ~2,500–3,500 g		20,274	5,480	1,458	1,014	28,226
	R	0.718	0.194	0.052	0.036	
	C	**0.534**	**0.577**	**0.624**	**0.598**	**0.548**
	U	0.394	0.106	0.028	0.020	
>7.7 lb; ~3,500 g		15,250	3,271	738	447	19,706
	R	0.774	0.166	0.037	0.023	
	C	**0.402**	**0.344**	**0.316**	**0.264**	**0.383**
	U	0.296	0.064	0.014	0.009	
Total		37,936	9,505	2,337	1,695	51,473
	R	0.737	0.185	0.045	0.033	1.000

[a] R, row; C, column; and U, unconditional

Let us consider the entries in the row 1, column 1 cell. The first two entries below the frequency of the cell are conditional probabilities. The value 0.681 (= 2412/3541) is the probability based on the row total; that is, it is the probability of a woman having begun her prenatal care during the first trimester given that the baby's birth weight was less than or equal to 5.5 lb. The value 0.064 (= 2412/37936) is the probability based on the column total; that is, it is the probability of a birth weight of less than or equal to 5.5 lb given that the woman had begun her prenatal care during the first trimester. The last value, 0.047 (= 2412/51473), is the unconditional probability; that is, it is based on the grand total of 51,473 live births. It is the probability of the intersection of a birth weight less than or equal to 5.5 lb with the prenatal care having been begun during the first trimester.

As Table 4.3 shows, at least three different probabilities, or observed proportions if the data are a sample, can be calculated for the entries in the two-way table. The choice of which probability (row, column, or unconditional) to use depends on the purpose of the investigation. In this case, the data may have been tabulated to determine whether or not the timing of the initiation of the prenatal care had any effect on the birth weight of the infant. If this is the purpose of the study, the column-based probabilities may be the more appropriate to use and report. The column-based calculations give the probabilities of the different birth weight categories conditional on when the prenatal care was begun. The row-based calculations give the probability of the trimester prenatal care was initiated given the birth weight category. However, these row-based probabilities are of no interest because the birth weight cannot affect the timing of the prenatal care. The unconditional probabilities are less informative in this situation, as they also reflect the row and column totals. For example, compare the unconditional probabilities in the first and third columns in the first row — 0.047 and 0.003. Even though we have seen that there is little difference in the corresponding column-based probabilities of 0.064 and 0.060, these unconditional values are very different. The value of 0.047 is larger mainly because there are 37,936 live births in the first column compared to only 2337 live births in the third column. However, the unconditional probabilities may be useful in planning and allocating resources for maternal and child health services programs.

Using the column-based values, women who began their prenatal care during the first trimester had a probability of a low birth weight baby of 0.064. This value is compared to 0.079, the probability of a low birth weight baby for those who began their prenatal care during their second trimester, to 0.060 for those who began their prenatal care during the third trimester, and to 0.138 for those who received no prenatal care. There is little difference in the probabilities of a low birth weight baby among women who received prenatal care. However, the probability of a low birth weight baby is about twice as large for women who received no prenatal care compared to women who received prenatal care. The effect of prenatal care is most clearly evident in the probability of having a baby with a birth weight of greater than 7.7 lb. In this category, the probabilities are 0.402, 0.344, 0.316, and 0.264 for the first, second, and third trimesters and no prenatal care, respectively.

Based on the trend in the probabilities of a birth weight of greater than 7.7 lb, one might conclude that there is an effect of prenatal care. However, to do so is inappropriate without further information. First, although these births can be viewed as constituting a population — that is, all the live births in Harris County in 1986 — they could also be viewed as a sample in time, one year selected from many, or in place, one county selected from many. From the perspective that these births are a sample, there is sampling variation to be taken into account, and this will be covered in Chapter 11. Second, and more important, these data do not represent a true experiment. Chapter 6 presents more on experiments, but, briefly, the women were not randomly assigned to the different prenatal care groups — that is, to the first, second, or third trimester groups or to the no prenatal care group. Thus, the women in these groups may differ on variables related to birth weight — for example, smoking habits, amount of weight gained, and dietary behavior. Without further examination of these other factors, it is wrong to conclude that the variation in the probabilities of birth weights is due to the time when prenatal care was begun.

Example 4.1

Suppose that a couple has two children and one of them is a boy. What is the probability that both children are boys? For a couple with two children, there are four possible outcomes: boy and boy, boy and girl, girl and boy, girl and girl. If one of the two children is a boy, then there are three possible outcomes, excluding the (girl and girl) outcome. Therefore, the probability of having two boys is 1/3 (one of three possible outcomes). Applying the conditional probability rule, Equation (4.4), we can calculate this probability by (1/4) / (1 – 1/4).

4.2.3 Independent Events

Suppose we were satisfied that there are no additional factors of interest in the examination of prenatal care and birth weight and only the data in Table 4.2 were to be used to determine whether or not there was a relation between when prenatal care was initiated and birth weight. Row C in Table 4.3 shows the column-based probabilities — that is, those conditional on which trimester care was begun or whether care was received — and these are the probabilities to be used in the study.

If there was no relationship between the prenatal care variable and the birth weight variable — that is, if these two variables were *independent* — what values should the column-based probabilities have? If these variables are independent, this means that the birth weight probability distribution is the same in each of the columns. The last column in Table 4.3 gives the birth weight probability distribution, and this is the distribution that will be in each of the columns if the birth weight and prenatal care variables are independent. Table 4.4 shows the birth weight probability distribution for the situation when these two variables are independent.

Table 4.4 Probabilities conditional on trimester under the assumption of independence of birth weight level and trimester of first prenatal care: Harris County, Texas, 1986.

	Trimester Prenatal Care Began				
Birth Weight	**1st**	**2nd**	**3rd**	**No Care**	**Total**
≤ 5.5 lb; ~2500 g	0.069	0.069	0.069	0.069	0.069
5.6–7.7 lb; ~2500–3500 g	0.548	0.548	0.548	0.548	0.548
> 7.7 lb; ~3500 g	0.383	0.383	0.383	0.383	0.383
Total	1.000	1.000	1.000	1.000	1.000

The entries in Table 4.4 are conditional probabilities — for example, of a birth weight less than or equal to 5.5 lb (A) given that prenatal care began during the first trimester (B) under the assumption of independence. Hence, under the assumption of independence of A and B, the probability of A given B is equal to the probability of A. In symbols, this is

$$\Pr\{A \mid B\} = P\{A\}$$

when A and B are independent. Combining this formula with the formula for the probability of the intersection — that is,

$$\Pr\{A \cap B\} = \Pr\{A \mid B\} \cdot \Pr\{B\}$$

yields

$$\Pr\{A \cap B\} = \Pr\{A\} \cdot \Pr\{B\}$$

when A and B are independent.

Example 4.2

For a couple with one child, there are two possible outcomes: a boy or a girl. It is assumed that the probability of a girl is the same as the probability of a boy — that is, 0.5. For a couple with two children, there are four possible outcomes, as seen in Example 4.1. The probability of each outcome is 0.25 (one out of the four possible outcomes). We have to realize that the probability of having one boy and one girl is 0.5, accounting for two of four possible outcomes. However, U.S. vital statistics consistently show that about 105 boys are born per 100 girls (a sex ratio at birth of 105; Mathews and Hamilton 2005), which suggests that the probability of having a boy is 0.51 and the probability of having a girl is 0.49. If we use these values of the case of two children, the probabilities of the four possible outcomes are 0.26 (= 0.51 [0.51]), 0.25 (= 0.51 [1 − 0.51]), 0.25 (= [1 − 0.51] 0.51), and 0.24 (= [1 − 0.51] [1 − 0.51]), respectively. The four probabilities add up to 1. The reason for multiplying two probabilities will become clearer in Example 4.3.

Example 4.3

Since the gender of the second child is independent of that of the first child, the probability of two boys in row, based on vital statistics, is 0.51(0.51) = 0.26, as shown in Example 4.2. When considering diseases, it is unlikely that the disease status of one person is independent of that of another person for many infectious diseases. However, it is likely that the disease status of one person is independent of that of another for many chronic diseases. For example, let π be the probability that a person has Alzheimer's disease. One person's Alzheimer's status should be independent of another's status. Therefore, the probability of two persons having Alzheimer's disease is the product of the probabilities of either having the disease — that is, $\pi \cdot \pi$.

Establishing the dependence (a relation exists) or independence (no relation) of variables is what much of health research is about. For example, in the disease context, is disease status related to some variable? If there is a relation (dependency), the variable is said to be a risk factor for the disease. The identification of risk factors leads to strategies for preventing or reducing the occurrence of the disease.

Example 4.4

Let us apply these definitions of probabilities to the example used for the randomized response technique in Chapter 2. In Figure 2.2 there were 12 yes responses among 36 individuals to whom the randomized response technique was administered. We can denote as the probability of yes, Pr(Y) = 12/36 = 1/3. We know this observed probability is a combination of probabilities under two circumstances — that is, Pr($\underline{\text{H}}$ead and $\underline{\text{D}}$runken driving) + Pr($\underline{\text{T}}$ail and $\underline{\text{B}}$orn in September or October). In symbols, this relationship is expressed as

$$\Pr\{Y\} = \Pr\{H \cap D\} + \Pr\{T \cap B\}.$$

The two probabilities of intersection in the right hand side of equation can be expressed in terms of conditional probabilities, applying Equation (4.6) as follows:

$$\Pr\{Y\} = \Pr\{D \mid H\} \cdot \Pr\{H\} + \Pr\{B \mid T\} \cdot \Pr\{T\}.$$

We know that Pr(H) = ½ and Pr(T) = 1 – Pr(H) = ½. Pr($D \mid H$) is unknown quantity, and we want to estimate this conditional probability. Pr($B \mid T$) is known — that is, 2 months out of 12 months, 2/12 = 1/6.

If we solve the above equation for the unknown probability, Pr($D \mid H$), then we have

$$\Pr(D|H) = \frac{\Pr(Y) - \Pr(B|T)\cdot\Pr(T)}{\Pr(H)} = \frac{\left(\frac{1}{3}\right) - \left(\frac{1}{6}\right)\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)} = \left(\frac{2}{3}\right) - \left(\frac{1}{6}\right) = \left(\frac{1}{2}\right).$$

The result is 50 percent, which is the same as in Figure 2.2.

4.3 Definitions from Epidemiology

There are many quantities used in epidemiology that are defined in terms of probabilities, particularly conditional probabilities. Several of these useful quantities are defined in this section and used in the next section to illustrate Bayes' rule.

4.3.1 Rates and Probabilities

Various rates and relative numbers are used in epidemiology. A rate is generally interpreted as a probability — as a measure of likelihood that a specified event occurs to a specified population. *Prevalence* of a disease is the probability of having the disease. It is the number of people with the disease divided by the number of people in the defined population. The observed proportion of those with the disease in a sample is the sample estimate of prevalence. When the midyear population is used for the denominator, it is possible that the numerator contains persons not included in the denominator. For example, persons with the disease that move into the area in the second half of the year are not counted in the denominator, but they are counted in the numerator. When prevalence or other quantities use midyear population or person-years lived values, they are not really probabilities or proportions, although this distinction usually is unimportant. However, this distinction is important when estimating the probability of dying from the age-specific death rate as will be discussed later in conjunction with the life table.

Incidence of a disease is the probability that a person without the disease will develop the disease during some specified interval of time. It is the number of new cases of the disease that occur during the specified time interval divided by the number of people in the population who do not already have the disease.

Prevalence provides an idea of the current magnitude of the disease problem, whereas incidence informs as to whether or not the disease problem is getting worse.

Example 4.5

We consider the incidence and prevalence rates of AIDS based on the data from *Health, United States*, *2004* (NCHS 2004, Tables 1 and 52). By the end of 2002, 829,998 cases of AIDS had been reported to the Centers for Disease Control and Prevention, and of those, 42,478 cases were reported in 2002. The estimated U.S. population as of July 1, 2002, was 288,369,000. Based on the cases reported in 2002 and the estimated midyear population of 2002, the 2002 incidence rate of AIDS in the United States was 0.00014730 (42478/288369000) or 14.7 per 100,000 population. Since the rate is low, the rate is expressed as the number of cases per 100,000.

Based on the preceding data, it is difficult to estimate the prevalence rate because there is no information on the number of individuals with AIDS who had died prior to 2002. The AIDS death rate was reported starting in 1987. It steadily increased from 5.6 per 100,000 to 16.2 per 100,000 in 1995 and steadily declined to 4.9 per 100,000 in 2002. Based on an average death rate of AIDS for the last two and half decades, it is roughly estimated that about 80 percent of those diagnosed prior to 2002 had died by the end of 2001. Thus, of the 874,230 reported cases, we are assuming that 630,016 (0.8{874230 – 42479}) had died, leaving 199,982 persons with AIDS in 2002. The prevalence rate of AIDS then was 0.00069349 (= 199982/288369000) or 69.3 per 100,000 population.

A *specific rate* of a disease is the disease rate for people with a specified characteristic, such as age, race, sex, occupation, and so on. It is the conditional probability of a person having the disease given that the person has the characteristic. For example, an age-specific death rate is a death rate conditional on a specified age group, as seen in Chapter 3.

4.3.2 Sensitivity, Specificity, and Predicted Value Positive and Negative

Laboratory test results are part of the diagnostic process for determining if a patient has some disease. Unfortunately in many cases, a positive test result — that is, the existence of an unusual value — does not guarantee that a patient has the disease. Nor does a negative test result, the existence of a typical value, guarantee the absence of the disease. To provide some information on the accuracy of testing procedures, their developers use two conditional probabilities: sensitivity and specificity.

The *sensitivity* of a test (symptom) is the probability that there was a positive result (the symptom was present) given that the person has the disease. The *specificity* of a test (symptom) is the probability that there was a negative result (the symptom was absent) given that the person does not have the disease. Note that one minus sensitivity is the false negative rate, and one minus specificity is the false positive rate. Thus, large values of sensitivity and specificity imply small false negative and false positive rates.

Sensitivity and specificity are probabilities of the test result conditional on the disease status. These are values that the developer of the test has estimated during extensive testing in hospitals and clinics. However, as a potential patient, we are more interested in the probability of disease status conditional on the test result. Names given to two conditional probabilities that address the patient's concerns are predicted value positive and predicted value negative. *Predicted value positive* is the probability of disease given a positive test result, and *predicted value negative* is the probability of no disease given a negative test result.

These four quantities can be expressed succinctly in symbols. Let T^+ represent a positive test result and T^- represent a negative result. The presence of disease is indicated by D^+ and its absence is indicated by D^-. Then these four quantities can be expressed as conditional probabilities as follows:

Sensitivity	$\Pr\{T^+ \mid D^+\}$
Specificity	$\Pr\{T^- \mid D^-\}$
Predicted value positive	$\Pr\{D^+ \mid T^+\}$
Predicted value negative	$\Pr\{D^- \mid T^-\}$

All four of these probabilities should be large for a screening test to be useful to the screener and to the screenee. Discussions of these and related issues are plentiful in the epidemiologic literature (Weiss 1986).

It is possible to estimate these probabilities. One way is to select a large sample of the population and subject the sample to a screening or diagnostic test as well as to a standard clinical evaluation. The standard clinical evaluation is assumed to provide the true disease status. Then the sample persons can be classified into one of the four cells in the 2 by 2 table in Table 4.5. For example, hypertension status is first screened by the sphygmomanometer in the community and by a comprehensive clinical evaluation in the clinic. Or persons are screened for mental disorders first by the DIS (Diagnostic Interview Schedule) and then by a comprehensive psychiatric evaluation. The results from a two-stage diagnostic procedure would look like Table 4.5.

Table 4.5 Disease status by test results for a large sample from the population.

Disease	Test Result		
Status	Positive	Negative	Total
Presence	a	b	$a + b$
Absence	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Sensitivity is estimated by $a/(a + b)$, specificity is estimated by $d/(c + d)$, predicted value positive is estimated by $a/(a + c)$, and predicted value negative is estimated by $d/(b + d)$. Similarly, the false positive rate is estimated by $c/(c + d)$ and the false negative rate by $b/(a + b)$.

For many diseases of interest, the prevalence is so low that there would be few persons with the disease in the sample. This means that the estimates of sensitivity and the predicted value positive would be problematic. Therefore, some alternate sample design must be used to estimate these conditional probabilities. When a large number of people are screened by a test in a community and a sample of persons with positive test results and those with negative test results are subjected to clinical evaluations, the predicted value positive and the predicted value negative can be directly calculated from the results of clinical evaluations, and sensitivity and specificity can be indirectly estimated. Conversely, when sensitivity and specificity are directly estimated by applying the test to persons with the disease and persons without the disease in the clinic setting, the predicted value positive and the predicted value negative can be indirectly estimated if the prevalence rate of disease is known. These indirect estimation procedures are explained in the next section.

4.3.3 Receiver Operating Characteristic Plot

In evaluating a diagnostic test for a certain disease, we need to consider relative importance of sensitivity and specificity. For incidence, if the disease in question is likely to lead to death and the preferred treatment has few side effects, then it will be more important to make sensitivity as large as possible. On the other hand, if the disease is not too serious and the known treatment has considerable side effects, then more weight might be given to specificity. The cost of the treatment given to those with positive test results could also come into consideration. In many situations, we need to consider both sensitivity and specificity. But sensitivity and specificity are relative to how we define the status of disease. Different cut-off points in the definition of the condition would give different results.

Here we illustrate how the sensitivity and specificity of a test change with respect to the cut-off point chosen for indicating a positive test result. Let us consider the case of using the serum calcium level as a test for detect hyperparathyroidism (Lundgren et al. 1997). The following data show the level of serum calcium and the status of hyperparathyroidism:

		Serum Calcium Levels mg/dL					
		8 mg/dL	**9 mg/dL**	**10 mg/dL**	**11 mg/dL**	**12 mg/dL**	**Total**
Disease Status	Negative	40	7	4	2	1	54
	Positive	2	3	5	8	17	35
	Total	42	10	9	10	18	89

If we consider 9 mg/dL or more as positive test result, the data can be summarized as follows:

		Serum Calcium Levels mg/dL		
		9 mg/dL or more	**Less than 9 mg/dL**	**Total**
Disease Status	Negative	14	**40**	54
	Positive	**33**	2	35
	Total	47	42	89

From this summary, the estimated sensitivity is 0.94 (= 33/35) and the specificity is 0.74 (= 40/54). As the cut-point changes, the sensitivity and specificity of the diagnostic test also change. As we increase the cut-point for serum calcium levels, the sensitivity of the test decreases and the specificity increases as shown here.

Cut-Point	**Sensitivity**	**Specificity**
<8 mg/dL | 8 mg/dL	35 / 35 = 1.00	0 / 54 = 0.00
8 mg/dL | 9 mg/dL	33 / 35 = 0.94	40 / 54 = 0.74
9 mg/dL | 10 mg/dL	30 / 35 = 0.86	47 / 54 = 0.87
10 mg/dL | 11 mg/dL	25 / 35 = 0.71	51 / 54 = 0.94
11 mg/dL | 12 mg/dL	17 / 35 = 0.48	53 / 54 = 0.98
12 mg/dL | >12 mg/dL	0 / 35 = 0.00	54 / 54 = 1.00

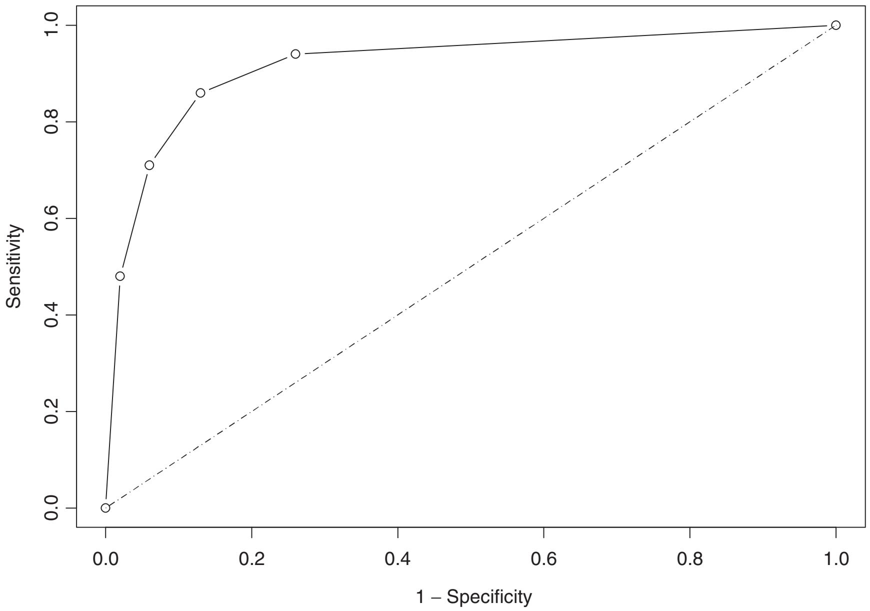

Figure 4.1 The receiver operating characteristic plot for serum calcium values and hyperparathyroidism.

We generally use the Receiver Operating Characteristic (ROC) plot to examine the tradeoff between sensitivity and specificity. This is a plot of sensitivity versus 1 – specificity. Figure 4.1 shows the ROC plot for the preceding data. By looking at the curve relative to a 45-degree line, we notice that as the curve extends farther away from the line, the accuracy of the diagnostic test improves, and as the curve draws nearer to the 45-degree line, the diagnostic test's accuracy becomes worse. Therefore, we can consider the area under the ROC curve as a measure of a diagnostic test's discrimination or the test's ability to correctly classify individuals with and without the disease. An excellent test would have an area under the curve of nearly 1.00, while a poor test would have an area under the curve of nearly 0.50.

4.4 Bayes' Theorem

We wish to find the predicted value positive and predicted value negative using the known values for disease prevalence, sensitivity, and specificity. Let us focus on predicted value positive — that is, $\Pr\{D^+ \mid T^+\}$ — and see how it can be expressed in terms of sensitivity, $\Pr\{T^+ \mid D^+\}$, specificity, $\Pr\{T^- \mid D^-\}$, and disease prevalence, $\Pr\{D^+\}$.

We begin with the definition of the predicted value positive:

$$\Pr\{D^+|T^+\} = \frac{\Pr\{D^+ \cap T^+\}}{\Pr\{T^+\}}.$$

Applying Equation (4.6), $\Pr\{A \cap B\} = \Pr\{B \mid A\} \cdot \Pr\{A\}$, the probability of the intersection of D^+ and T^+ can also be expressed as

$$\Pr\{D^+ \cap T^+\} = \Pr\{T^+ \mid D^+\}\ \Pr\{D^+\}.$$

On substitution of this expression for the probability of the intersection in the definition of the predicted value positive, we have

$$\Pr\{D^+|T^+\} = \frac{\Pr\{T^+|D^+\}\cdot\Pr\{D^+\}}{\Pr\{T^+\}} \tag{4.7}$$

which shows that predicted value positive can be obtained by dividing the product of sensitivity and prevalence by $\Pr\{T^+\}$.

Applying Equation (4.1), $\Pr\{T^+\}$ can be expressed as the sum of the probabilities of the intersection of T^+ with the two possible outcomes of the disease status: D^+ and D^-; that is,

$$\Pr\{T^+\} = \Pr\{T^+ \cap D^+\} + \Pr\{T^+ \cap D^-\}.$$

Applying Equation (4.5) to the two probabilities of intersections, we now have

$$\Pr\{T^+\} = \Pr\{T^+ \mid D^+\}\Pr\{D^+\} + \Pr\{T^+ \mid D^-\}\Pr\{D^-\}.$$

Substituting this expression in Equation (4.7), the predictive value positive is

$$\Pr\{D^+|T^+\} = \frac{\Pr\{T^+|D^+\}\cdot\Pr\{D^+\}}{\Pr\{T^+|D^+\}\cdot\Pr\{D^+\}+\Pr\{T^+|D^-\}\cdot\Pr\{D^-\}}. \tag{4.8}$$

Note that the numerator and the first component of the denominator is the product of sensitivity and disease prevalence. The second component of the denominator is the product of (1 – specificity) and (1 – disease prevalence). Predicted value negative follows immediately, and it is

$$\Pr\{D^-|T^-\} = \frac{\Pr\{T^-|D^-\}\cdot\Pr\{D^-\}}{\Pr\{T^-|D^-\}\cdot\Pr\{D^-\}+\Pr\{T^-|D^+\}\cdot\Pr\{D^+\}}.$$

These two formulas are special cases of the theorem discovered by Reverend Thomas Bayes (1702–1761). In terms of the events A and B_i, Bayes' theorem is

$$\Pr\{B_i|A\} = \frac{\Pr\{A|B_i\}\cdot\Pr\{B_i\}}{\sum_i \Pr\{A|B_i\}\cdot\Pr\{B_i\}}.$$

Example 4.6

Consider the use of the count of blood vessels in breast tumors. A high density of blood vessels indicates a patient who is at high risk of having cancer spread to other organs (Weidner et al. 1992). Use of the count of blood vessels appears to be worthwhile in women with very small tumors and no lymph node involvement — the node-negative case. Suppose that during the development stage of this procedure, its sensitivity was estimated to be 0.85; that is, of the women who had cancer spread to other organs, 85 percent of them had a high count of blood vessels in their breast tumors. The specificity of the test was estimated to be 0.90; that is, of the women for whom there was no spread of cancer, 90 percent of them had a low count of blood vessels in their tumors. Assume that the prevalence of cancers spread from breast cancers is 0.02. Given these assumed values, what is the predicted value positive (PVP) of counting the number of blood vessels in the small tumors?

Using Equation (4.8),

$$\Pr\{D^+|T^+\} = \frac{\Pr\{T^+|D^+\} \cdot \Pr\{D^+\}}{\Pr\{T^+|D^+\} \cdot \Pr\{D^+\} + \Pr\{T^+|D^-\} \cdot \Pr\{D^-\}},$$

the answer is

$$PVP = \frac{\mathit{prevalence} \cdot \mathit{sensitivity}}{[\mathit{prevalence} \cdot \mathit{sensitivity}] + [(1 - \mathit{prevalence}) \cdot (1 - \mathit{specificity})]}$$

$$= \frac{(0.02) \cdot (0.85)}{(0.02) \cdot (0.85) + (1 - 0.02) \cdot (1 - 0.90)} = \frac{0.017}{0.115} = 0.148.$$

Using the preceding assumed values for sensitivity, specificity, and prevalence, there is approximately a 15 percent chance of having cancer spread from a small breast tumor given a high density of blood vessels in the tumor. This value may be too low for the test to be useful. If the true values for specificity or prevalence are higher than the values just assumed, then the PVP will also be higher. For example, if the prevalence is 0.04 instead of 0.02, then the PVP is 0.262 instead of 0.148.

Example 4.7

Let us recast the question in Example 4.6 using frequencies instead of probabilities. Suppose that 20 out of every 1000 women with breast tumors have cancer spread to other organs (the prevalence of cancer spread is 0.02). Of these 20 women with cancer spread, 17 will have a high count of blood vessels (sensitivity of the test was estimated to be 0.85). Of the remaining 980 women without cancer spread, 882 will have a low count of blood vessels in their tumors (specificity of the test was estimated to be 0.90). Then what percent of women with a high density of blood vessels do actually have cancer spread (predicted value positive)?

This question can be answered easily without using the Bayes' formula. Looking at the frequencies just stated, the total number of women with high-density blood vessels is the sum of 17 from those with cancer spread and 98 (980 minus 882) from those without cancer spread. The sum is 115. Of these, 17 saw their cancers spread. Therefore, the predicted value positive is 0.148 (= 17/115), which is the same value obtained by the formula in Example 4.5. You can see it more clearly in the following 2 by 2 table:

	Blood Vessel Count		
Cancer Spread	**High**	**Low**	**Total**
Yes	(17)**	3	(20)*
No	98	(882)***	980
Total	115	885	1000

*Prevalence rate of 0.02 Predicted value positive: 17/115 = 0.148
**Sensitivity of 0.85 Predicted value negative: 882/885 = 0.997
***Specificity of 0.90

This example demonstrates that Bayes' theorem is to enhance and expedite our reasoning rather than to be memorized blindly.

4.5 Probability in Sampling

Sampling means selecting a few units from all the possible observational units in the population. To infer from the sample to the population, we need to know the probability of selection. A sample selected with unknown probability of selection cannot be linked appropriately to the population from which the sample was drawn. A sample drawn with known probability of selection is called a probability sample. We examine the simplest probability sample that assigns an equal probability of selection to every unit of observation in the population. More complex sample selection designs will be discussed in Chapter 6.

4.5.1 Sampling with Replacement

A sample that allows duplicate selections is called a sample with replacement. Allowance of duplicate selection implies that sample selections are independent — each selection is not dependent on previous selections. To understand the probability of selection in a sample with replacement, let us consider the case of selecting three units from a population of four units (*A*, *B*, *C*, and *D*). There are 64 (= 4^3) ways of selecting such samples as listed in Table 4.6.

Table 4.6 Possible samples of drawing 3 from (A, B, C and D) with replacement.

AAA	ACA	BAA	BCA*	CAA	CCA	DAA	DCA*
AAB	ACB*	BAB	BCB	CAB*	CCB	DAB*	DCB*
AAC	ACC	BAC*	BCC	CAC	CCC	DAC*	DCC
AAD	ACD*	BAC*	BCD*	CAD*	CCD	DAD	DCD
ABA	ADA	BBA	BDA*	CBA*	CDA*	DBA*	DDA
ABB	ADB*	BBB	BDB	CBB	CDB*	DBB	DDB
ABC*	ADC*	BBC	BDC*	CBC	CDC	DBC*	DDC
ABD*	ADD	BBD	BDD	CBD*	CDD	DBD	DDD

*Samples without duplications

Since selections are independent, the probability of selecting each of these samples is 1/64 (= $[1/4]^3$). As shown in the table, the total number of samples without duplications is 24(4 × 3 × 2); that is, there are 4 ways to fill the first position of the sample, 3 ways to the second position, and 2 ways to fill the third position. The probability of selection with replacement samples that do not contain duplication is 0.375 (= 24/64). The probability of obtaining samples with duplications is 0.625 (= 1 – 0.375).

When selecting n units from N units in the population, there are N^n possible samples with replacement. Of these, $N(N-1)\ldots(N-n+1)$ samples contain no duplications. Then the probability of obtaining with replacement samples that contain duplications is

$$\Pr(duplications) = 1 - \frac{N(N-1)\cdots(N-n+1)}{N^n}. \tag{4.9}$$

Example 4.8

How likely is it that at least two students in a class of 23 will share the same birthday? The chance may be better than we might expect. If we assume that the birthdays of 23 students are independent and that each day out of 365 days in a year, eliminating February 29, is equally likely to be a student's birthday, the situation is equivalent to selection of a random sample of 23 days from the 365 days using the sampling with replacement procedure. The probability can be calculated using Equation (4.9) — that is,

$$1-\frac{365(364)\cdots(343)}{365^{23}}=0.507.$$

The calculation may require the use of a computer (the SAS program is available on the website). If the size of class increases to 50, the probability increases to 0.97.

Let us consider the probability of selecting a particular unit. In the list of samples in Table 4.6, unit A appears 16 times in the first position of the sample, 16 times in the second position, and 16 times in the third position. Then the probability of A being selected into the sample is $[3(16) / 4^3] = (48 / 64) = (3 / 4)$. Thus, in general, the selection probability of a unit is n / N.

4.5.2 Sampling without Replacement

A sample that does not allow duplications is called a sample without replacement. In sampling without replacement, a selection of a unit is no longer independent because the selection is conditional on the unit being not selected in a previous draw. In this sampling, once a subject is selected, it is removed from the population, and the number of units in the population is decreased by one unit. Does this decrease in the denominator as a unit is selected invalidate the equal probability of selection for subsequent units? The following example addresses this.

Suppose that a class has 30 students, and a random sample of 5 students is to be selected without allowing duplicate selections. The probability of selection for the first draw will be 1 / 30, and that for the student selected second will be 1 / 29, since one student was already selected. This line of thinking seems to suggest that random sampling without replacement is not an equal probability sampling model. Is anything wrong in our thinking?

We have to realize that the selection probability of 1 / 29 for the second draw is a conditional probability. The student selected in the second draw is available for selection only if the student were not selected in the first draw. The probability of not being selected in the first draw is 29 / 30. Thus, the event of being selected during the second draw is the intersection of the events of not being selected during the first draw and being selected during the second draw. Applying $\{\Pr\{A \cap B\} = \Pr\{A \mid B\} \cdot \Pr\{B\}$, the probability of this intersection is (1 / 29) (29 / 30), which yields 1 / 30. The same argument can be made for subsequent draws, as shown in Table 4.7.

Table 4.7 Calculation of inclusion probabilities in drawing an SRS of 5 from 30 without replacement.

Order of Draw	Conditional Probability (1)	Probability Not Selected in Previous Draws (2)	Product of (1) & (2)
1	1/30	1	1/30
2	1/29	29/30	1/30
3	1/28	(29/30)(28/29) = 28/30	1/30
4	1/27	(29/30)(28/29)(27/28) = 27/30	1/30
5	1/26	(29/30)(28/29)(27/28)(26/27) = 26/30	1/30

The demonstration in Table 4.7 indicates that the probability of being selected in any draw is 1 / 30, and hence the equal probability of selection also holds for sampling without replacement. Now we can state that the probability for a particular student to be included in the sample will be 5 / 30, since the student can be drawn in any one of the five draws. Thus, in general, the selection probability of a unit without replacement is n / N, the same as in the case of replacement sampling.

A sampling procedure that assigns n / N chance of being selected into the sample to every unit in the population is called *simple random sampling*, regardless of whether sampling is done with or without replacement. We usually use sampling without replacement. The distinction between sampling with and without replacement is moot when selecting a sample from large populations because the chance of selecting a unit more than once would be very small. The statement that each of the possible samples is equally likely implies that each unit in the population has the same probability of being included in the sample as demonstrated in this and the previous section.

4.6 Estimating Probabilities by Simulation

Our approach to finding probabilities has been to enumerate all possible outcomes and to base the calculation of probabilities on this enumeration. This approach works well with simple phenomena, but it is difficult to use with complex events. Another way of assessing probabilities is to simulate the random phenomenon by using repeated sampling. With the wide availability of microcomputers, the simulation approach has become a powerful tool to approach many statistical problems.

Example 4.9

Let us reconsider the question posed in Example 4.8. In a class of 30, what will be the chance of finding at least 2 students sharing the same birthday? It should be higher than the 50 percent that we found among 23 students in Example 4.8. Let us find an answer by simulation. We need to make the same assumptions as in Example 4.8. Selecting 30 days from 365 days using the sampling procedure, we can use the random number table in Appendix B. For example, we can read 30 three-digit numbers between 1 and 365 from the table and check to see if any duplicate numbers are selected. We can repeat the operation many times and see how many of the trials produced duplicates. Since this manual simulation would require considerable time, we can use a computer program (see **Program Note 4.1** on the website). The results of 10 simulations are shown in Table 4.8.

Table 4.8 Simulation to find the probability of common birthdays among 30 students.

	Simulations									
Student	**1**	**2***	**3***	**4***	**5***	**6***	**7**	**8***	**9***	**10***
1	4	2	3	44	8	3	7	5	8	12
2	10	30	10*	52	21	4	47	7	18	19
3	21	46	10*	72	24	22	48	7	27	31
4	47	67	15	85	76	23	54	18	45	48
5	48	97	23	106	91	27	80	23	50	65
6	64	100	26	116	100	42	82	37	66	80
7	65	105	35	120	113	57	93	54	90	82
8	78	106	41	123	124	64	119	59	91	103
9	93	106	53	132	143*	72	123	64	94	116
10	95	109	73	143	143*	104	137	89	97	169
11	101	133	78	151	147	107	138	109	104	175
12	115	140	86	180	150	119	140	120	132	182
13	154	145	87	181	155	132	162	138	149	193
14	165	158	163	188	166	152	179	143	153	195
15	167	191	166	208	172	167	185	173	180	208
16	185	209*	176	231	200	210	191	201	187	217
17	193	209*	186	248	205	229	199	209*	188	247
18	220	220	200	249	241	230	203	209*	189	249
19	232	223	209	255	243	233	213	215	193	261
20	242	229	220	259*	248	236	232	223	196	262*
21	257	241	251	259*	250	253	238	224	242	262*
22	282	249	260	267	263	307	252	231	250	305
23	284	268	264	270	281	321	259	239	324	307
24	285	286	265	285	283	326	267	259	333	309
25	288	317	283	286	307	327	272	274	338	321
26	299	323	295	288	310	334	287	335	354	326
27	309	335*	297	296	311	336	295	342	360*	328
28	346	335*	300	310	326	343*	308	352	360*	330
29	347	336	352	327	335	343*	313	357	360*	347
30	357	356	355	352	336	362	363	358	360*	356

Eight of these 10 trials have duplicates, which suggests that there is an 80 percent probability of finding at least one common birthday among 30 students. Not shown are the results of 10 additional trials in which 6 of the 10 had duplicates. Combining these two sets of 10 trials, the probability of finding common birthdays among 30 students is estimated to be 70 percent (= [8 + 6] / 20). Using $\Pr(duplications) = 1 - \frac{N(N-1)\cdots(N-n+1)}{N^n}$ we get 70.6 percent. Using 20 replicates is usually not enough to have a lot of confidence in the estimate; we usually would like to have at least hundreds of replicates.

Let us consider another example.

Example 4.10

Population and family planning program planners in Asian countries have been dealing with the effects of the preference for a son on population growth. If all couples continue to have children until they have two sons, what is the average number of children they would have? To build a probability model for this situation, we assume that genders of successive children are independent and the chance of a son is 1 / 2. To simulate the number of children a couple has, we select single digits from the random number table, considering odd numbers as boys and even numbers as girls. Random numbers are read until the second odd number is encountered, and the number of values required to obtain two odd values is noted. Table 4.9 shows the results for 20 trials (couples).

The average number of children based on this very small simulation is estimated to be 4.25 (= 85 / 20). Additional trials would provide an estimate closer to the true value of four children.

Table 4.9 Simulation of child-bearing until the second son is born.

Trial	Digits	No. of Digits	Trial	Digits	No. of Digits
1	19	2	11	37	2
2	2239	4	12	367	3
3	503	3	13	6471	4
4	4057	4	14	509	3
5	56287	5	15	940001	6
6	13	2	16	927	3
7	96409	5	17	277	3
8	125	3	18	544264882425	12
9	31	2	19	3629	4
10	425448285	9	20	045467	6
			Total number of digits		85
			Average = 85/20 = 4.25		

4.7 Probability and the Life Table

Perhaps the oldest probability model that has been applied to a problem related to health is the life table. The basic idea was conceived by John Graunt (1620–1674), and the first life table, published in 1693, was constructed by Edmund Halley (1656–1742). Later Daniel Bernoulli (1700–1782) extended the model to determine how many years would be added to the average life span if smallpox were eliminated as a cause of death. Now the life table is used in a variety of fields — for example, in life insurance calculations, in clinical research, and in the analysis of processes involving attrition, aging and wearing out of industrial products.

We are presenting the life table here to show an additional application of the probability rules described above. Table 4.10 is the abridged life table for the total U.S. population in 2002. It is based on information from all death certificates filed in the 50 states and the District of Columbia. It is called an abridged life table because it uses age-groupings instead of single years of age. If single years of age are used, it is called a complete life table. Prior to 1997, a complete life table was construed only for a census year and for all off-census years abridged life tables were constructed. Beginning with 1997 mortality data, a complete life table was constructed every year, and abridged tables are derived from the complete tables. Previously, the annual life tables were closed at age 85, but they have been extended to age 100 based on old-age mortality data from the Medicare program. Other types of life tables are available from the National Center for Health Statistics. A brief history and sources for life tables for the United States can be found in Appendix C.

Table 4.10 Abridged life table for the total U.S. population, 2002.

	Probability of Dying Between Ages *x* and *x* + *n*	Number Surviving to Age *x*	Number Dying Between Ages *x* and *x* + *n*	Person-Years Lived Between Ages *x* and *x* + *n*	Total Number of Person-Years Lived Above Age *x*	Expectation of Life at Age *x*
Age	$_nq_x$	l_x	nd_x	$_nL_x$	T_x	e_x
0–1	0.006971	100,000	697	99,389	7,725,787	77.3
1–5	0.001238	99,303	123	396,921	7,626,399	76.8
5–10	0.000759	99,180	75	495,706	7,229,477	72.9
10–15	0.000980	99,105	97	495,311	6,733,771	67.9
15–20	0.003386	99,008	335	494,345	6,238,460	63.0
20–25	0.004747	98,672	468	492,189	5,744,116	58.2
25–30	0.004722	98,204	464	489,871	5,251,927	53.5
30–35	0.005572	97,740	545	487,395	4,762,056	48.7
35–40	0.007996	97,196	777	484,164	4,274,661	44.0
40–45	0.012066	96,419	1,163	479,362	3,790,497	39.3
45–50	0.017765	95,255	1,692	472,292	3,311,135	34.8
50–55	0.025380	93,563	2,375	462,186	2,838,843	30.3
55–60	0.038135	91,188	3,478	447,838	2,376,658	26.1
60–65	0.058187	87,711	5,104	426,603	1,928,820	22.0
65–70	0.088029	82,607	7,272	395,866	1,502,217	18.2
70–75	0.133076	75,335	10,025	352,791	1,106,350	14.7
75–80	0.201067	65,310	13,132	294,954	753,560	11.5
80–85	0.304230	52,178	15,874	222,013	458,606	8.8
85–90	0.447667	36,304	16,252	140,041	236,593	6.5
90–95	0.599618	20,052	12,024	67,822	96,552	4.8
95–100 ...	0.739020	8028	5933	23,056	28,730	3.6
100+	1.000000	2095	2095	5675	5675	2.7

Source: Arias, 2004

One use of the life table is to summarize the life experience of the population. A direct way of creating a life table is to follow a large cohort — say, 100,000 infants born on the same day — until the last member of this cohort dies. For each person the exact length of life can be obtained by counting the number of days elapsed from the date of birth. This yields 100,000 observations of the length of life. The random variable is the length of life in years or even in days. We can display the distribution of this random variable and calculate the mean, the median, the first and third quartiles, and the minimum and maximum. Since most people die at older ages, we expect that the distribution is skewed to the left, and hence the median length of life is larger than the mean length of life. The mean length of life is the life expectancy. We can tabulate the data using the age intervals 0–1, 1–5, 5–10, 10–15, . . . , 95–100, and 100 or over. All the intervals are the same length — five years — except for the first two and the last interval. The first interval is of a special interest, since quite a few infants die within it. From this tabulation, we can also calculate the relative frequency distribution by dividing the frequencies by 100,000. These relative frequencies give the probability of dying in each age interval. This probability distribution can be used to answer many practical questions regarding life expectancy. For instance, what is a 20-year-old person's probability of surviving to the retirement age of 65?

However, acquiring such data poses a problem. It would take more than 100 years to collect it. Moreover, information obtained from such data may be of some historical interest but not useful in answering current life expectancy questions, since current life expectancy may be different from that of earlier times. To solve this problem, we have to find ways to use current mortality information to construct a life table. The logical current mortality data for this purpose are the age-specific death rates. For the time being, we assume that age-specific death rates measure the probability of dying in each age interval. Note that these rates are conditional probabilities. The death rate for the 5- to 10-year-old age group is computed on the condition that its members survived the previous age intervals.

As presented in Chapter 3, the age-specific death rate is calculated by dividing the number of deaths in a particular age group by the midyear population in that age group. This is not exactly a proportion, whereas a probability is. Therefore, the first step in constructing a life table is to convert the age-specific death rates to the form of a probability. One possible conversion is based on the assumption that the deaths were occurring evenly throughout the interval. Under this assumption, we expect that one-half of the deaths occurred during the first half of the interval. Thus, the number of persons at the beginning of an interval is the sum of the midyear population and one-half of the deaths that occurred during the interval. Then the conditional probability of dying during the interval is the number of deaths divided by the number of persons at the beginning of the interval. Actual conversions use more complicated procedures for different age groups, but we are not concerned about these details.

4.7.1 The First Four Columns in the Life Table

With this background, we are now ready to examine Table 4.10. The first column shows the age intervals between two exact ages. For instance, 5–10 indicates the five-year interval between the fifth and tenth birthdays. This age grouping is slightly different from those of under 5, 5–9, 10–14, and so on used in the Census publications. In the life table, age is considered a continuous variable, whereas in the Census, counting of people by age (ignoring the fractional year) is emphasized.

The second column shows the proportion of the persons alive at the beginning of the interval who will die before reaching the end of the interval. It is labeled as ${}_nq_x$, where the first subscript on the left denotes the length of the interval and the second subscript on the right denotes the exact age at the beginning of the interval. The first entry in the second column, ${}_1q_0$, is 0.006971, which is the probability of newborn infants dying during the first year of life. The second entry is ${}_4q_1$, which equals 0.001238. It is the conditional probability of dying during the interval between ages 1 and 5, provided the child survived the first year of life. The rest of the entries in this column are conditional probabilities of dying in a given interval for those who survived the preceding intervals. These conditional probabilities are estimated from the current age-specific death rates. Note that the last entry of column 2 is 1.000000, indicating everybody dies sometime after age 100.

Thus, we have a series of conditional probabilities of dying. Given these conditional probabilities of dying, we can also find the conditional probabilities of surviving. The probability of surviving the first year of life will be

$$(1 - {}_1q_0) = 1 - 0.006971 = 0.993029.$$

Likewise, the conditional probability of surviving the interval between exact ages 1 and 5, provided infants had survived the first year of life will be

$$(1 - {}_4q_1) = 1 - 0.001238 = 0.998762.$$

Surviving the first five years of life is the intersection of surviving the 0–1 interval and the 1–5 interval. The probability of this intersection can be obtained as the product of the probability of surviving the 0–1 interval and the conditional probability of surviving the 1–5 interval given survival during the 0–1 interval — that is,

$$\Pr\{\text{surviving the intervals 0–1 and 1–5}\} = (1 - {}_1q_0)\ (1 - {}_4q_1)$$
$$= (1 - 0.006971)\ (1 - 0.001238) = (0.993029)\ (0.998762) = 0.991800.$$

Similarly, the probability of surviving the first 10 years of life, the first three intervals, will be

$$(1 - {}_1q_0)\ (1 - {}_4\mathrm{q}_1)\ (1 - {}_5q_5).$$

Using this approach, we can calculate the survival probabilities from birth to the beginning of any subsequent age intervals. These survival probabilities are reflected in the third column, the number alive, l_x, at the beginning of the interval which begins at x years of age, out of a cohort of 100,000. Note that the entries in this column may differ slightly from the product of the survival probabilities and 100,000 because, although only four digits to the right of the decimal point are shown in the second column, more digits are used in the calculations. The first entry in this column is l_0, called the *radix*, is the size of the birth cohort. The second entry, the number alive at the beginning of the interval beginning at 1 year of age, l_1, is found by taking the product of the number alive at the beginning of the previous interval and the probability of surviving that interval — that is,

$$l_1 = l_0\ (1 - {}_1q_0) = l_0 - (l_0 - {}_1q_0) = l_0 - {}_1d_0.$$

This quantity, l_1, is equivalent to taking the number alive at the beginning of the previous period minus the number that died during that period, ${}_1d_0$. The numbers that died during each interval are shown in the fourth column, which is labeled as ${}_nd_x$.

The number who died during the four-year age interval from 1 to 5 is ${}_4d_1$. This is found by taking the product of the number alive at the beginning of this interval, l_1, and the probability of dying during the interval, ${}_4q_1$ — that is, ${}_4d_1 = l_1({}_4q_1)$. The number alive at the beginning of the interval of 5 to 10 years of age, l_5, can be found by subtracting the number who died during the previous age interval, ${}_4d_1$, from the number alive at the beginning of the previous interval, l_1 — that is, $l_5 = l_1 - {}_4d_1$. Repeating this operation, the rest of the entries in the third and fourth columns can be obtained. The fourth column can also be obtained directly from the third column. For example,

$${}_1d_0 = l_0 - l_1,\ {}_4d_1 = l_1 - l_5, \text{ etc.}$$

Note that the last entry of the third column is the same as the last entry in the fourth column because all the survivors at age 100 will die subsequently. Note further that the l_x value in each row is a cumulative total of ${}_nd_x$ values in that and all subsequent rows.

Dividing the entries in the third and fourth columns by 100,000, we obtain the probabilities of surviving from birth to the beginning of the current interval and dying during the current interval, respectively. Note that the entries in the fourth column sum to 100,000, meaning that the probability of dying sums to one. As we expected, the distribution is negatively skewed, with the larger probabilities of dying at older ages.

4.7.2 Some Uses of the Life Table

Before looking at expected values in the life table, we wish to show how the first four columns, particularly the third column, can be used to answer some questions regarding life chances.

Example 4.11

What is the probability of surviving from one age to a subsequent age, say from age 5 to age 20? This is a conditional probability, conditional on the survival to age 5. The intersection of the events of surviving to age 20 and surviving to age 5 is surviving to age 20. Thus, the probability of this intersection is the probability of surviving from birth to age 20. This is the number alive at the beginning of the 20–25 interval divided by the number alive at the beginning — that is, l_{20}/l_0. The probability of surviving from birth to age 5 is l_5/l_0. Therefore, the conditional survival probability from age 5 to age 20 is found by dividing the probability of the intersection by the probability of surviving to age 5 — that is,

$$\left(\frac{l_{20}}{l_0}\right)\Big/\left(\frac{l_5}{l_0}\right)=\left(\frac{l_{20}}{l_5}\right)=\frac{98672}{99180}=0.994878.$$

The survival probabilities from any age to an older age can be calculated in a similar fashion.

We know the conditional probability of dying in any single interval. However, we may be interested in the probability of dying during a period formed by the first two or more consecutive intervals.

Example 4.12

What is the probability of dying during the first 5 years of life? This probability can be found by subtracting the probability of surviving the first 5 years from 1 — that is,

$$\begin{aligned}1-(1-{}_1q_0)(1-{}_4q_1)&=1-\left(\frac{l_1}{l_0}\right)\left(\frac{l_5}{l_1}\right)=1-\frac{l_5}{l_0}.\\&=1-0.99180=0.0820\end{aligned}$$

This is simply 1 minus the ratio of the number alive at the beginning of the final interval of interest and 100,000.

Example 4.13

A similar question relates to the probability of dying during a period formed by two or more consecutive intervals given that one had already survived several intervals. For example, what is the probability that a 30-year-old person will die between the ages of 50 and 60? This conditional probability is found by dividing the probability of the intersection of the event of dying between the ages of 50 and 60 and the event of surviving until 30 by the probability of the event of surviving until 30 years of age. The intersection of dying between 50 and 60 and surviving until 30 is dying between 50 and 60. The probability of dying between 50 and 60 is the number of persons dying, l_{50} minus l_{60}, divided by the total number, l_0. The probability of surviving until age 30 is simply l_{30} divided by l_0. Therefore, the probability of dying between 50 and 60 given survival until 30 is

$$\left(\frac{l_{50}-l_{60}}{l_0}\right)\bigg/\left(\frac{l_{30}}{l_0}\right)=\frac{l_{50}-l_{50}}{l_{30}}=\frac{93563-87711}{97740}=0.059873.$$

Example 4.14

Another slightly more complicated question concerns the joint survival of persons. Suppose that a 40-year-old person has a 5-year-old child. What will be the probability that both the parent and child survive 25 more years until the parent's retirement? If we assume that the survival of the parent and that of the child are independent, we can calculate the desired probability by multiplying the individual survival probabilities. Applying the rule for the probability of surviving from one age to a subsequent age from the first question, this is

$$\left(\frac{l_{65}}{l_{40}}\right)\cdot\left(\frac{l_{30}}{l_5}\right)=\frac{82607}{96419}\cdot\frac{97740}{99180}=0.856750\cdot 0.985481=0.844311.$$

The probability that the parent will die but the child will survive during the 25 years is

$$\left(1-\frac{l_{65}}{l_{40}}\right)\cdot\left(\frac{l_{30}}{l_5}\right)=(1-0.856750)\cdot 0.985481=0.141170.$$

The probability that the parent will survive but the child will die during the 25 years is

$$\left(\frac{l_{65}}{l_{40}}\right)\cdot\left(1-\frac{l_{30}}{l_5}\right)=0.856750\cdot(1-0.985481)=0.012439.$$

The probability that both the parent and the child will die during the 25 years is

$$\left(1-\frac{l_{65}}{l_{40}}\right)\cdot\left(1-\frac{l_{30}}{l_5}\right)=(1-0.856750)\cdot(1-0.985481)=0.002080.$$

These four probabilities sum to 1 because those four events represent all the possible outcomes in considering the life and death of two persons.

4.7.3 Expected Values in the Life Table

The last three columns contain the information for various expected values in the life table. The fifth column of the life table, denoted by ${}_nL_x$, shows the person-years lived during each interval. For instance, the first entry in the fifth column is 99,389, which is the total number of person-years of life contributed by 100,000 infants during the first year of life. This value consists of 99,303 years contributed by the infants that survived the full year. The remaining 86 (= 99389 – 99303) person-years came from 697 infants who died during the year. The value of 86 years is based on actual mortality data coupled with mathematical smoothing. It cannot be found from the first four columns in the table. The value of 86 years is much less than 348.5 expected if the deaths had been distributed uniformly during the year. This value also suggests that most of the deaths occurred during the first half of the interval. The second entry of the fifth column is much larger than the first entry, mainly reflecting that the length of the second interval is greater than the length of the first interval. Each person surviving this second interval contributed 4 person years of life.

The fifth column is often labeled as the "stationary population in the age interval." The label of stationary population is based on a model of the long-term process of birth and death. If we assume 100,000 infants are born every year for 100 years, with each birth cohort subject to the same probabilities of dying specified in the second column of the life table, then we expect that there will be 100,000 people dying at the indicated ages every year. This means that the number of people in each age group will be the numbers shown in the fifth column. This hypothetical population will maintain the same size, since the number of births is the same as the number of deaths and it also keeps the same age distribution. That is, the size and structure of population is invariant, and hence this is called a stationary population.

The sixth column of the life table, denoted by T_x, shows cumulative totals of ${}_nL_x$ values starting from the last age interval. The T_x value in each interval indicates the number of person years remaining in that and all subsequent age intervals. For example, the T_{95} value of 28,730 is the sum of ${}_5L_{95}$ (= 23056) and ${}_\infty L_{100}$ (= 5675).

The last column of the life table, denoted by e_x, shows the life expectancies at various ages, which are calculated by $e_x = T_x/l_x$. The first entry of the last column is the life expectancy for newborn infants, and all subsequent entries are conditional life expectancies. Conditional life expectancies are more useful information than the expectancies figured for newborn infants. For instance, those who survived to age 100 are expected to live 2.7 years more ($e_{100} = 2.7$), the last entry of the last column, whereas newborn infants are expected to live 0.06 years beyond age 100 ($T_{100} / l_0 = 5675/100000 = 0.06$).

Example 4.15

Based on T_x values, more complicated conditional life expectancies can be calculated. For instance, suppose that a 30-year-old person was killed in an industrial accident and had been expected to retire at age 65 if still alive. For how many years of unearned income should that person's heirs be compensated? The family may request a compensation for 35 years. However, based on the life table, the company argues for a smaller number of years. The total number of years of life remaining during the interval from 30 to 65 is T_{30} minus T_{65}, and there are l_{30} persons remaining at age 30 to live those years. Therefore, the average number of years of life remaining is found by

$$\frac{T_{30} - T_{65}}{l_{30}} = \frac{4762056 - 1502217}{97740} = 33.4.$$

Example 4.16

The notion of stationary population can be used to make certain inferences for population planning and manpower planning. The birth rate of the stationary population can be obtained by dividing 100,000 by the total years of life lived by the stationary population, or

$$\frac{l_0}{T_0} = \frac{100000}{7725787} = \frac{1}{77.6} = 0.013$$

or 13 per 1000 population. The death rate should be the same. But note that the birth rate equals the reciprocal of the life expectancy at birth ($1/e_0$). In other words, the birth rate (replacement rate) and death rate (attrition rate) are entirely determined by the life expectancy under the stationary population assumption.

4.7.4 Other Expected Values in the Life Table

The most widely used figures from the life table are life expectancies. These are average values. As discussed in Chapter 3, the mean value may not represent the distribution appropriately in some circumstances. Let us find the median length of life at birth. To find the median, the second quartile, we must find the value such that 50 percent of the radix falls below it. By examining column 3 in the life table, we find that 52,178 persons are alive at the beginning of the age interval 80–85, whereas only 36,304 are alive at the beginning of the interval 85–90. Since 50,000 is between 52,178 and 36,304, we know that the median is somewhere between 80 and 85 years of age. If we assume that the 15,874 deaths are uniformly distributed over this age interval, we can find the median by interpolation. We add a proportion of the five years, the length of the interval, to the age at the beginning of the interval, 80 years. The proportion is the ratio of the difference between 52,178 and 50,000 to the 15,304 deaths that occurred in the interval. The calculation is

$$median = 80 + 5 \cdot \left(\frac{52178 - 50000}{15874} \right) = 80.69.$$

As expected, the mean is smaller than the median. Perhaps, it is more enlightening to know that one-half of a birth cohort will live to age 81 than to know that an average length of life is about 77 years.

The corresponding calculations for the first and third quartiles are

$$Q_1 = 70 + 5 \cdot \left(\frac{75335 - 75000}{10025} \right) = 70.17$$

$$Q_3 = 85 + 5 \cdot \left(\frac{36304 - 25000}{16252} \right) = 88.48.$$

Conclusion

Probability has been defined as the relative frequency of an event in an infinite number of trials or in a population. Its use has been demonstrated in a number of examples, and a number of rules for the calculation of probabilities have been presented. The use of probabilities and the rules for calculating probabilities have been applied to the life table, a basic tool in public health research.

Now that we have an understanding of probability, we shall examine particular probability distributions in the next chapter.

EXERCISES

4.1 Choose the most appropriate answer.

a. If you get 10 straight heads in tossing a fair coin, a tail is _________ on the next toss.
___ more likely
___ less likely
___ neither more likely nor less likely

b. In the U.S. life table, the distribution of the length of life (or age at death) is ___.
___ skewed to the left
___ skewed to the right
___ symmetric

c. A test with high sensitivity is very good at _________.
___ screening out patients who do not have the disease.
___ detecting patients with the disease.
___ determining the probability of the disease.

d. In the U.S. life table the life expectancy (mean) is _________ the median length of life.
___ the same as
___ greater than
___ less than

e. ${}_4q_1$ is called a _________ because an infant cannot die in this interval unless it survived the first year of life.
___ personal probability
___ marginal probability
___ conditional probability

f. In the life table, the mean length of life for those who died during ages 0–1 is _________.
___ about 1/2 year
___ more than 1/2 year
___ less than 1/2 year

4.2 The following table gives estimates of the probabilities that a randomly chosen adult in the United States falls into each of six gender-by-education categories (based on relative frequencies from the NHANES II, NCHS 1982). The three education categories used are (1) less than 12 years, (2) high school graduate, and (3) more than high school graduation.

Categories of Education			
Gender	(1)	(2)	(3)
Female	0.166	0.194	0.164
Male	0.149	0.140	0.187

a. What is the estimate of the probability that an adult is a high school graduate (categories 2 and 3)?

b. What is the estimate of the probability that an adult is a female?

c. From the NHANES II data, it is also estimated that the probability that a female is taking a vitamin supplement is 0.426. What is the estimate of the probability that the adult is a female and taking a vitamin supplement?

d. From the NHANES II, it is also estimated that the probability of adults taking a vitamin supplement is 0.372. What is the estimate of the probability that a male is taking a vitamin supplement?

4.3 Suppose that the failure rate (failing to detect smoke when smoke is present) for a brand of smoke detector is 1 in 2000. For safety, two of these smoke detectors are installed in a laboratory.

a. What is the probability that smoke is not detected in the laboratory when smoke is present in the laboratory?

b. What is the probability that both detectors sound an alarm when smoke is present in the laboratory?

c. What is the probability that one of the detectors sounds the alarm and the other fails to sound the alarm when smoke is present in the laboratory?

4.4 Suppose that the probability of conception for a married woman in any month is 0.2. What is the probability of conception in two months?

4.5 A new contraceptive device is said to have only a 1 in 100 chance of failure. Assume that the probability of conception for a given month, without using any contraceptive, is 20 percent. What is the probability of having at least one unwanted pregnancy if a woman were to use this device for 10 years? [Hint: This would be the complement of the probability of avoiding pregnancy for 10 years or 120 months. The probability of conception for any month with the use of the new contraceptive device would 0.2 * (1 – 0.99). This and related issues are examined by Keyfitz 1971.]

4.6 In a community, 5500 adults were screened for hypertension by the use of a standard sphygmomanometer, and 640 were found to have a diastolic blood pressure of 90 mmHg or higher. A random sample of 100 adults from those with diastolic blood pressure of 90 mmHg or higher and another random sample of 100 adults from those with blood pressure less than 90 mmHg were subjected to more intensive clinical evaluation for hypertension, and 73 and 13 of the respective samples were confirmed as being hypertensive.

a. What is an estimate of the probability that an adult having blood pressure greater than or equal to 90 at the initial screening will actually be hypertensive (predicted value positive)?

b. What is an estimate of the probability that an adult having blood pressure less than 90 at the initial screening will not actually be hypertensive (predicted value negative)?

c. What is an estimate of the probability that an adult in this community is truly hypertensive (prevalence rate of hypertension)?

d. What is an estimate of the probability that a hypertensive person will be found to have blood pressure greater than or equal to 90 at the initial screening (sensitivity)?

e. What is an estimate of the probability that a person without hypertension will have blood pressure less than 90 at the initial screening (specificity)?

4.7 What is the average number of children per family if every couple were to have children until a son is born? Simulate using the random number table in Appendix B or a random number generator in any statistical software.

4.8 Calculate the following probabilities from the 2002 U.S. Abridged Life Table.

a. What is the probability that a 35-year-old person will survive to retirement at age 65?

b. What is the probability that a 20-year-old person will die between ages 55 and 65?

4.10 Calculate the following expected values from the 2002 U.S. Abridged Life Table.

a. How many years is a newborn expected to live before his fifth birthday?

b. How many years is a 20-year-old person expected to live after retirement at age 65? Repeat the calculation for a 60-year-old person. How would you explain the difference?

4.11 Suppose that a couple wants to have children until they have a girl or until they have four children.

a. What is the probability that they have at least two boys?

b. What is the expected number of children?

4.12 The following are tallies of the first digits of the 50 states' populations in the 2000 U.S. Census:

Digit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
Frequencies	14	6	4	7	6	5	3	3	2	50

a. Why do you think digit 1 appears most frequently and digit 9 least frequently?

b. Tabulate the first digits of numerical data that appeared on the front page of today's newspaper, and see whether your findings conform to Benford's law (Hill 1999) [Pr(first significant digit = d) = $\log_{10}(1 + 1 / d)$, $d = 1, 2, \ldots 9$].

4.13 About 1 percent of women have breast cancer. A cancer screening method can detect 80 percent of genuine cancers with a false alarm rate of 10 percent. What is the probability that women producing a positive test result really have breast cancer?

4.14 Suppose that a factory hires 500 men at age 25 and 200 women at age 25 each year. The factory maintains the fixed number of workforce. From the 2002 life tables, the following values are available: For men: $l_{25} = 97746$; $l_{65} = 78556$; $e_{25} = 51.0$; $e_{65} = 16.6$. For women: $l_{20} = 98922$; $l_{65} = 86680$; $e_{20} = 60.7$; $e_{65} = 19.5$.

a. What would be the expected number of retirees at age 65?

b. What would be the expected number of total employees?

REFERENCES

Arias, E. *United States Life Tables, 2002. National Vital Statistics Reports*, Vol. 53, No. 6. Hyattsville, MD: National Center for Health Statistics, 2004.

Harris County Health Department. Mark Canfield, ed. *The Health Status of Harris County Residents: Births, Deaths and Selected Measures of Public Health, 1980–1986*, 1990.

Hill, T. P. "The Difficulty of Faking Data," *Chance* 12(3):27–31, 1999.

Keyfitz, N. "How Birth Control Affects Births," *Social Biology* 18:109–121, 1971.

Lundgren, E., J. Rastad, E. Thrufjell, G. Akerstrom, and S. Ljunghall. "Population-based screening for primary hyperparathyroidism with serum calcium and parathyroid hormone values in menopausal women," *Surgery* 121(3):287–294, 1997.

Mathews, T. J., and B. E. Hamilton. *Trend Analysis of the Sex Ratio at Birth in the United States. National Vital Statistics Reports*, Vol. 53, No. 20. Hyattsville, MD: National Center for Health Statistics, 2005.

National Center for Health Statistics, *Second National Health and Nutrition Examination Survey (NHANES II) 1982*. Tabulation of data for adults 18 years of age and over by the authors.

National Center for Health Statistics. *Health, United States, 1991 and Prevention Profile*. Hyattsville, MD: Public Health Service. DHHS Pub. No. 92–1232, 1992.

National Center for Health Statistics. *Health, United States, 2004 with Chartbook on Trends in the Health of Americans*. Hyattsville, MD: DHHS Pub. No. 2004–1232, 2004.

U.S. Bureau of the Census. The 2000 Census of Population and Housing, Summary Tape File 1A.

Weidner, N., J. Forkman, F. Pozza, P. Bevilacqua, E. N. Allred, D. H. Morre, S. Meli, and G. Gasparini. "Tumor Angiogenesis: A New Significant and Independent Prognostic Indicator in Early State Breast Carcinoma." *Journal of the National Cancer Institute* 84(24):1875–1887, December 16, 1992, issue.

Weiss, N. S. *Clinical Epidemiology: The Study of Outcome of Disease*, New York: Oxford University Press, 1986.