CHAPTER THREE

PROBABILITY

SECTION 3.1 Introduction

In Chapter 2 various techniques for concisely describing data were outlined. But we usually want to do more with data than just describe them. In particular, we might want to test certain specific inferences about the behavior of the data.

EXAMPLE 3.1

Cancer One theory concerning the etiology of breast cancer states that women in a given age group who give birth to their first child relatively late in life (after 30) are at greater risk for eventually developing breast cancer over some time period t than are women who give birth to their first child early in life (before 20). Because women in the upper social classes tend to have children later, this theory has been used to explain why these women have a higher risk of developing breast cancer than women in the lower social classes. To test this hypothesis, we might identify 2000 women from a particular census tract who are currently aged 45–54 and have never had breast cancer, of whom 1000 had their first child before the age of 20 (call this group A) and 1000 after the age of 30 (call this group B). These 2000 women might be followed for 5 years and asked if they had a new case of breast cancer during this period. Suppose that there are 4 new cases of breast cancer out of 1000 in group A and 5 new cases out of 1000 in group B. ■■■

Is this sufficient evidence to confirm a difference in risk between the two groups? Most people would feel uneasy about coming to this conclusion on the basis of such a limited amount of data.

Suppose we had a more ambitious plan and sampled 10,000 women from groups A and B, respectively, and found 40 new cases in group A and 50 new cases in group B and asked the same question. Although we might be more comfortable with the conclusion because of the larger sample size, we would still have to admit that there was some possibility that this apparent difference in the rates could be due to chance.

The problem is that we need a conceptual framework to make these decisions but have not explicitly stated what the framework is. This framework is provided by the underlying concept of **probability**. In this chapter probability is defined and some rules for working with probabilities are introduced. Understanding of probability is essential in the calculation and interpretation of p-values in the statistical tests of subsequent chapters. It also permits a discussion of sensitivity, specificity, and predictive values of screening tests, which are discussed in Section 3.7.

SECTION 3.2 Definition of Probability

EXAMPLE 3.2

Obstetrics Suppose we are interested in the probability of a male live childbirth (or livebirth) among all livebirths in the United States. Conventional wisdom tells us that this probability should be close to .5. We can explore this subject by looking at some vital-statistics data, as presented in Table 3.1 [1]. The probability of a male livebirth based on 1965 data is .51247, based on 1965–1969 data .51248, and based on 1965–1974 data .51268. These are **empirical** probabilities based on a finite amount of data. In principle, the sample size could be expanded indefinitely and an increasingly more precise estimate of this probability obtained.

TABLE 3.1
Probability of a male livebirth during the period 1965–1974

Time period	Number of male livebirths (a)	Total number of livebirths (b)	Empirical probability of a male livebirth (a/b)
1965	1,927,054	3,760,358	0.51247
1965–1969	9,219,202	17,989,361	0.51248
1965–1974	17,857,857	34,832,051	0.51268

This principle leads to the following definition of probability:

DEFINITION 3.1

The **sample space** is the set of all possible outcomes. In referring to probabilities of events, an **event** is any set of outcomes of interest. The **probability** of an event is the relative frequency (see p. 25) of this set of outcomes over an indefinitely large (or infinite) number of trials. ■

EXAMPLE 3.3

Pulmonary Disease The tuberculin skin test is a routine screening test used to detect tuberculosis. The results of this test can be categorized as either positive, negative, or uncertain. If the probability of a positive test is .1, it means that if a large number of such tests were performed, about 10% of them would be positive. The actual percentage of positive tests will be increasingly close to .1 the larger the number of tests performed. ■■■

EXAMPLE 3.4

Cancer The probability of developing a new case of breast cancer in 30 years in 40-year-old women who have never had breast cancer is approximately 1/11. This probability means that over a large sample of 40-year-old women who have never had breast cancer, approximately 1 in 11 will develop the disease over 30 years, with this proportion becoming increasingly close to 1 in 11 as the number of women sampled increases. ■■■

In real life, experiments cannot be performed an infinite number of times. Instead, probabilities of events are estimated from the empirical probabilities obtained from large samples (as was done in Examples 3.2–3.4). In other instances, theoretical probability models are constructed from which probabilities of many different kinds of events can be computed. One of the important issues in statistical inference is to compare empirical probabilities with theoretical probabilities, that is, to assess the goodness of fit of probability models. This topic is covered in Section 10.12.

EXAMPLE 3.5

Cancer The probability of developing a new case of stomach cancer over a 1-year period for 45–49-year-old women based on Connecticut Tumor Registry data from 1963–1965 is 14 per 100,000 [2]. Suppose we have studied cancer rates in a small group of Connecticut nurses over this period and wish to compare how close the rates from this limited sample are to the tumor-registry figures. The figure 14 per 100,000 would be the best estimate of the probability prior to collecting any data, and we would then see how closely our new sample data conformed with this probability. ■■■

From Definition 3.1 and from the preceding examples, we can deduce that probabilities have the following basic properties:

3.1

(1) The probability of an event E, denoted by $Pr(E)$, always satisfies $0 \leq Pr(E) \leq 1$.

(2) If outcomes A and B are two events that cannot both happen at the same time, then $Pr(A \text{ or } B \text{ occurs}) = Pr(A) + Pr(B)$.

EXAMPLE 3.6 **Hypertension** Let A be the event that a person has normotensive diastolic blood-pressure (DBP) readings (i.e., DBP <90), and let B be the event that a person has borderline DBP readings (i.e., DBP ≥ 90 and <95). Suppose that $Pr(A) = .7$, $Pr(B) = .1$. Let C be the event that a person has DBP <95. Then,

$$Pr(C) = Pr(A) + Pr(B) = .8$$

because the events A and B cannot occur at the same time. ■■■

DEFINITION 3.2

Two events A and B are **mutually exclusive** if they cannot both happen at the same time. ■

Thus, the events A and B in Example 3.6 are mutually exclusive.

EXAMPLE 3.7 **Hypertension** Let x be DBP, C be the event that $x \geq 90$, and D be the event that $75 \leq x \leq 100$. The events C and D are *not* mutually exclusive, since they both occur when $90 \leq x \leq 100$. ■■■

SECTION 3.3 Some Useful Probabilistic Notation

DEFINITION 3.3

The symbol { } is used as shorthand for the phrase "the event." ■

DEFINITION 3.4

$A \cup B$ is the event that either A or B occurs or they both occur. ■

Figure 3.1 diagrammatically depicts $A \cup B$ both for the case where A and B are and are not mutually exclusive.

EXAMPLE 3.8 **Hypertension** Let the events A and B be defined as in Example 3.6; that is, $A = \{x < 90\}$, $B = \{90 \leq x < 95\}$, where x = DBP. Then, $A \cup B = \{x < 95\}$. ■■■

EXAMPLE 3.9 **Hypertension** Let the events C and D be defined as in Example 3.7; that is,

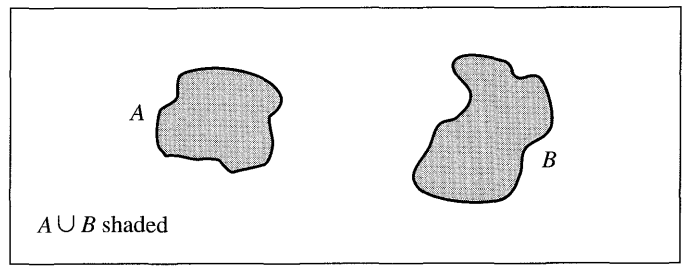
$$C = \{x \geq 90\} \qquad D = \{75 \leq x \leq 100\}$$

Then,

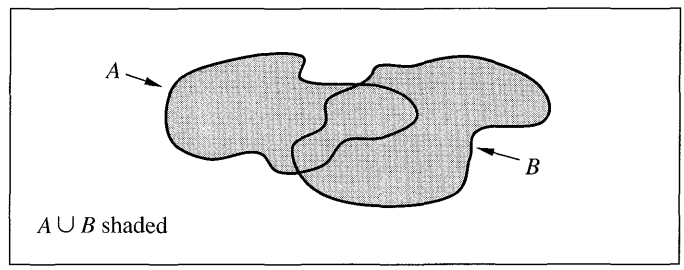
$$\{C \cup D\} = \{x \geq 75\}$$ ■■■

DEFINITION 3.5

$\{A \cap B\}$ is the event that both A and B occur simultaneously. $\{A \cap B\}$ is depicted diagrammatically in Figure 3.2. ■



(a)



(b)

FIGURE 3.1 Diagrammatic representation of $A \cup B$; (a) A, B mutually exclusive; (b) A, B not mutually exclusive

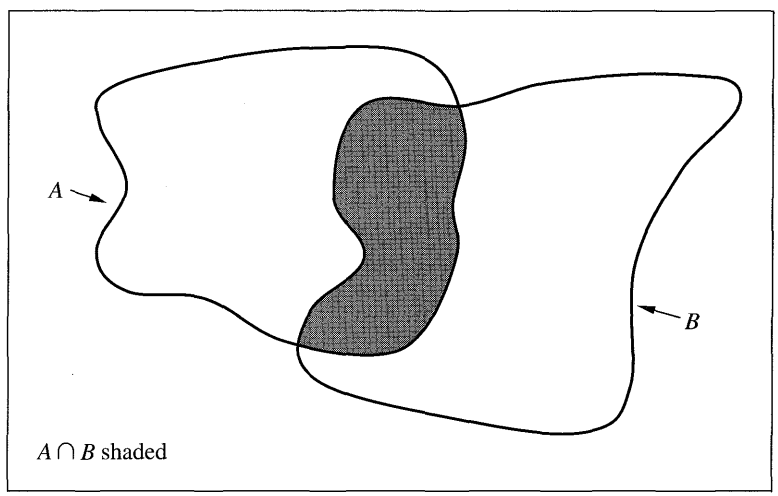


FIGURE 3.2 Diagrammatic representation of $A \cap B$

EXAMPLE 3.10 **Hypertension** Let the events C and D be defined as in Example 3.7; that is,

$$C = \{x \geq 90\} \qquad D = \{75 \leq x \leq 100\}$$

Then,

$$\{C \cap D\} = \{90 \leq x \leq 100\}$$

■■■

Notice that $\{A \cap B\}$ is not well defined for the events A and B in Example 3.6, since both A and B cannot occur simultaneously. This situation is true for any mutually exclusive events.

DEFINITION 3.6

$\bar{A}$ is the event that A does not occur. It is sometimes referred to as the **complement** of A. Notice that $Pr(\bar{A}) = 1 - Pr(A)$, since $\bar{A}$ occurs only when A does not occur. The event $\bar{A}$ is depicted diagrammatically in Figure 3.3. ■

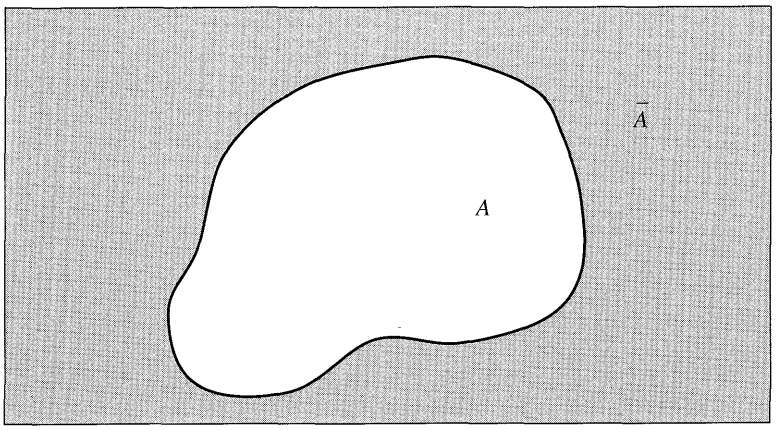


FIGURE 3.3 Diagrammatic representation of $\bar{A}$

EXAMPLE 3.11

Hypertension Let the events A and C be defined as in Examples 3.6 and 3.7; that is,

$$A = \{x < 90\} \qquad C = \{x \geq 90\}$$

Then, $C = \bar{A}$, since C can only occur when A does not occur. Notice that

$$Pr(C) = Pr(\bar{A}) = 1 - .7 = .3$$

Thus, if 70% of people have DBP <90, then 30% of people must have DBP ≥90. ■■■

SECTION 3.4 The Multiplication Law of Probability

In the preceding section, events in general were described. In this section, certain specific types of events are discussed.

EXAMPLE 3.12

Hypertension, Genetics Suppose we are conducting a hypertension-screening program in the home. Consider all possible pairs of DBP measurements of the mother and father within a given family, assuming that the mother and father are not genetically related. This sample space consists of all pairs of numbers of the form (X, Y), where $X > 0$, $Y > 0$. Certain specific events might be of interest in this context. In particular, we might be interested in whether the mother or father is hypertensive, which is described, respectively, by the events A = {mother's DBP ≥ 95}, B = {father's DBP ≥ 95}. These events are depicted graphically in Figure 3.4.

Suppose we know that $Pr(A) = .1$, $Pr(B) = .2$. What can we say about $Pr(A \cap B) = Pr(\text{mother's DBP} \geq 95 \text{ and father's DBP} \geq 95) = Pr(\text{both mother and father are hypertensive})$? We can say nothing unless we are willing to make certain assumptions. ■■■

DEFINITION 3.7

Two events A and B are referred to as **independent events** if

$$Pr(A \cap B) = Pr(A) \times Pr(B)$$ ■

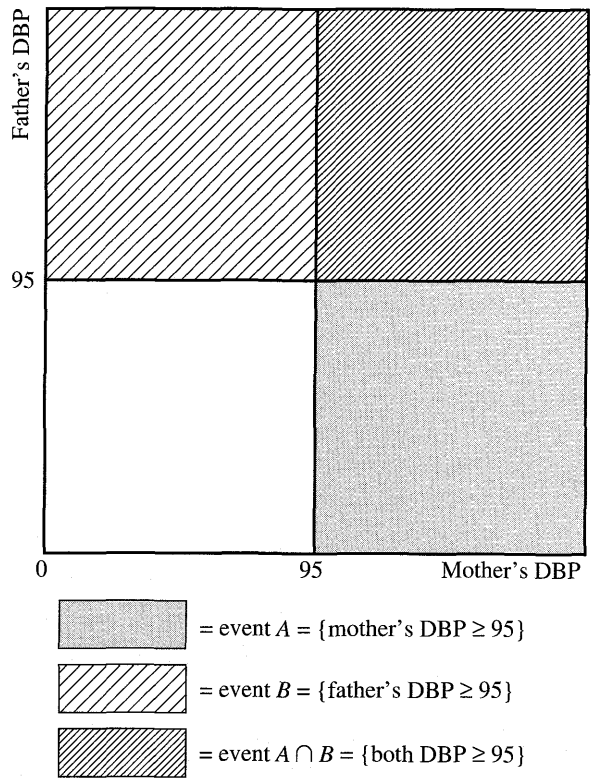


FIGURE 3.4
Possible diastolic blood-pressure measurements of the mother and father within a given family

EXAMPLE 3.13 **Hypertension, Genetics** Compute the probability that both the mother and father are hypertensive if the events in Example 3.12 are independent.

SOLUTION If A and B are independent events, then

$$Pr(A \cap B) = Pr(A) \times Pr(B) = .1(.2) = .02$$ ■■■

One way to interpret this example is to assume that the hypertensive status of the mother does not depend at all on the hypertensive status of the father. Thus, if these events are independent, then in 10% of all households where the father is hypertensive the mother is also hypertensive, and in 10% of all households where the father is *not* hypertensive the mother is hypertensive. We would expect these two events to be independent if the primary determinants of elevated blood pressure were genetic. However, if the primary determinants of elevated blood pressure were, to some extent, environmental, then we would expect that the mother would be more likely to have elevated blood pressure (A true) if the father had elevated blood pressure (B true) than if the father did not have elevated blood pressure (B not true). In this latter case the events would not be independent. The implications of this situation are discussed later in this chapter.

If two events are not independent, then they are said to be dependent.

DEFINITION 3.8

Two events A, B are **dependent** if

$$Pr(A \cap B) \neq Pr(A) \times Pr(B)$$

Example 3.14 is a classic example of dependent events.

EXAMPLE 3.14

Hypertension, Genetics Consider all possible diastolic blood-pressure measurements from a mother and her first-born child. Let

$$A = \{\text{mother's DBP} \geq 95\} \qquad B = \{\text{first-born child's DBP} \geq 80\}$$

Suppose $Pr(A \cap B) = .05 \quad Pr(A) = .1 \quad Pr(B) = .2$

Then $Pr(A \cap B) = .05 > Pr(A) \times Pr(B) = .02$

and the events A, B would be dependent.

This outcome would be expected, since the mother and first-born child both share the same environment and are genetically related. In other words, the first-born child is more likely to have elevated blood pressure in households where the mother is hypertensive than in households where the mother is not hypertensive.

EXAMPLE 3.15

Sexually Transmitted Disease Suppose two doctors, A and B, diagnose all patients coming into a VD clinic for syphilis. Let the events $A^+ = \{\text{doctor } A \text{ makes a positive diagnosis}\}$, $B^+ = \{\text{doctor } B \text{ makes a positive diagnosis}\}$. Suppose that doctor A diagnoses 10% of all patients as positive, doctor B diagnoses 17% of all patients as positive, and both doctors diagnose 8% of all patients as positive. Are the events A^+, B^+ independent?

SOLUTION We are given that

$$Pr(A^+) = .1 \qquad Pr(B^+) = .17 \qquad Pr(A^+ \cap B^+) = .08$$

Thus, $Pr(A^+ \cap B^+) = .08 > Pr(A^+) \times Pr(B^+) = .1(.17) = .017$

and the events are dependent. This result would be expected, since there should be a similarity between how two doctors diagnose patients for syphilis.

Definition 3.7 can be generalized to the case of $k(>2)$ independent events. This is often referred to as the multiplication law of probability.

3.2 If $A_1, \ldots, A_k$ are mutually independent events, then $Pr(A_1 \cap A_2 \cap \cdots \cap A_k) = Pr(A_1) \times Pr(A_2) \times \cdots \times Pr(A_k)$. This principle is referred to as the **multiplication law of probability**.

SECTION 3.5 The Addition Law of Probability

We have seen from the definition of probability that if A and B are mutually exclusive events, then $Pr(A \cup B) = Pr(A) + Pr(B)$. A more general formula for $Pr(A \cup B)$ can be developed when the events A and B are not necessarily mutually exclusive. This formula is referred to as the addition law of probability and is stated as follows:

3.3 Addition Law of Probability

If A and B are any events, then

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

This principle is depicted diagrammatically in Figure 3.5. Thus, to compute $Pr(A \cup B)$, add the probabilities of A and B separately and then subtract the overlap, which is $Pr(A \cap B)$.

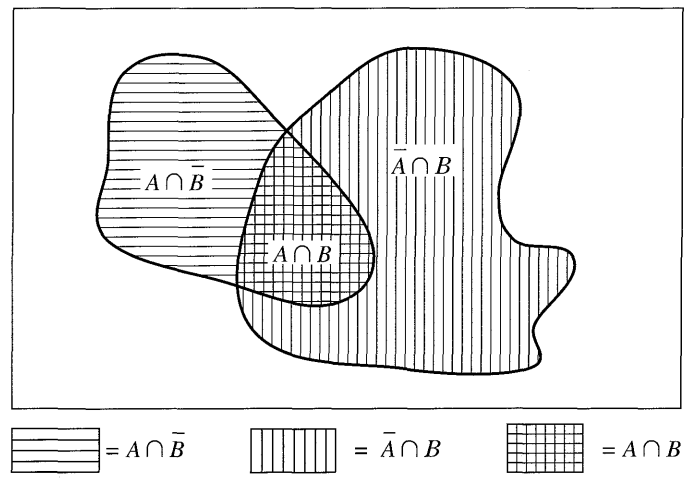


FIGURE 3.5 Diagrammatic representation of the addition law of probability

To derive this result, note that the event $A \cup B$ can be subdivided into three mutually exclusive components, namely, $A \cap \bar{B}$, $\bar{A} \cap B$, $A \cap B$, that, in words, are the events A occurs and B does not occur, A does not occur and B occurs, and both A and B occur. If $A \cup B$ occurs, then exactly one of these events must occur. Therefore,

$$Pr(A \cup B) = Pr(A \cap \bar{B}) + Pr(\bar{A} \cap B) + Pr(A \cap B)$$

However, if A occurs, then it must occur either with B $(A \cap B)$ or without B $(A \cap \bar{B})$ occurring. Therefore,

$$Pr(A) = Pr(A \cap B) + Pr(A \cap \bar{B})$$

If $Pr(A \cap B)$ is subtracted from both sides of the equation,

$$Pr(A \cap \bar{B}) = Pr(A) - Pr(A \cap B)$$

Similarly, if the roles of A and B are interchanged,

$$Pr(\bar{A} \cap B) = Pr(B) - Pr(A \cap B)$$

Finally, by substituting into the expression for $Pr(A \cup B)$,

$$\begin{aligned} Pr(A \cup B) &= Pr(A) - Pr(A \cap B) + [Pr(B) - Pr(A \cap B)] + Pr(A \cap B) \\ &= Pr(A) + Pr(B) - Pr(A \cap B) \end{aligned}$$

EXAMPLE 3.16 **Sexually Transmitted Disease** Consider the data given in Example 3.15. Suppose a patient is referred for further lab tests if either doctor A or B makes a positive diagnosis. What is the probability that a patient will be referred for further lab tests?

SOLUTION The event that either doctor makes a positive diagnosis can be represented by $\{A^+ \cup B^+\}$. We know that

$$Pr(A^+) = .1 \qquad Pr(B^+) = .17 \qquad Pr(A^+ \cap B^+) = .08$$

Therefore, from the addition law of probability,

$$Pr(A^+ \cup B^+) = Pr(A^+) + Pr(B^+) - Pr(A^+ \cap B^+) = .1 + .17 - .08 = .19$$

Thus, 19% of all patients will be referred for further lab tests. ■■■

There are special cases of the addition law that are of interest. First, if the events A and B are *mutually exclusive*, then $Pr(A \cap B) = 0$ and the addition law reduces to $Pr(A \cup B) = Pr(A) + Pr(B)$. This property is given in **(3.1)** for probabilities over any two mutually exclusive events. Second, if the events A and B are *independent*, then by definition $Pr(A \cap B) = Pr(A) \times Pr(B)$ and $Pr(A \cup B)$ can be rewritten as $Pr(A) + Pr(B) - Pr(A) \times Pr(B)$. This leads to the following important special case of the addition law.

3.4 **Addition Law of Probability for Independent Events**

If two events A and B are independent, then

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)]$$

This special case of the addition law can be interpreted as follows: The event $A \cup B$ can be separated into two mutually exclusive events: {A occurs} and {B occurs and A does not occur}. Furthermore, because of the independence of A and B, the probability of the latter event can be written as $Pr(B) \times [1 - Pr(A)]$. This probability is depicted diagrammatically in Figure 3.6.

EXAMPLE 3.17 **Hypertension** Refer to Example 3.12, where

$$A = \{\text{mother's DBP} \geq 95\} \qquad \text{and} \qquad B = \{\text{father's DBP} \geq 95\}$$

$Pr(A) = .1$, $Pr(B) = .2$, and assume that A and B are independent events. Suppose a "hypertensive household" is defined as one in which either the mother or the father is hypertensive, and hypertension is defined for the mother and father, respectively, in terms of the events A and B. What is the probability of a hypertensive household?

SOLUTION Pr(hypertensive household) is

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)] = .1 + .2(.9) = .28$$

Thus, 28% of all households will be hypertensive. ■■■

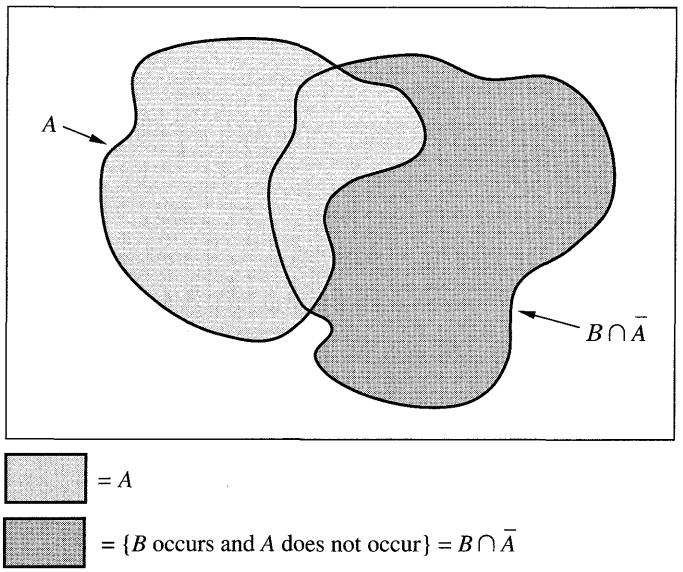


FIGURE 3.6 Diagrammatic representation of the addition law of probability for independent events

It is possible to extend the addition law to more than two events. In particular, if there are three events A, B, and C, then

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C)$$

This result can be generalized to an arbitrary number of events, although this is beyond the scope of this text (see [3]).

SECTION 3.6 Conditional Probability

Suppose we want to compute the probability of several events occurring simultaneously. If the events are independent, then the multiplication law of probability can be used to accomplish this. If some of the events are dependent, then some quantitative measure of dependence is needed in order to extend the multiplication law to the case of dependent events. Consider the following example:

EXAMPLE 3.18

Pulmonary Disease In many places of employment, prospective employees are customarily given a screening test for tuberculosis (TB) before starting employment. The definitive test for the detection of TB is the chest X-ray. Unfortunately, the chest X-ray is somewhat expensive to administer and exposure to the radiation from the X-ray is an undesirable side effect of the test. A common procedure to avoid giving everyone a chest X-ray is to perform a less expensive test, the skin test, with the hope that only people who are positive on the skin test can possibly have TB. The ideal situation would be if the probability of having TB among all those with positive skin tests (SKT) were 1 and the probability of having TB among all those with negative skin tests were 0. The two events $\{SKT^+\}$, $\{TB\}$ would then be completely dependent; that is, the result of the screening test would automatically determine the disease state. The opposite extreme is achieved when the events $\{SKT^+\}$, $\{TB\}$ are completely independent. In this case the probability of TB is the same whether or not the skin test is positive, and the skin test would not be useful in screening for TB and should not be given. ■■■

These concepts can be quantified in the following way. Let $A = \{SKT^+\}$, $B = \{TB\}$ and suppose that we are interested in the probability of TB (B) given that the skin test is positive (A). This probability can be written as $Pr(A \cap B)/Pr(A)$.

DEFINITION 3.9

The quantity $Pr(A \cap B)/Pr(A)$ is defined as the **conditional probability of B given A**, which is written as $Pr(B|A)$. ■

However, from Section 3.4 we know that, by definition, if two events are independent, then $Pr(A \cap B) = Pr(A) \times Pr(B)$. If both sides are divided by $Pr(A)$, then $Pr(B) = Pr(A \cap B)/Pr(A) = Pr(B|A)$. Similarly, we can show that if A and B are independent events, then $Pr(B|\overline{A}) = Pr(B|A) = Pr(B)$. This relationship leads to the following alternative interpretation of independence in terms of conditional probabilities:

3.5

(1) If A and B are independent events, then $Pr(B|A) = Pr(B) = Pr(B|\overline{A})$.

(2) If two events A, B are dependent, then $Pr(B|A) \neq Pr(B) \neq Pr(B|\overline{A})$ and $Pr(A \cap B) \neq Pr(A) \times Pr(B)$.

DEFINITION 3.10

The **relative risk (RR)** of B given A is

$$Pr(B|A)/Pr(B|\overline{A})$$

■

Notice that if two events A, B are independent, then the relative risk will be 1. If two events A, B are dependent, then the relative risk will be different from 1. Heuristically, the more the dependence between events increases, the further the relative risk is from 1.

EXAMPLE 3.19

Pulmonary Disease Suppose that 1 person in 10,000 from those with negative skin tests has TB, or $Pr(B|\overline{A}) = .0001$, whereas 1 person in 100 from those with positive skin tests has TB, or $Pr(B|A) = .01$. The two events would be highly dependent here, since

$$RR = Pr(B|A)/Pr(B|\overline{A}) = .01/.0001 = 100$$

In words, people with positive skin tests are 100 times as likely to have TB as those with negative skin tests. This is the rationale for using the skin test as a screening test for TB. If the events A and B were independent, then the relative risk would be 1; that is, people with positive or negative skin tests would be equally likely to have TB and the test would not be useful as a screening test. ■■■

EXAMPLE 3.20

Sexually Transmitted Disease Using the data in Example 3.15, find the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a positive diagnosis. What is the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a negative diagnosis? What is the relative risk of $\{B^+\}$ given $\{A^+\}$?

SOLUTION $$Pr(B^+|A^+) = Pr(B^+ \cap A^+)/Pr(A^+) = .08/.1 = .8$$

Thus, doctor B will confirm doctor A's positive diagnosis 80% of the time. Similarly,

$$Pr(B^+|A^-) = Pr(B^+ \cap A^-)/Pr(A^-) = Pr(B^+ \cap A^-)/.9$$

We must compute $Pr(B^+ \cap A^-)$. We know that if doctor B diagnoses a patient as positive, then doctor A either does or does not diagnose the patient as positive. Thus,

$$Pr(B^+) = Pr(B^+ \cap A^+) + Pr(B^+ \cap A^-)$$

since the events $\{B^+ \cap A^+\}$ and $\{B^+ \cap A^-\}$ are mutually exclusive. If we subtract $Pr(B^+ \cap A^+)$ from both sides of the equation, then

$$Pr(B^+ \cap A^-) = Pr(B^+) - Pr(B^+ \cap A^+) = .17 - .08 = .09$$

Therefore, $$Pr(B^+|A^-) = .09/.9 = .1$$

Thus, when doctor A diagnoses a patient as negative, doctor B will contradict the diagnosis 10% of the time. The relative risk of the event $\{B^+\}$ given $\{A^+\}$ is

$$Pr(B^+|A^+)/Pr(B^+|A^-) = .8/.1 = 8$$

This indicates that doctor B is 8 times as likely to diagnose a patient as positive when doctor A diagnoses the patient as positive than when doctor A diagnoses the patient as negative. These results quantify the dependence between the two doctors' diagnoses. ■■■

The conditional ($Pr(B|A)$, $Pr(B|\overline{A})$) and unconditional ($Pr(B)$) probabilities mentioned previously can be related in the following way:

3.6 For any events A and B,

$$Pr(B) = Pr(B|A) \times Pr(A) + Pr(B|\overline{A}) \times Pr(\overline{A})$$

This formula tells us that the unconditional probability of B is the sum of the conditional probability of B given A *times* the unconditional probability of A *plus* the conditional probability of B given A *not* occurring *times* the unconditional probability of A *not* occurring.

To derive this, we note that if the event B occurs, it must occur either with A or without A. Therefore,

$$Pr(B) = Pr(B \cap A) + Pr(B \cap \overline{A})$$

From the definition of conditional probability, we see that

$$Pr(B \cap A) = Pr(A) \times Pr(B|A)$$

and

$$Pr(B \cap \overline{A}) = Pr(\overline{A}) \times Pr(B|\overline{A})$$

By substitution, it follows that

$$Pr(B) = Pr(B|A)Pr(A) + Pr(B|\overline{A})Pr(\overline{A})$$

EXAMPLE 3.21

Pulmonary Disease Let A and B be defined as in Example 3.19 and suppose that 1% of the general population will have a positive skin test. What is the probability of tuberculosis in the general population?

SOLUTION

$$Pr(B) = Pr(\text{TB}) = Pr(\text{TB}|\text{SKT}^+) \times Pr(\text{SKT}^+) + Pr(\text{TB}|\text{SKT}^-) \times Pr(\text{SKT}^-)$$
$$= .01(.01) + 10^{-4}(.99) = .000199 \simeq .0002 = 2 \times 10^{-4}$$

Thus, the unconditional probability of TB in the general population (2×10^{-4}) is a weighted average of the conditional probability of TB given a positive skin test ($.01 = 100 \times 10^{-4}$) and the conditional probability of TB given a negative skin test (10^{-4}). ■■■

In **(3.6)** the probability of the event B is expressed in terms of the two events A and $\overline{A}$. In many instances the probability of an event B will need to be expressed in terms of more than two mutually exclusive events, denoted by $A_1, A_2, \ldots, A_k$.

DEFINITION 3.11

A set of events $A_1, \ldots, A_k$ is **exhaustive** if at least one of the events must occur. ■

Assume that the events $A_1, \ldots, A_k$ are mutually exclusive and exhaustive, that is, at least one of the events $A_1, \ldots, A_k$ must occur and no two events can occur simultaneously. Thus, exactly one of the events $A_1, \ldots, A_k$ must occur.

3.7 Total Probability Rule

Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. The unconditional probability of B ($Pr(B)$) can then be written as a weighted average of the conditional probabilities of B given A_i ($Pr(B|A_i)$) as follows:

$$Pr(B) = \sum_{i=1}^{k} Pr(B|A_i) \times Pr(A_i)$$

To show this, we note that if B occurs, then it must occur together with one and only one of the events, $A_1, \ldots, A_k$. Therefore,

$$Pr(B) = \sum_{i=1}^{k} Pr(B \cap A_i)$$

Also, from the definition of conditional probability,

$$Pr(B \cap A_i) = Pr(A_i) \times Pr(B|A_i)$$

By substitution, we obtain Equation **(3.7)**.

An application of the total probability rule is given in the following example:

EXAMPLE 3.22

Ophthalmology We are planning a 5-year study of cataract in a population of 5000 people 60 years of age and older. We know from census data that 45% of this population are ages 60–64, 28% are ages 65–69, 20% are ages 70–74, and 7% are age 75 or older. We also know from the Framingham Eye Study that 2.4%, 4.6%, 8.8%, and 15.3% of the people in those respective age groups will develop cataract over the next 5 years [4]. What percentage of our population will develop cateract over the next 5 years, and how many cataracts does this percentage represent?

SOLUTION Let A_1 = {ages 60–64}, A_2 = {ages 65–69}, A_3 = {ages 70–74}, A_4 = {age 75+}. These events are mutually exclusive and exhaustive, since exactly one event must occur for each person in our population. Furthermore, from the conditions of the problem, we know that $Pr(A_1) = .45$, $Pr(A_2) = .28$, $Pr(A_3) = .20$, $Pr(A_4) = .07$, $Pr(B|A_1) = .024$, $Pr(B|A_2) = .046$, $Pr(B|A_3) = .088$, and $Pr(B|A_4) = .153$. Finally, using the total probability rule,

$$\begin{aligned} Pr(B) &= Pr(B|A_1) \times Pr(A_1) + Pr(B|A_2) \times Pr(A_2) \\ &\quad + Pr(B|A_3) \times Pr(A_3) + Pr(B|A_4) \times Pr(A_4) \\ &= .024(.45) + .046(.28) + .088(.20) + .153(.07) = .052 \end{aligned}$$

Thus 5.2% of our population will develop cataract over the next 5 years, which represents a total of $5000 \times .052 = 260$ persons with cataract. ■■■

The definition of conditional probability allows the multiplication law of probability to be extended to the case of dependent events.

3.8 **Generalized Multiplication Law of Probability**

If $A_1, \ldots, A_k$ are an arbitrary set of events, then

$$Pr(A_1 \cap A_2 \cap \cdots \cap A_k) = Pr(A_1) \times Pr(A_2|A_1) \times Pr(A_3|A_2 \cap A_1) \times \cdots \times Pr(A_k|A_{k-1} \cap \cdots \cap A_2 \cap A_1)$$

If the events are independent, then the conditional probabilities on the right-hand side of **(3.8)** reduce to unconditional probabilities, and the generalized multiplication law reduces to the multiplication law for independent events given in **(3.2)**. Equation **(3.8)** also generalizes the relationship $Pr(A \cap B) = Pr(A) \times Pr(B|A)$ given in Definition 3.9 for two events to the case of more than two events.

SECTION 3.7 Bayes' Rule and Screening Tests

The tuberculosis skin-test example given in Example 3.18 illustrates the general concept of the predictive value of a screening test, which can be defined as follows:

DEFINITION 3.12

The **predictive value positive** (PV^+) of a screening test is the probability that a person has disease given that the test is positive

$$Pr(\text{disease}|\text{test}^+)$$

The **predictive value negative** (PV^-) of a screening test is the probability that a person does *not* have disease given that the test is negative

$$Pr(\text{no disease}|\text{test}^-)$$

EXAMPLE 3.23 **Pulmonary Disease** Find the predictive values positive and negative for the tuberculosis skin test given the data in Example 3.19.

SOLUTION We see that

$$PV^+ = Pr(B|A) = .01$$

whereas

$$PV^- = Pr(\overline{B}|\overline{A}) = 1 - Pr(B|\overline{A}) = .9999$$

Thus, if the skin test is negative, the person is virtually certain not to have disease ($PV^- \approx 1$); whereas if the skin test is positive, the person still has only a small chance of having disease ($PV^+ = .01$). ■■■

A symptom or a set of symptoms can also be regarded as a screening test for disease. The higher the predictive value of the screening test or symptoms, the more valuable the test. Ideally, we would like to find a set of symptoms such that both PV^+ and PV^- are 1. Then we would be able to accurately diagnose disease for each patient.

Clinicians often cannot directly measure the predictive value of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people. These measures are defined as follows:

DEFINITION 3.13

The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has disease. ■

DEFINITION 3.14

The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is not present given that the person does not have disease. ■

DEFINITION 3.15

A **false negative** is defined as a person who tests out as negative but who is actually positive. A **false positive** is defined as a person who tests out as positive but who is actually negative. ■

It is important that both the sensitivity and specificity be high for a symptom to be effective in predicting disease.

EXAMPLE 3.24 **Cancer** Suppose that the disease is lung cancer and the symptom is cigarette smoking. If we assume that 90% of people with lung cancer and 50% of people without lung cancer (essentially the entire general population) are smokers, then the sensitivity and specificity are .9 and .5, respectively. Obviously cigarette smoking cannot be used by itself as a diagnostic tool for predicting lung cancer, because there will be too many false positives (normal people who are smokers). ■■■

EXAMPLE 3.25 **Cancer** Suppose that the disease is breast cancer in women and the symptom is having a family history of breast cancer (i.e, either a mother or a sister with breast cancer). If we assume that 5% of women with breast cancer have a family history of breast cancer whereas only 2% of women without breast cancer have such a history, then the sensitivity is .05 and the specificity is .98 = (1 − .02). A family history of breast cancer cannot be used by itself to diagnose breast cancer because there will be too many false negatives (i.e., women without a family history who have the disease). ■■■

How can the sensitivity and specificity of a symptom (or set of symptoms), which are quantities a physician can estimate, be used to compute predictive values, which are quantities a physician needs to make appropriate diagnoses?

Let A = symptom and B = disease. From Definitions 3.12, 3.13, and 3.14, we have

$$\text{Predictive value positive} = \text{PV}^+ = Pr(B|A)$$

$$\text{Predictive value negative} = \text{PV}^- = Pr(\overline{B}|\overline{A})$$

$$\text{Sensitivity} = Pr(A|B)$$

$$\text{Specificity} = Pr(\overline{A}|\overline{B})$$

Let $Pr(B)$ = probability of disease in the reference population. We wish to compute $Pr(B|A)$ and $Pr(\overline{B}|\overline{A})$ in terms of the other quantities. This relationship is known as Bayes' rule.

3.9 Bayes' Rule

Let A = symptom and B = disease.

$$\text{PV}^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})}$$

In words, this can be written as

$$\text{PV}^+ = \frac{x \times \text{sensitivity}}{x \times \text{sensitivity} + (1 - x) \times (1 - \text{specificity})}$$

where $x = Pr(B)$ = prevalence of disease in the reference population. Similarly

$$\text{PV}^- = \frac{(1 - x) \times \text{specificity}}{(1 - x) \times \text{specificity} + x \times (1 - \text{sensitivity})}$$

To derive this, we have from the definition of conditional probability,

$$\text{PV}^+ = Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B \cap A) = Pr(A|B) \times Pr(B)$$

Finally, from the total probability rule,

$$Pr(A) = Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})$$

If the expressions for $Pr(B \cap A)$ and $Pr(A)$ are substituted into the equation for PV^+, we obtain

$$\text{PV}^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})}$$

That is, PV^+ can be expressed as a function of sensitivity, specificity, and probability of disease in the reference population. A similar derivation can be used for PV^-.

EXAMPLE 3.26 **Hypertension** Suppose that 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What is the predictive value positive and predictive value negative of the machine, assuming that 20% of the adult population is hypertensive?

SOLUTION The sensitivity = .84 and specificity = 1 − .23 = .77. Thus, from Bayes' rule it follows that

$$\begin{aligned} PV^+ &= .2(.84)/[.2(.84) + .8(.23)] \\ &= .168/.352 = .48 \end{aligned}$$

Similarly,

$$\begin{aligned} PV^- &= .8(.77)/[.8(.77) + .2(.16)] \\ &= .616/.648 = .95 \end{aligned}$$

Thus, a negative result from the machine is very predictive, since we are 95% sure that such a person is normotensive. However, a positive result is not very predictive, since we are only 48% sure that such a person is hypertensive. ■■■

In Example 3.26 there were only two possible disease states: hypertensive and normotensive. In clinical medicine there are often more than two possible disease states. We would like to be able to predict the most likely disease state given a specific symptom (or set of symptoms). We will assume that the probability of having these symptoms for each disease state is known from clinical experience, as is the probability of each of the disease states in the reference population. This leads us to the generalized Bayes' rule:

3.10 **Generalized Bayes' Rule**

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive disease states, that is, at least one disease state must occur and no two disease states can occur at the same time. Let A represent the presence of a symptom or set of symptoms. Then

$$Pr(B_i|A) = Pr(A|B_i) \times Pr(B_i) \Big/ \left[\sum_{j=1}^{k} Pr(A|B_j) \times Pr(B_j)\right]$$

This result is obtained in a similar manner to that of Bayes' rule for two disease states in **(3.9)**. Specifically, from the definition of conditional probability, note that

$$Pr(B_i|A) = \frac{Pr(B_i \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B_i \cap A) = Pr(A|B_i) \times Pr(B_i)$$

From the total probability rule,

$$Pr(A) = Pr(A|B_1) \times Pr(B_1) + \cdots + Pr(A|B_k) \times Pr(B_k)$$

If the expressions for $Pr(B_i \cap A)$ and $Pr(A)$ are substituted we obtain

$$Pr(B_i|A) = \frac{Pr(A|B_i) \times Pr(B_i)}{\sum_{j=1}^{k} Pr(A|B_j) \times Pr(B_j)}$$

EXAMPLE 3.27 **Pulmonary Disease** Suppose that a 60-year-old male who has never smoked cigarettes presents with symptoms consisting of a chronic cough and occasional breathlessness to a physician. The physician becomes concerned and orders the patient admitted to the hospital for a lung biopsy. Suppose that the results of the lung biopsy are consistent with either lung cancer or sarcoidosis, a fairly common, nonfatal lung disease. In this case

$$\text{Symptoms } A = \{\text{chronic cough, results of lung biopsy}\}$$

$$\text{Disease state } B_1 = \text{normal}$$

$$B_2 = \text{lung cancer}$$

$$B_3 = \text{sarcoidosis}$$

Suppose that $Pr(A|B_1) = .001 \quad Pr(A|B_2) = .9 \quad Pr(A|B_3) = .9$

and that in 60-year-old, never-smoking males

$$Pr(B_1) = .99 \quad Pr(B_2) = .001 \quad Pr(B_3) = .009$$

The first set of probabilities $Pr(A|B_i)$ could be obtained from clinical experience with the previous diseases, whereas the latter set of probabilities $Pr(B_i)$ would have to be obtained from age-sex-smoking specific prevalence rates for the diseases in question. The interesting question now is what are the probabilities $Pr(B_i|A)$ of the three disease states given the previous symptoms?

SOLUTION Bayes' rule can be used to answer this question. Specifically,

$$\begin{aligned} Pr(B_1|A) &= Pr(A|B_1) \times Pr(B_1) \Big/ \left[\sum_{j=1}^{3} Pr(A|B_j) \times Pr(B_j)\right] \\ &= .001(.99)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00099/.00999 = .099 \end{aligned}$$

$$\begin{aligned} Pr(B_2|A) &= .9(.001)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00090/.00999 = .090 \end{aligned}$$

$$\begin{aligned} Pr(B_3|A) &= .9(.009)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00810/.00999 = .811 \end{aligned}$$

Thus, although the unconditional probability of sarcoidosis is very low (.009), the conditional probability of the disease given these symptoms and this age-sex-smoking group is .811. Also, although the symptoms are consistent with both lung cancer and sarcoidosis, the latter is much more likely among patients in this age-sex-smoking group. ■■■

EXAMPLE 3.28 **Pulmonary Disease** Now, suppose that the patient in Example 3.27 was a smoker of two packs of cigarettes per day for 40 years. Then, assume that $Pr(B_1) = .98$, $Pr(B_2) = .015$, $Pr(B_3) = .005$ in this type of person. What are the probabilities of the three disease states given these symptoms for this type of patient?

SOLUTION

$$\begin{aligned} Pr(B_1|A) &= .001(.98)/[.001(.98) + .9(.015) + .9(.005)] \\ &= .00098/.01898 = .052 \end{aligned}$$

$$Pr(B_2|A) = .9(.015)/.01898 = .01350/.01898 = .711$$

$$Pr(B_3|A) = .9(.005)/.01898 = .237$$

Thus, in this type of patient, lung cancer is the most likely diagnosis. ■■■

SECTION 3.8 Prevalence and Incidence

In clinical medicine, the terms *prevalence* and *incidence* are used to denote probabilities in a special context and are used frequently in this text.

DEFINITION 3.16

The **prevalence** of a disease is the probability of currently having that disease regardless of the duration of time one has had the disease. It is obtained by dividing the number of people who currently have the disease by the number of people in the study population. ■

EXAMPLE 3.29

Hypertension The prevalence of hypertension in 1974 among all people 17 years of age and older was reported to be 15.7% as assessed by a government study [5]. It was computed by dividing the number of people who had elevated blood pressure and were 17 years of age and older (22,626) by the total number of people 17 years of age and older in the study population (144,380). ■■■

DEFINITION 3.17

The **cumulative incidence** of a disease is the probability that an individual with no prior disease will develop a new case of the disease over some specified time period. ■

In Chapter 13 we will distinguish between *cumulative incidence*, which is defined over a long period of time, and *incidence density*, which is defined over a very short (or instantaneous) period of time. For simplicity, prior to Chapter 13 the abbreviated term *incidence* will be used to denote *cumulative incidence*.

EXAMPLE 3.30

Cancer The annual incidence rate of breast cancer in 40–44-year-old Connecticut women over the time period January 1, 1970, through December 31, 1970, was approximately 1 per 1000 [2]. This rate means that about 1 woman in 1000 of the 40–44-year-old women who had never had breast cancer on January 1, 1970, would have developed a new case of breast cancer by December 31, 1970. ■■■

SECTION 3.9 Summary

In this chapter, probabilities and how to work with them using the addition and multiplication laws were discussed. An important distinction was made between independent events, which are unrelated to each other, and dependent events, which tend to occur simultaneously. The general concepts of conditional probability and relative risk were introduced to quantify the dependence between two events. These ideas were then applied to the special area of screening populations for disease. In particular, the notions of sensitivity, specificity, and predictive value, which are used to define the accuracy of screening tests, were developed as applications of conditional probability.

On some occasions, only sensitivities and specificities are available and we wish to compute the predictive value of screening tests. This task can be accomplished using Bayes' rule. Indeed, Bayes' rule can be used generally to change the direction of conditional probabilities. Finally, prevalence and incidence, which are probabilistic parameters that are often used to describe the magnitude of disease in a population, were defined.

In the next two chapters, these general principles of probability are applied to derive some of the important probabilistic models often used in biomedical research, including the binomial, Poisson, and normal models. These models will be used eventually to test hypotheses about data.

PROBLEMS

Consider a family with a mother, father, and two children. Let A_1 = {mother has influenza}, A_2 = {father has influenza}, A_3 = {first child has influenza}, A_4 = {second child has influenza}, B = {at least one child has influenza}, C = {at least one parent has influenza}, D = {at least one person in the family has influenza}.

* **3.1** What does $A_1 \cup A_2$ mean?

* **3.2** What does $A_1 \cap A_2$ mean?

* **3.3** Are A_3 and A_4 mutually exclusive?

* **3.4** What does $A_3 \cup B$ mean?

* **3.5** What does $A_3 \cap B$ mean?

* **3.6** Express C in terms of A_1, A_2, A_3, A_4.

* **3.7** Express D in terms of B and C.

* **3.8** What does $\overline{A}_1$ mean?

* **3.9** What does $\overline{A}_2$ mean?

* **3.10** Represent $\overline{C}$ in terms of A_1, A_2, A_3, A_4.

* **3.11** Represent $\overline{D}$ in terms of B and C.

Suppose that an influenza epidemic strikes a city. In 10% of families the mother has influenza; in 10% of families the father has influenza; and in 2% of families both the mother and father have influenza.

3.12 Are the events A_1, A_2 independent?

Suppose that there is a 20% chance that each child will get influenza, whereas in 10% of two-child families, both children get the disease.

3.13 What is the probability that at least one child will get influenza?

Hypertension

Multiple drugs are often used in treating hypertension. Suppose that 10% of patients taking antihypertensive agent A experience gastrointestinal (GI) side effects, whereas 20% of patients taking antihypertensive agent B experience such side effects.

3.14 If the side effects of the two agents are assumed to be independent events, then what is the probability that a patient taking the two agents simultaneously will experience GI side effects?

Refer to Problem 3.12.

3.15 What is the conditional probability that the father has influenza given that the mother has influenza?

3.16 What is the conditional probability that the father has influenza given that the mother does not have influenza?

Mental Health

Estimates of the prevalence of Alzheimer's disease have recently been provided by Pfeffer et al. [6]. The estimates are given in Table 3.2.

TABLE 3.2 Prevalence of Alzheimer's disease (cases per 100 population)

Age group	Males	Females
65–69	1.6	0.0
70–74	0.0	2.2
75–79	4.9	2.3
80–84	8.6	7.8
85+	35.0	27.9

Suppose an unrelated 77-year-old man, 76-year-old woman, and 82-year-old woman are selected from a community.

3.17 What is the probability that all three of these individuals have Alzheimer's disease?

3.18 What is the probability that at least one of the women has Alzheimer's disease?

3.19 What is the probability that at least one of the three individuals has Alzheimer's disease?

3.20 What is the probability that exactly one of the three individuals has Alzheimer's disease?

3.21 Suppose we know that one of the three individuals has Alzheimer's disease, but we don't know which one. What is the conditional probability that the affected individual is a woman?

3.22 Suppose we know that two of the three individuals have Alzheimer's disease. What is the conditional probability that they are both women?

3.23 Suppose we know that two of the three individuals have Alzheimer's disease. What is the conditional probability that they are both less than 80 years old?

Suppose the probability that both members of a married couple will have the disease, where each member is 75–79 years old, is .0015.

3.24 What is the conditional probability that the man will be affected given that the woman is affected? How does this value compare to the prevalence in Table 3.2? Why should it be the same (or different)?

3.25 What is the conditional probability that the woman will be affected given that the man is affected? How does this value compare to the prevalence in Table 3.2? Why should it be the same (or different)?

3.26 What is the probability that at least one member of the couple is affected?

Suppose a study of Alzheimer's disease is proposed in a retirement community with persons 65+ years of age, where the age-sex distribution is as shown in Table 3.3.

TABLE 3.3 Age-sex distribution of retirement community

	Male (%)[a]	Female (%)
65–69	5	10
70–74	9	17
75–79	11	18
80–84	8	12
85+	4	6

[a]Percentage of total population.

3.27 What is the expected overall prevalence of Alzheimer's disease in the community, if the prevalence estimates in Table 3.2 for specific age-sex groups holds?

3.28 If there are 1000 persons 65+ years of age in the community, then what is the expected number of cases of Alzheimer's disease in the community?

Occupational Health

A study is conducted on male workers 50–69 years old working in a chemical plant. We are interested in comparing the mortality experience of the workers in the plant with national mortality rates. Suppose that of the 500 workers in this age group in the plant, 35% are 50–54, 30% are 55–59, 20% are 60–64, and 15% are 65–69.

* **3.29** If the annual national mortality rates are 0.9% in 50–54-year-old men, 1.4% in 55–59-year-old men, 2.2% in 60–64-year-old men, and 3.3% in 65–69-year-old men, then what is the projected annual mortality rate in the plant as a whole?

The SMR (standardized mortality ratio) is often used in occupational studies as a measure of risk. It is defined as 100% *times* the observed number of events in the exposed group *divided by* the expected number of events in the exposed group (based on some reference population).

* **3.30** If 15 deaths are observed over 1 year among the 500 workers, then what is the SMR?

Genetics

Suppose that a disease is inherited via a **dominant** mode of inheritance and that one of two parents is affected with the disease whereas one is not. The implications of this mode of inheritance are that the probability is $\frac{1}{2}$ that any particular offspring will get the disease.

3.31 What is the probability that in a family with two children, both siblings are affected?

3.32 What is the probability that exactly one sibling is affected?

3.33 What is the probability that neither sibling will be affected?

3.34 Suppose that the older child is affected. What is the probability that the younger child will be affected?

3.35 If A, B are two events such that A = {older child is affected}, B = {younger child is affected}, then are the events A, B independent?

Suppose that a disease is inherited via an **autosomal recessive** mode of inheritance. The implications of this mode of inheritance are that the children in a family each have a probability of $\frac{1}{4}$ of inheriting the disease.

3.36 What is the probability that in a family with two children, both siblings are affected?

3.37 What is the probability that exactly one sibling is affected?

3.38 What is the probability that neither sibling is affected?

Suppose that a disease is inherited via a **sex-linked** mode of inheritance. The implications of this mode of inheritance are that each male offspring has a 50% chance of inheriting the disease, whereas the female offspring have no chance of getting the disease.

3.39 In a family with one male and one female offspring, what is the probability that both siblings are affected?

3.40 What is the probability that exactly one sibling is affected?

3.41 What is the probability that neither sibling is affected?

3.42 Answer Problem 3.39 for families with two male siblings.

3.43 Answer Problem 3.40 for families with two male siblings.

3.44 Answer Problem 3.41 for families with two male siblings.

Suppose that in a family with two male siblings, both siblings are affected with a genetically inherited disease. Suppose also that, although the genetic history of the family is unknown, only a dominant, recessive, or sex-linked mode of inheritance is possible.

3.45 Assume that the dominant, recessive, and sex-linked modes of inheritance follow the probability laws given in Problems 3.31, 3.36, and 3.39 and that, without prior knowledge about the family in question, each is equally likely to occur. What is the probability of each mode of inheritance in this family?

3.46 Answer Problem 3.45 for a family with two male siblings where only one sibling is affected.

3.47 Answer Problem 3.45 for a family with one male and one female sibling where both siblings are affected.

3.48 Answer Problem 3.47 where only the male sibling is affected.

Obstetrics

The following data are derived from the 1973 Final Natality Statistics Report issued by the National Center for Health Statistics [7]. These data are pertinent to live births only.

Suppose that infants are classified as low birthweight if they have a birthweight ≤ 2500 g and as normal birthweight if they have a birthweight ≥ 2501 g. Suppose that infants are also classified by length of gestation in the following four categories: <20 weeks, 20–27 weeks, 28–36 weeks, >36 weeks. Assume that the probabilities of the different periods of gestation are as given in Table 3.4.

TABLE 3.4 Distribution of length of gestation

Length of gestation	Probability
<20 weeks	.0004
20–27 weeks	.0059
28–36 weeks	.0855
>36 weeks	.9082

Also assume that the probability of being low birthweight given that the length of gestation is <20 weeks is .540, the probability of being low birthweight given that the length of gestation is 20–27 weeks is .813, the probability of being low birthweight given that the length of gestation is 28–36 weeks is .379, and the probability of being low birthweight given that the length of gestation is >36 weeks is .035.

* **3.49** What is the probability of having a low birthweight infant?

3.50 Show that the events (length of gestation ≤ 27 weeks) and (low birthweight) are not independent.

* **3.51** What is the probability of having a length of gestation ≤ 36 weeks given that a child is low birthweight?

Pulmonary Disease

A 1974 paper by Colley et al. looked at the relationship between parental smoking and the incidence of pneumonia and/or bronchitis in children in the first year of life [8]. One important finding of the paper was that 7.8% of children with nonsmoking parents had episodes of pneumonia and/or bronchitis in the first year of life, whereas, respectively, 11.4% of children with one smoking parent and 17.6% of children with two smoking parents had such an episode. Suppose that in the general population both parents are smokers in 40% of households, one parent smokes in 25% of households, and neither parent smokes in 35% of households.

3.52 What percentage of children in the general population will have pneumonia and/or bronchitis in the first year of life?

A group of families in which both parents smoke at the time of the first prenatal visit decide, after counseling by the nurse practitioner, to give up smoking. Suppose that in 10% of these families both parents resume smoking and in 30% of these families one parent resumes smoking. In the remainder of the families both parents have not resumed smoking at the time of birth of the child. Assume also that the smoking status of the parents at the time of the birth is maintained during the first year of life of the child.

3.53 What is the probability of pneumonia and/or bronchitis in children from families in this group?

3.54 Among families where both parents smoke, what percentage of cases of pneumonia and/or bronchitis have been prevented by this type of counseling?

Pulmonary Disease

The familial aggregation of respiratory disease is a well-established clinical phenomenon. However, whether this aggregation is due to genetic or environmental factors or both is somewhat controversial. An investigator wishes to study a particular environmental factor, namely, the relationship of cigarette-smoking habits in the parents to the presence or absence of asthma in their oldest child living in the household in the 5–9-year-old age range (referred to below as their offspring). Suppose that the investigator finds that (i) if both the mother and father are current smokers, then the probability of their offspring having asthma is .15; (ii) if the mother is a current smoker and the father is not, then the probability of their offspring having asthma is .13; (iii) if the father is a current smoker and the mother is not, then the probability of their offspring having asthma is .05; (iv) if neither parent is a current smoker, then the probability of their offspring having asthma is .04.

* **3.55** Suppose that the smoking habits of the parents are independent and that the probability that the mother is a current smoker is .4, whereas the probability that the father is a current smoker is .5. What is the probability that both the father and the mother are current smokers?

* **3.56** What is the probability that the father is a current smoker if the mother is not a current smoker?

Suppose, alternatively, that if the father is a current smoker, then the probability that the mother is a current smoker is .6; whereas if the father is not a current smoker, then the probability that the mother is a current smoker is .2. Also assume that statements (i), (ii), (iii), and (iv) above hold.

* **3.57** If the probability that the father is a current smoker is .5, what is the probability that the father is a current current smoker *and* that the mother is not a current smoker?

* **3.58** Are the current smoking habits of the father and the mother independent? Why or why not?

* **3.59** Find the unconditional probability that the offspring will have asthma under the assumptions in Problems 3.57 and 3.58.

* **3.60** Suppose that a child has asthma. What is the probability that the father is a current smoker?

* **3.61** What is the probability that the mother is a current smoker if the child has asthma?

* **3.62** Answer Problem 3.60 if the child does not have asthma.

* **3.63** Answer Problem 3.61 if the child does not have asthma.

* **3.64** Are the child's asthma status and the father's smoking status independent? Why or why not?

* **3.65** Are the child's asthma status and the mother's smoking status independent? Why or why not?

Pulmonary Disease

Smoking cessation is an important dimension in public health programs aimed at the prevention of cancer and heart and lung diseases. For this purpose data were accumulated starting in 1962 on a group of current smoking men as part of the Normative Aging Study, a longitudinal study of the Veterans Administration in Boston. No interventions were attempted on this group of men, but the data in Table 3.5 were obtained as to annual quitting rates among initially healthy men who remained healthy during the entire period [9]:

TABLE 3.5 Annual quitting rates of men who smoked, from the Normative Aging Study, 1962–1975

Time period	Light smokers (≤ one pack per day) average annual quitting rate per 100 persons	Heavy smokers (> one pack per day) average annual quitting rate per 100 persons
1962–1966	3.1	2.0
1967–1970	7.1	5.0
1971–1975	4.7	4.1

Note that the quitting rates increased during the period of 1967 to 1970, which was around the time of the first Surgeon General's report on cigarette smoking.

3.66 Suppose a man was a light smoker on January 1, 1962. What is the probability that he quit smoking by the end of 1975 (a 14-year period)? (Assume that he remained a light smoker until just prior to quitting.)

3.67 Answer Problem 3.66 for a heavy smoker on January 1, 1962 (assume that he remained a heavy smoker until just prior to quitting).

Pulmonary Disease

Research into cigarette-smoking habits, smoking prevention, and cessation programs necessitates accurate measurement of smoking behavior. However, decreasing social acceptability of smoking appears to engender significant underreporting. Chemical markers for cigarette use can provide objective indicators of smoking behavior. One widely used noninvasive marker is the level of saliva thiocyanate (SCN). In a Minneapolis school district, 1332 students in the eighth grade (ages 12–14) participated in a study [10] whereby they

(1) Viewed a film illustrating how recent cigarette use could be readily detected from small samples of saliva

(2) Provided a personal sample of saliva thiocyanate

(3) Provided a self-report on the number of cigarettes smoked per week

The results are given in Table 3.6.

TABLE 3.6 Relationship between saliva thiocyanate levels (SCN) and self-reported cigarettes smoked per week

Self-reported cigarettes smoked in last week	Number of students	Percent with SCN ≥ 100 μg/mL
None	1163	3.3
1–4	70	4.3
5–14	30	6.7
15–24	27	29.6
25–44	19	36.8
45+	23	65.2

Source: Reprinted with permission from the *American Journal of Public Health*, *71*(12), 1320, 1981.

Suppose the self-reports are completely accurate and are representative of the amount that eighth-grade students smoke in the general community. We are considering using an SCN level of $\geq$100 μg/mL as a test criterion for identifying cigarette smokers. Regard a student as positive if he or she smokes 1 or more cigarettes per week.

* **3.68** What is the sensitivity of the test for light-smoking students (i.e., students who smoke $\leq$14 cigarettes per week)?

* **3.69** What is the sensitivity of the test for moderate-smoking students (i.e., students who smoke 15–44 cigarettes per week)?

* **3.70** What is the sensitivity of the test for heavy-smoking students (i.e., students who smoke $\geq$45 cigarettes per week)?

* **3.71** What is the specificity of the test?

* **3.72** What is the predictive value positive of the test?

* **3.73** What is the predictive value negative of the test?

Suppose we regard the self-reports of all students who report some cigarette consumption as valid but estimate that 10% of students who report no cigarette consumption actually smoke 1–4 cigarettes per week and an additional 2% smoke 5–14 cigarettes per week.

* **3.74** If we assume that the percentage of students with SCN $\geq$100 μg/mL in these two subgroups is the same as in those who truly report 1–4 and 5–14 cigarettes per week, then what effect would this underreporting have on the predictive value positive of the test (i.e., would the true predictive value positive be the same, higher, or lower than that computed in 3.72)?

* **3.75** Compute the predictive value positive under these altered assumptions.

Hypertension

Laboratory measures of cardiovascular reactivity are receiving increasing attention. Much of the expanded interest is based on the belief that these measures, obtained under challenge from physical and psychological stressors, may yield a more biologically meaningful perspective of cardiovascular function than more traditional static measures. Typically, measurement of cardiovascular reactivity involves the use of an automated blood-pressure monitor to examine the changes in blood pressure before and after a stimulating experience (such as playing a video game). For this purpose, BP measurements were made with the Vita-Stat machine both before and after playing a video game. Similar measurements were obtained using manual methods for obtaining blood pressure. A person was classified as a "reactor" if his or her diastolic blood pressure (DBP) increased by 10 mm Hg or more after playing the game and as a nonreactor otherwise. The results are given in Table 3.7.

TABLE 3.7 Classification of cardiovascular reactivity using an automated and manual sphygmomanometer

	Δ DBP, manual	
Δ DBP, automated	<10	⩾10
<10	51	7
⩾10	15	6

3.76 If the manual measurements are regarded as the "true" measure of reactivity, then what is the sensitivity of automated BP measurements?

3.77 What is the specificity of automated BP measurements?

3.78 If the population tested is representative of the general population, then what are the predictive values positive and negative using this test?

Otolaryngology

The data set in Table 3.8 is based on 214 children with acute otitis media (OME) who participated in a randomized clinical trial [11]. Each child had OME at the beginning of the study in either one (unilateral cases) or both (bilateral cases) ears. Each child was randomly assigned to receive a 14-day course of one of two antibiotics, either cefaclor (CEF) or amoxicillin (AMO). The focus here is on the 203 children whose middle-ear status was determined at a 14-day follow-up visit. The data in Table 3.8 are presented in Data Set EAR.DAT (on the data disk).

3.79 Does there seem to be any difference in the effect of the antibiotics on clearance of otitis media? Try to express your results in terms of relative risk. Consider separate analyses for unilateral and bilateral cases. Also consider an analysis combining the two types of cases.

3.80 The investigators recorded the age of the children because they felt this might be an important factor in determining outcome. Were they right? Try to express your results in terms of relative risk.

3.81 While controlling for age, propose an analysis comparing the effectiveness of the two antibiotics. Express your results in terms of relative risk.

3.82 Another issue in this trial is the possible dependence between ears for the bilateral cases. Can you comment on this issue based on the data collected?

The concept of a **randomized clinical trial** is discussed more completely in Chapter 6. The analysis of **contingency-table data** is studied in Chapter 10, where many of the formal methods for analyzing this type of data are discussed.

Cardiovascular Disease

In Table 3.9 data on the relationship between various symptoms and disease states in patients suspected of having congenital heart disease are presented. In this table, for simplicity, only a subset of the symptoms (7) and disease states (7) are presented. In the original report [12], 50 symptoms and 33 disease states were considered. In the Data Set DISEASE.DAT, the prevalence of each of the disease states and the conditional probability of each of the symptoms given each of the disease states are presented. The documentation for this data set is given in the Data Set DISEASE.DOC. (All data sets are on the data disk.)

3.83 Write a computer program to compute the probability of each of the disease states given the presence or absence of any combination of the 50 symptoms. Note that some of the symptoms are mutually exclusive and thus cannot occur simultaneously; for example, symptom 1 = age 1 month to 1 year and symptom 2 = age 1–20 years. Also, some of the symptoms have to be considered as a group. Read the original report for details concerning these points.

3.84 Test your program using some of the examples given in the article.

TABLE 3.8 Format for EAR.DAT

Column	Variable	Format or code
1–3	ID	
5	Clearance by 14 days	1 = yes/0 = no
7	Antibiotic	1 = CEF/2 = AMO
9	Age	1 = <2 yrs/2 = 2–5 yrs/3 = 6+ yrs
11	Ear	1 = 1st ear/2 = 2nd ear

TABLE 3.9 Prevalence of symptoms and diagnoses for patients suspected of having congenital heart disease

		Symptoms						
Diagnosis	**Prevalence**	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y_1	.155	.49	.50	.01	.10	.05	.05	.01
Y_2	.126	.50	.50	.02	.50	.02	.40	.70
Y_3	.084	.55	.05	.25	.90	.05	.10	.95
Y_4	.020	.45	.45	.01	.95	.10	.10	.95
Y_5	.098	.10	.00	.20	.70	.01	.05	.40
Y_6	.391	.70	.15	.01	.30	.01	.15	.30
Y_7	.126	.60	.10	.30	.70	.10	.20	.70

Note: Y_1 = normal
Y_2 = atrial septal defect without pulmonary stenosis or pulmonary hypertension[a]
Y_3 = ventricular septal defect with valvular pulmonary stenosis
Y_4 = isolated pulmonary hypertension[a]
Y_5 = transposed great vessels
Y_6 = ventricular septal defect without pulmonary hypertension[a]
Y_7 = ventricular septal defect with pulmonary hypertension[a]
X_1 = age 1–20 years old
X_2 = age > 20 years old
X_3 = mild cyanosis
X_4 = easy fatigue
X_5 = chest pain
X_6 = repeated respiratory infections
X_7 = EKG axis more than 110°

[a]Pulmonary hypertension is defined as pulmonary artery pressure ≥ systemic arterial pressure.
Source: Reprinted with permission of *The American Medical Association* from *The Journal of the American Medical Association, 177*(3), 177–183, 1961. Copyright 1961, American Medical Association.

Gynecology

A drug company is developing a new pregnancy-test kit for use on an outpatient basis. The company uses the pregnancy test on 100 women who are known to be pregnant, of whom 95 are positive using the test. The company uses the pregnancy test on 100 other women who are known to *not* be pregnant, of whom 99 are negative using the test.

* **3.85** What is the sensitivity of the test?

* **3.86** What is the specificity of the test?

The company anticipates that of the women who will use the pregnancy-test kit, 10% will actually be pregnant.

* **3.87** What is the predictive value positive of the test?

* **3.88** Suppose the "cost" of a false negative (2c) is twice that of a false positive (c) (since for a false negative prenatal care would be delayed during the first trimester of pregnancy). If the standard home pregnancy-test kit (made by another drug company) has a sensitivity of 0.98 and a specificity of .98, then which test (the new or standard) will cost least per woman using it in the general population and by how much?

Mental Health

The Chinese Mini-Mental Status Test (CMMS) is a test consisting of 114 items intended to identify people with Alzheimer's disease and senile dementia among people in China [13]. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and nurses and a definitive diagnosis of dementia was made. Table 3.10 shows the results obtained on the subgroup of people with at least some formal education.

Suppose a cutoff value of ≤20 on the test is used to identify people with dementia.

3.89 What is the sensitivity of the test?

3.90 What is the specificity of the test?

TABLE 3.10 Relationship of clinical dementia to outcome on the Chinese Mini-Mental Status Test

CMMS score	Nondemented	Demented
0–5	0	2
6–10	0	1
11–15	3	4
16–20	9	5
21–25	16	3
26–30	18	1
	46	16

Demography

A study based on data collected from the Medical Birth Registry of Norway looked at fertility rates according to survival outcomes of previous births [14]. The data are presented in Table 3.11.

3.91 What is the probability of having a livebirth (L) at a second birth given that the outcome of the first pregnancy was a stillbirth (D), that is, death?

3.92 Answer Problem 3.91 if the outcome of the first pregnancy was a livebirth.

3.93 What is the probability of 0, 1, and 2+ additional pregnancies if the first birth was a stillbirth?

3.94 Answer Problem 3.93 if the first birth was a livebirth.

TABLE 3.11 Relationship of fertility rates to survival outcome of previous births in Norway

	First birth	Continuing to second birth	Second birth outcome		Continuing to third birth	Third birth outcome	
Perinatal outcome	*n*	*n*		*n*	*n*		*n*
D	7,022	5,924	D	368	277	D	39
						L	238
			L	5,556	3,916	D	115
						L	3,801
L	350,693	265,701	D	3,188	2,444	D	140
						L	2,304
			L	262,513	79,450	D	1,005
						L	78,445

Note: D = dead, L = alive at birth and for at least one week.

References

[1] National Center for Health Statistics (1976, February 13). *Monthly vital statistics report, advance report, final natality statistics (1974)*, *24*(11) (Suppl. 2).

[2] Doll, R., Muir, C., & Waterhouse, J. (Eds.). (1970). *Cancer incidence in five continents II*. Berlin: Springer-Verlag.

[3] Feller, W. (1960). *An introduction to probability theory and its applications*. New York: Wiley.

[4] Podgor, M. J., Leske, M. C., & Ederer, F. (1983). Incidence estimates for lens changes, macular changes, open-angle glaucoma, and diabetic retinopathy. *American Journal of Epidemiology*, *118*(2), 206–212.

[5] National Center for Health Statistics (1976, November 8). *Advance data from vital and health statistics, 2*.

[6] Pfeffer, R. I., Afifi, A. A., & Chance, J. M. (1987). Prevalence of Alzheimer's Disease in a retirement community. *American Journal of Epidemiology*, *125*(3), 420–436.

[7] National Center for Health Statistics (1975, January 30). *Monthly vital statistics report, final natality statistics (1973)*, *23*(11) (Suppl.).

[8] Colley, J. R. T., Holland, W. W., & Corkhill, R. T. (1974). Influence of passive smoking and parental phlegm on pneumonia and bronchitis in early childhood. *Lancet*, *II*, 1031.

[9] Garvey, A. J., Bossé, R., Glynn, R. J., & Rosner, B. (1983). Smoking cessation in a prospective study of healthy adult males: Effects of age, time period, and amount smoked. *American Journal of Public Health*, *73*(4), 446–450.

[10] Luepker, R. V., Pechacek, T. F., Murray, D. M., Johnson, C. A., Hund, F., & Jacobs, D. R. (1981). Saliva thiocyanate: A chemical indicator of cigarette smoking in adolescents. *American Journal of Public Health*, *71*(12), 1320.

[11] Mandel, E., Bluestone, C. D., Rockette, H. E., Blatter, M. M., Reisinger, K. S., Wucher, F. P., & Harper, J. (1982). Duration of effusion after antibiotic treatment for acute otitis media: Comparison of cefaclor and amoxicillin. *Pediatric Infectious Diseases*, *1*, 310–316.

[12] Warner, H., Toronto, A., Veasey, L. G., & Stephenson, R. (1961). A mathematical approach to medical diagnosis. *JAMA*, *177*(3), 177–183.

[13] Katzman, R., Zhang, M. Y., Ouang-Ya-Qu, Wang, Z. Y., Liu, W. T., Yu, E., Wong, S. C., Salmon, D. P., & Grant, I. (1988). A Chinese version of the Mini-Mental State Examination; impact of illiteracy in a Shanghai dementia survey. *Journal of Clinical Epidemiology*, *41*(10), 971–978.

[14] Skjaerven, R., Wilcox, A. J., Lie, R. T., & Irgens, L. M. (1988). Selective fertility and the distortion of perinatal mortality. *American Journal of Epidemiology*, *128*(6), 1352–1363.