Chapter 3

Probability

3.2 Definition of Probability

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	.51247
1965–1969	9,219,202	17,989,361	.51248
1965–1974	17,857,857	34,832,051	.51268

The probability of live male births in 1965 was 0.51247 ; in 1965 – 69 was 0.51248; *in* 1965 – 74 *was* 0.51268.

- The **sample space** is the set of all possible outcomes.
- An **event** is any set of outcomes of interest.
- The **probability** of an event is the relative frequency of this set of outcomes over an indefinitely large (or infinite) number of trials.

- The probability of an event E, denoted by $Pr(E)$, always satisfies
  $0 \leq Pr(E) \leq 1$
- 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 (DBP < 90), and let B be the event that a person has borderline DBP readings ($90 \leq$ DBP < 95). Suppose that $Pr(A) = .7$, and $Pr(B) = .1$. Let Z be the event that a person has a DBP < 95. Then

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

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

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

EXAMPLE 3.7

Hypertension Let X be DBP, C be the event $X \geq 90$, and D be the event $75 \leq X \leq 100$. Events C and D are *not* mutually exclusive, because they both occur when $90 \leq X \leq 100$.

3.3 Some Useful Probabilistic Notations

- The symbol { } is used as shorthand for the phrase "the event."
- $A \cup B$ is the event that either A or B occurs, or they both occur.

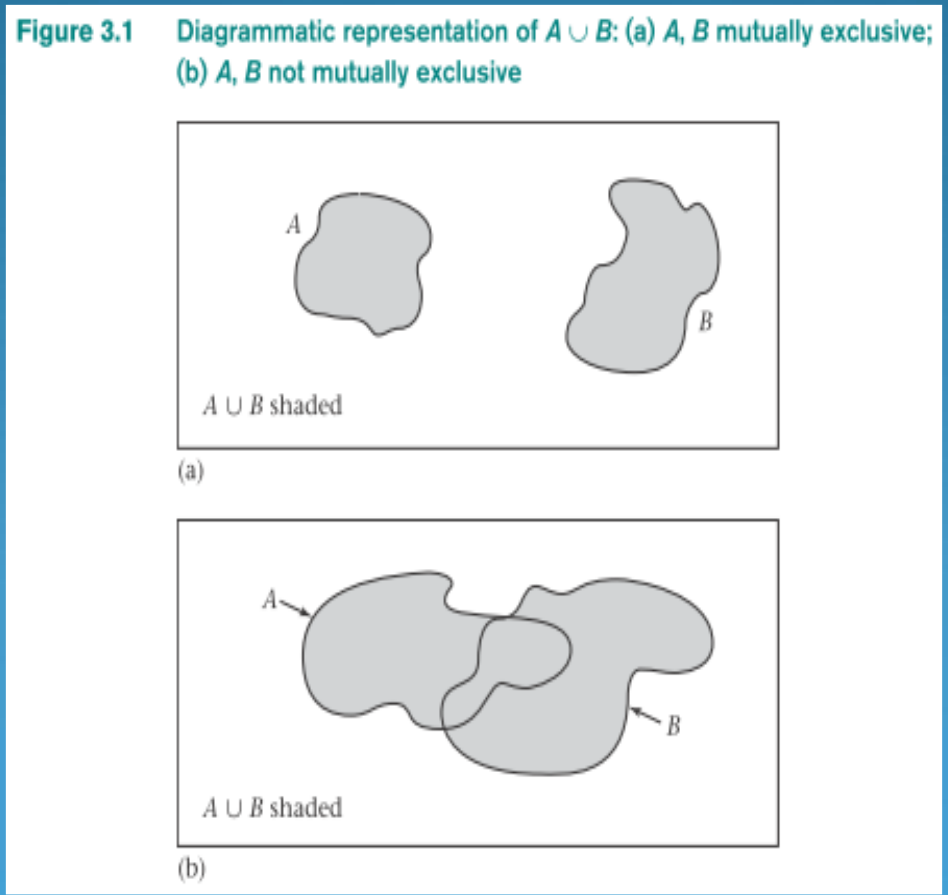
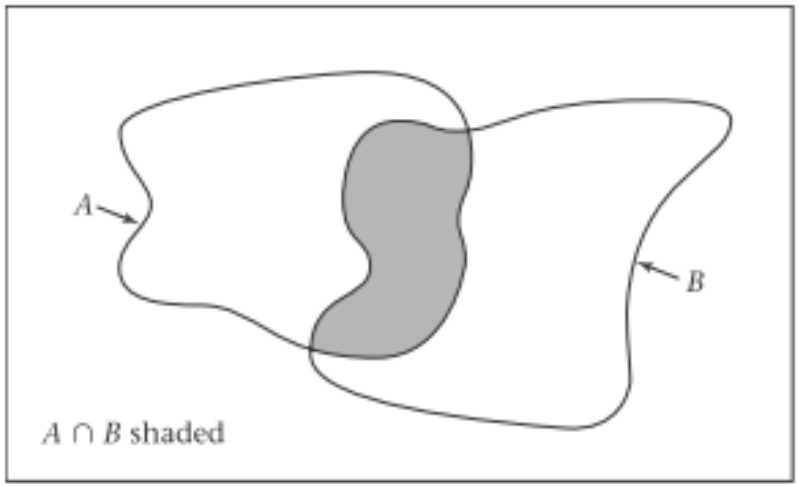


Figure 3.1 Diagrammatic representation of $A \cup B$: (a) A, B mutually exclusive; (b) A, B not mutually exclusive

$A \cap B$ is the event that both A and B occur simultaneously.

Figure 3.2 Diagrammatic representation of $A \cap B$



EXAMPLE 3.8 **Hypertension** Let events A and B be defined as in Example 3.6: $A = \{X < 90\}$, $B = \{90 \le X < 95\}$, where $X = \text{DBP}$. Then $A \cup B = \{X < 95\}$.

EXAMPLE 3.9 **Hypertension** Let events C and D be defined as in Example 3.7:

$$C = \{X \ge 90\} \quad D = \{75 \le X \le 100\}$$

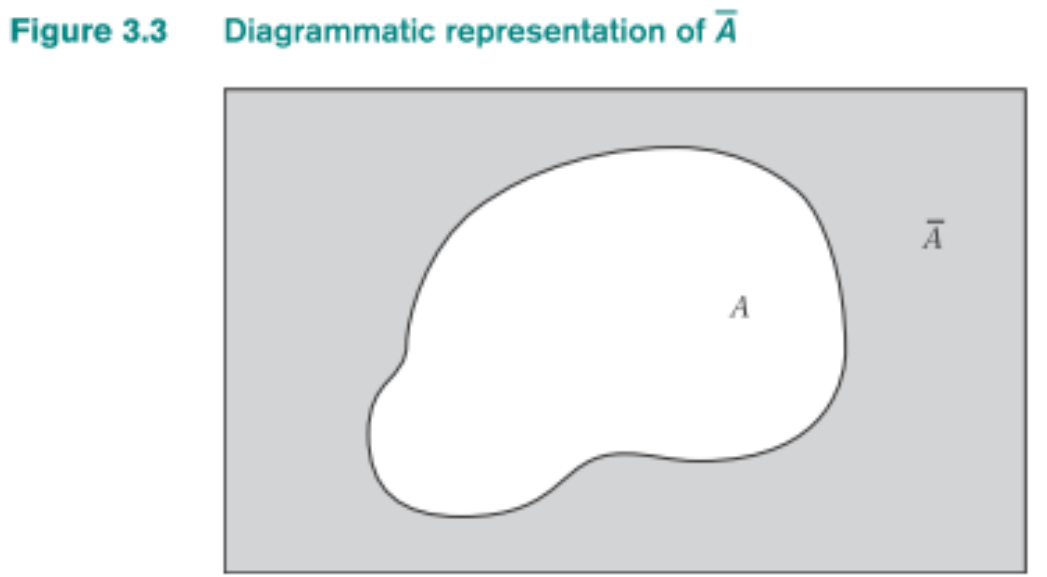
Then $C \cup D = \{X \ge 75\}$

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

$$C = \{X \ge 90\} \quad D = \{75 \le X \le 100\}$$

Then $C \cap D = \{90 \le X \le 100\}$

$\overline{A}$ is the event that A does not occur. It is called the complement of A. Notice that $Pr(\overline{A}) = 1 - Pr(A)$, because $\overline{A}$ occurs only when A does not occur.



Example:
If A and C are two events. $A = \{ X < 90\}$; $C = \{ X \geq 90\}$.
Then, $C = \overline{A}$, because C can only occur when A does not occur.

3.4 Multiplication Law of Probability

Events A and B are called independent events if
$Pr(A \cap B) = Pr(A) \times Pr(B)$

If $Pr(A) = 0.1$ and $Pr(B) = 0.2$
Probability that both mother and father are hypertensive,

$Pr(A \cap B) = Pr(A) \times Pr(B)$
$= 0.1(0.2) = .02$

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 events $A = \{\text{mother's DBP} \geq 90\}$, $B = \{\text{father's DBP} \geq 90\}$. These events are diagrammed in Figure 3.4.

Suppose we know that $Pr(A) = .1$, $Pr(B) = .2$. What can we say about $Pr(A \cap B) = Pr(\text{mother's DBP} \geq 90 \text{ and father's DBP} \geq 90) = Pr(\text{both mother and father are hypertensive})$? We can say nothing unless we are willing to make certain assumptions.

Two events A, B are dependent if $Pr(A \cap B) \neq Pr(A) \times Pr(B)$

EXAMPLE 3.14

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

$$A = \{\text{mother's DBP} \geq 90\} \quad B = \{\text{first-born child's DBP} \geq 80\}$$

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

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

and the events A, B would be dependent.

EXAMPLE 3.15

Sexually Transmitted Disease Suppose two doctors, A and B, test all patients coming into a clinic for syphilis. Let events $A^+ = \{\text{doctor A makes a positive diagnosis}\}$ and $B^+ = \{\text{doctor B makes a positive diagnosis}\}$. Suppose 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

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

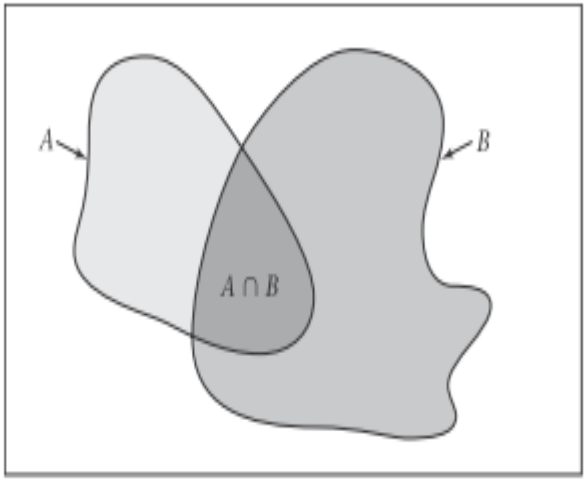
Thus, $Pr(A^+ \cap B^+) = .08 > Pr(A^+) \times Pr(B^+) = .1(.17) = .017$

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

3.5 Addition Law of Probability

If A and B are any events, then $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Figure 3.5 Diagrammatic representation of the addition law of probability



Special case: If events A and B are mutually exclusive, then $Pr(A \cap B) = 0$ And the addition law reduces to $Pr(A \cup B) = Pr(A) + Pr(B)$.

EXAMPLE 3.16

Sexually Transmitted Disease Consider the data 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.

EXAMPLE 3.17

Hypertension Look at Example 3.12, where

$$A = \{\text{mother's DBP} \geq 90\} \quad \text{and} \quad B = \{\text{father's DBP} \geq 90\}$$

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

3.6 Conditional Probability

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

- If A and B are independent events, then $\Pr(B|A) = \Pr(B) = \Pr(B|\bar{A})$
- If two events A, B are dependent, then $\Pr(B|A) \neq \Pr(B) \neq \Pr(B|\bar{A})$ and $\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$

Def.: The **relative risk (RR)** of B given A is $\Pr(B|A)/\Pr(B|\bar{A})$

- If A and B are independent, then RR is 1.
- If A and B are dependent, then RR is different from 1.
- The more the dependence between events increases, the further the RR will be from 1.

Relative risk (RR) compares the probability of a health outcome in an exposed group with the probability of the same outcome in an unexposed group. It is commonly used in health research to compare the risk of developing a disease between two groups.

EXAMPLE 3.19

Cancer Suppose that among 100,000 women with negative mammograms 20 will be diagnosed with breast cancer within 2 years, or $Pr(B|\bar{A}) = 20/10^5 = .0002$, whereas 1 woman in 10 with positive mammograms will be diagnosed with breast cancer within 2 years, or $Pr(B|A) = .1$. The two events A and B would be highly dependent, because

$$RR = Pr\left(B|A\right)\Big/Pr\left(B|\bar{A}\right) = .1/.0002 = 500$$

In other words, women with positive mammograms are 500 times more likely to develop breast cancer over the next 2 years than are women with negative mammograms. This is the rationale for using the mammogram as a screening test for breast cancer. If events A and B were independent, then the RR would be 1; women with positive or negative mammograms would be equally likely to have breast cancer, and the mammogram would not be useful as a screening test for breast cancer.

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 *RR* 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 diagnoses 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 confirm the diagnosis. Thus,

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

because 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 *RR* 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.

Total-Probability Rule

For any events A and B, $\Pr(B) = \Pr(B|A) \times \Pr(A) + \Pr(B|\bar{A}) \times \Pr(\bar{A})$

If event B occurs, it must occur either with A or without A.

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

From definition of conditional probability,

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

On substitution,

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

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.

EXAMPLE 3.21

Cancer Let A and B be defined as in Example 3.19, and suppose that 7% of the general population of women will have a positive mammogram. What is the probability of developing breast cancer over the next 2 years among women in the general population?

Solution:

$$\begin{aligned} Pr(B) &= Pr(\text{breast cancer}) \\ &= Pr(\text{breast cancer} \mid \text{mammogram}^+) \times Pr(\text{mammogram}^+) \\ &+ Pr(\text{breast cancer} \mid \text{mammogram}^-) \times Pr(\text{mammogram}^-) \\ &= .1(.07) + .0002(.93) = .00719 = 719/10^5 \end{aligned}$$

Thus, the unconditional probability of developing breast cancer over the next 2 years in the general population ($719/10^5$) is a weighted average of the conditional probability of developing breast cancer over the next 2 years among women with a positive mammogram (.1) and the conditional probability of developing breast cancer over the next 2 years among women with a negative mammogram ($20/10^5$), with weights of 0.07 and 0.93 corresponding to mammogram$^+$ and mammogram$^-$ women, respectively.

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 is 60–64 years of age, 28% are 65–69 years of age, 20% are 70–74 years of age, and 7% are 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 these respective age groups will develop cataract over the next 5 years [4]. What percentage of the population in our study will develop cataract over the next 5 years, and how many people with cataract does this percentage represent?

Solution: Let $A_1 = \{\text{ages 60–64}\}$, $A_2 = \{\text{ages 65–69}\}$, $A_3 = \{\text{ages 70–74}\}$, $A_4 = \{\text{ages 75+}\}$. These events are mutually exclusive and exhaustive because each person in our population must be in one and only one age group. 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 $Pr(B|A_4) = .153$, where $B = \{\text{develop cataract in the next 5 years}\}$. Finally, using the total-probability rule,

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

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

3.7 Baye's Rule and Screening Tests

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

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

EXAMPLE 3.23

Cancer Find PV^+ and PV^- for mammography given the data in Example 3.19.

Solution: We see that $PV^+ = Pr(\text{breast cancer} \mid \text{mammogram}^+) = .1$

whereas $PV^- = Pr(\text{breast cancer}^- \mid \text{mammogram}^-)$

$$= 1 - Pr(\text{breast cancer} \mid \text{mammogram}^-) = 1 - .0002 = .9998$$

Thus, if the mammogram is negative, the woman is virtually certain *not* to develop breast cancer over the next 2 years ($PV^- \approx 1$); whereas if the mammogram is positive, the woman has a 10% chance of developing breast cancer ($PV^+ = .10$).

- A symptom or set of symptoms can be regarded as a screening test for a disease. Higher the PV of the symptoms, the more valuable the test will be.
- Clinicians often cannot directly measure the PV of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people.
- **Sensitivity** of a symptom (or a set of symptoms): the probability that the symptom is present given that the person has a disease.
- **Specificity** of a symptom (or a set of symptoms): the probability that the symptom is not present given that the person does not have a disease.
- **False negative**: negative test result when the disease or condition being tested for is actually present.
- **False positive**: positive test result when the disease or condition being tested for is not actually present.

EXAMPLE 3.24

Cancer Suppose the disease is lung cancer and the symptom is cigarette smoking. If we assume that 90% of people with lung cancer and 30% of people without lung cancer (essentially the entire general population) are smokers, then the sensitivity and specificity of smoking as a screening test for lung cancer are .9 and .7, respectively. Obviously, cigarette smoking cannot be used by itself as a screening criterion for predicting lung cancer because there will be too many false positives (people without cancer who are smokers).

EXAMPLE 3.25

Cancer Suppose the disease is breast cancer in women and the symptom is having a family history of breast cancer (either a mother or a sister with breast cancer). If we assume 5% of women with breast cancer have a family history of breast cancer but only 2% of women without breast cancer have such a history, then the sensitivity of a family history of breast cancer as a predictor of breast cancer 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 (women with breast cancer who do not have a family history).

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

Predictive value negative $= PV^{-} = Pr\left(\bar{B}|\bar{A}\right)$

Sensitivity $= Pr\left(A|B\right)$

Specificity $= Pr\left(\bar{A}|\bar{B}\right)$

EXAMPLE 3.26

Hypertension Suppose 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What are the PV^+ and PV^- of the machine, assuming 20% of the adult population is hypertensive?

Solution: The sensitivity = .84 and specificity = 1 – .23 = .77. Thus, from Bayes' rule it follows that

$$PV^+ = (.84)(.2)/[(.84)(.2)+(.23)(.8)]$$
$$= .168/.352 = .48$$

Similarly, $PV^- = (.77)(.8)/[(.77)(.8)+(.16)(.2)]$
$=.616/.648 = .95$

Thus, a negative result from the machine is reasonably predictive because we are 95% sure a person with a negative result from the machine is normotensive. However, a positive result is not very predictive because we are only 48% sure a person with a positive result from the machine is hypertensive.

EXAMPLE 3.27

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

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

$$\text{Disease state} \begin{cases} B_1 = \text{normal} \\ B_2 = \text{lung cancer} \\ B_3 = \text{sarcoidosis} \end{cases}$$

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 men

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

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-, gender-, and smoking-specific prevalence rates for the diseases in question. The interesting question now becomes 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,

$$Pr(B_1|A) = Pr(A|B_1) \times Pr(B_1) \Big/ \left[\sum_{j=1}^{3} Pr(A|B_j) \times Pr(B_j)\right]$$

$$= .001(.99)/[.001(.99) + .9(.001) + .9(.009)]$$

$$= .00099 / .00999 = .099$$

$$Pr(B_2|A) = .9(.001)/[.001(.99) + .9(.001) + .9(.009)]$$

$$= .00090 / .00999 = .090$$

$$Pr(B_3|A) = .9(.009)/[.001(.99) + .9(.001) + .9(.009)]$$

$$= .00810 / .00999 = .811$$

Thus, although the unconditional probability of sarcoidosis is very low (.009), the conditional probability of the disease given these symptoms and this age-gender-smoking group is .811. Also, although the symptoms and diagnostic tests are consistent with both lung cancer and sarcoidosis, the latter is much more likely among patients in this age-gender-smoking group (i.e., among never- smoking men).

EXAMPLE 3.28

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

Solution:

$$Pr(B_1|A) = .001(.98)/[.001(.98) + .9(.015) + .9(.005)]$$

$$= .00098 / .01898 = .052$$

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

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

Thus, in this type of patient (i.e., a heavy-smoking man) lung cancer is the most likely diagnosis.

Key Terms

Probability	A numerical measure of how likely an event is to occur.
Sample space	The set of all possible outcomes.
Event	A set of outcomes of interest.
Complement	The event that the original event does not occur.
Independent events	Events for which the probability of both equals the product of their individual probabilities.
Dependent events	Events for which the multiplication rule for independence does not hold.
Mutually exclusive events	Events that cannot occur at the same time.
Union (A ∪ B)	The event that A or B or both occur.
Intersection (A ∩ B)	The event that A and B both occur.
Conditional probability	The probability of an event given that another event has occurred; the chapter summary mentions this concept.
Sensitivity	Probability of a positive finding among people who have the disease.
Specificity	Probability of a negative finding among people who do not have the disease.
Positive predictive value (PV+)	Probability of disease given a positive test result.
Negative predictive value (PV–)	Probability of no disease given a negative test result.
False positive	A positive result in a person who does not have the disease.
False negative	A negative result in a person who has the disease.
Bayes' rule	A rule used to calculate conditional probabilities, including predictive values for screening tests.
Relative risk (RR)	The ratio of risk in an exposed group to risk in an unexposed group.

Quick Review

1. What is the difference between a sample space and an event?
2. What are the possible values of a probability?
3. What does $A \cup B$ mean? What does $A \cap B$ mean?
4. If $Pr(A) = 0.35$, calculate $Pr(A^c)$.
5. If A and B are independent, with $Pr(A) = 0.4$ and $Pr(B) = 0.5$, calculate $Pr(A \cap B)$.
6. If $Pr(A) = 0.30$, $Pr(B) = 0.40$, and $Pr(A \cap B) = 0.10$, calculate $Pr(A \cup B)$.
7. How does the addition rule simplify when A and B are mutually exclusive?
8. Explain the difference between sensitivity and positive predictive value.
9. What is the difference between a false positive and a false negative?
10. What does the notation $Pr(\text{disease} \mid \text{test positive})$ mean?
11. In an exposed group, 12 of 60 people develop a disease. In an unexposed group, 6 of 60 develop it. Calculate and interpret the relative risk.