CHAPTER 4

Probability and Probability Distributions

4.1 Introduction and Abstract of Research Study

We stated in Chapter 1 that a scientist uses inferential statistics to make statements about a population based on information contained in a sample of units selected from that population. Graphical and numerical descriptive techniques were presented in Chapter 3 as a means to summarize and describe a sample.

However, a sample is not identical to the population from which it was selected. We need to assess the degree of accuracy to which the sample mean, sample standard deviation, or sample proportion represent the corresponding population values.

Most management decisions must be made in the presence of uncertainty. Prices and designs for new automobiles must be selected on the basis of shaky forecasts of consumer preference, national economic trends, and competitive actions. The size and allocation of a hospital staff must be decided with limited information on patient load. The inventory of a product must be set in the face of uncertainty about demand. Probability is the language of uncertainty. Now let us examine probability, the mechanism for making inferences. This idea is probably best illustrated by an example.

Newsweek, in its June 20, 1998, issue, asks the question, "Who Needs Doctors? The Boom in Home Testing." The article discusses the dramatic increase in medical screening tests for home use. The home-testing market has expanded beyond the two most frequently used tests, pregnancy and diabetes glucose monitoring, to a variety of diagnostic tests that were previously used only by doctors and certified laboratories. There is a DNA test to determine whether twins are fraternal or identical, a test to check cholesterol level, a screening test for colon cancer, and tests to determine whether your teenager is a drug user. However, the major question that needs to be addressed is, How reliable are the testing kits? When a test indicates that a woman is not pregnant, what is the chance that the test is incorrect and the woman is truly pregnant? This type of incorrect result from a home test could translate into a woman not seeking the proper prenatal care in the early stages of her pregnancy.

Suppose a company states in its promotional materials that its pregnancy test provides correct results in 75% of its applications by pregnant women. We want to evaluate the claim, and so we select 20 women who have been determined by their physicians, using the best possible testing procedures, to be pregnant. The test is taken by each of the 20 women, and for all 20 women the test result is negative, indicating that none of the 20 is pregnant. What do you conclude about the company's claim on the reliability of its test? Suppose you are further assured that each of the 20 women was in fact pregnant, as was determined several months after the test was taken.

If the company's claim of 75% reliability was correct, we would have expected somewhere near 75% of the tests in the sample to be positive. However, none of the test results was positive. Thus, we would conclude that the company's claim is probably false. Why did we fail to state with certainty that the company's claim was false? Consider the possible setting. Suppose we have a large population consisting of millions of units, and 75% of the units are Ps for positives and 25% of the units are Ns for negatives. We randomly select 20 units from the population and count the number of units in the sample that are Ps. Is it possible to obtain a sample consisting of 0 Ps and 20 Ns? Yes, it is possible, *but* it is highly *improbable*. Later in this chapter we will compute the probability of such a sample occurrence.

To obtain a better view of the role that probability plays in making inferences from sample results to conclusions about populations, suppose the 20 tests result in 14 tests being positive—that is, a 70% correct response rate. Would you consider this result highly improbable and reject the company's claim of a 75% correct response rate? How about 12 positives and 8 negatives, or 16 positives and 4 negatives? At what point do we decide that the result of the observed sample is

so improbable, assuming the company's claim is correct, that we disagree with its claim? To answer this question, we must know how to find the probability of obtaining a particular sample outcome. Knowing this probability, we can then determine whether we agree or disagree with the company's claim. Probability is the tool that enables us to make an inference. Later in this chapter we will discuss in detail how the FDA and private companies determine the reliability of screening tests.

Because probability is the tool for making inferences, we need to define probability. In the preceding discussion, we used the term *probability* in its everyday sense. Let us examine this idea more closely.

Observations of phenomena can result in many different outcomes, some of which are more likely than others. Numerous attempts have been made to give a precise definition for the probability of an outcome. We will cite three of these.

classical interpretation

The first interpretation of probability, called the **classical interpretation of probability,** arose from games of chance. Typical probability statements of this type are, for example, "the probability that a flip of a balanced coin will show 'heads' is 1/2" and "the probability of drawing an ace when a single card is drawn from a standard deck of 52 cards is 4/52." The numerical values for these probabilities arise from the nature of the games. A coin flip has two possible outcomes (a head or a tail); the probability of a head should then be 1/2 (1 out of 2). Similarly, there are 4 aces in a standard deck of 52 cards, so the probability of drawing an ace in a single draw is 4/52, or 4 out of 52.

outcome
event

In the classical interpretation of probability, each possible distinct result is called an **outcome;** an **event** is identified as a collection of outcomes. The probability of an event E under the classical interpretation of probability is computed by taking the ratio of the number of outcomes, N_e, favorable to event E to the total number N of possible outcomes:

$$P(\text{even } E) = \frac{N_e}{N}$$

The applicability of this interpretation depends on the assumption that all outcomes are equally likely. If this assumption does not hold, the probabilities indicated by the classical interpretation of probability will be in error.

relative frequency interpretation

A second interpretation of probability is called the **relative frequency concept of probability;** this is an empirical approach to probability. If an experiment is repeated a large number of times and event E occurs 30% of the time, then .30 should be a very good approximation to the probability of event E. Symbolically, if an experiment is conducted n different times and if event E occurs on n_e of these trials, then the probability of event E is approximately

$$P(\text{even } E) \cong \frac{n_e}{n}$$

We say "approximate" because we think of the actual probability P(event E) as the relative frequency of the occurrence of event E over a very large number of observations or repetitions of the phenomenon. The fact that we can check probabilities that have a relative frequency interpretation (by simulating many repetitions of the experiment) makes this interpretation very appealing and practical.

The third interpretation of probability can be used for problems in which it is difficult to imagine a repetition of an experiment. These are "one-shot" situations. For example, the director of a state welfare agency who estimates the probability that

subjective interpretation

a proposed revision in eligibility rules will be passed by the state legislature would not be thinking in terms of a long series of trials. Rather, the director would use a **personal** or **subjective probability** to make a one-shot statement of belief regarding the likelihood of passage of the proposed legislative revision. The problem with subjective probabilities is that they can vary from person to person and they cannot be checked.

Of the three interpretations presented, the relative frequency concept seems to be the most reasonable one because it provides a practical interpretation of the probability for most events of interest. Even though we will never run the necessary repetitions of the experiment to determine the exact probability of an event, the fact that we can check the probability of an event gives meaning to the relative frequency concept. Throughout the remainder of this text we will lean heavily on this interpretation of probability.

Abstract of Research Study: Inferences about Performance-Enhancing Drugs among Athletes

The *Associated Press* reported the following in an April 28, 2005, article:

> CHICAGO—The NBA and its players union are discussing expanded testing for performance-enhancing drugs, and commissioner David Stern said Wednesday he is optimistic it will be part of the new labor agreement. The league already tests for recreational drugs and more than a dozen types of steroids. But with steroid use by professional athletes and the impact they have on children under increasing scrutiny, Stern said he believes the NBA should do more.

An article in *USA Today* (April 27, 2005) by Dick Patrick reports,

> Just before the House Committee on Government Reform hearing on steroids and the NFL ended Wednesday, ranking minority member Henry Waxman, D-Calif., expressed his ambiguity about the effectiveness of the NFL testing system. He spoke to a witness panel that included NFL Commissioner Paul Tagliabue and NFL Players Association executive director Gene Upshaw, both of whom had praised the NFL system and indicated there was no performance-enhancing drug problem in the league. "There's still one thing that puzzles me," Waxman said, "and that's the fact that there are a lot of people who are very credible in sports who tell me privately that there's a high amount of steroid use in football. When I look at the testing results, it doesn't appear that's the case. It's still nagging at me."

Finally, we have a report from ABC News (April 27, 2005) in which the drug issue in major league sports is discussed:

> A law setting uniform drug-testing rules for major U.S. sports would be a mistake, National Football League Commissioner Paul Tagliabue said Wednesday under questioning from House lawmakers skeptical that professional leagues are doing enough. "We don't feel that there is rampant cheating in our sport," Tagliabue told the House Government Reform Committee. Committee members were far less adversarial than they were last month, when Mark McGwire, Jose Canseco and other current and former baseball stars were compelled to appear and faced tough questions about steroid use. Baseball commissioner Bud Selig, who also appeared at that hearing, was roundly criticized for the punishments in his sport's policy, which lawmakers said was too lenient.

One of the major reasons the union leaders of professional sports athletes are so concerned about drug testing is that failing a drug test can devastate an athlete's career. The controversy over performance-enhancing drugs has seriously brought into question the reliability of the tests for these drugs. Some banned substances,

such as stimulants like cocaine and artificial steroids, are relatively easy to deal with because they are not found naturally in the body. If these are detected at all, the athlete is banned. Nandrolone, a close chemical cousin of testosterone, was thought to be in this category until recently. But a study has since shown that normal people can have a small but significant level in their bodies—0.6 nanograms per milliliter of urine. The International Olympic Committee has set a limit of 2 nanograms per milliliter. But expert Mike Wheeler, a doctor at St Thomas' Hospital, states that this is "awfully close" to the level at which an unacceptable number (usually more than .01%) of innocent athletes might produce positive tests.

The article, "Inferences about testosterone abuse among athletes," in a 2004 issue of *Chance* (vol. 17, pp. 5–8), discusses some of the issues involved with the drug testing of athletes. In particular, they discuss the issues involved in determining the reliability of drug tests. The article reports, "The diagnostic accuracy of any laboratory test is defined as the ability to discriminate between two types of individuals—in this case, users and nonusers. *Specificity* and *sensitivity* characterize diagnostic tests. . . . Estimating these proportions requires collecting and tabulating data from the two reference samples, users and nonusers, . . . Bayes' rule is a necessary tool for relating experimental evidence to conclusions, such as whether someone has a disease or has used a particular substance. Applying Bayes' rule requires determining the test's sensitivity and specificity. It also requires a pre-test (or prior) probability that the athlete has used a banned substance."

Any drug test can result in a false positive due to the variability in the testing procedure, biologic variability, or inadequate handling of the material to be tested. Even if a test is highly reliable and produces only 1% false positives but the test is widely used, with 80,000 tests run annually, the result would be that 800 athletes would be falsely identified as using a banned substance. The result is that innocent people will be punished. The trade-off between determining that an athlete is a drug user and convincing the public that the sport is being conducted fairly is not obvious. The authors' state, "Drug testing of athletes has two purposes: to prevent artificial performance enhancement (known as doping) and to discourage the use of potentially harmful substances." Thus, there is a need to be able to assess the reliability of any testing procedure.

In this chapter, we will explicitly define the terms *specificity, sensitivity,* and *prior probability*. We will then formulate *Bayes' rule* (which we will designate as Bayes' Formula). At the end of the chapter, we will return to this article and discuss the issues of *false positives* and *false negatives* in drug testing and how they are computed from our knowledge of the specificity and sensitivity of a drug test along with the prior probability that a person is a user.

4.2 Finding the Probability of an Event

In the preceding section, we discussed three different interpretations of probability. In this section, we will use the classical interpretation and the relative frequency concept to illustrate the computation of the probability of an outcome or event. Consider an experiment that consists of tossing two coins, a penny and then a dime, and observing the upturned faces. There are four possible outcomes:

TT: tails for both coins
TH: a tail for the penny, a head for the dime
HT: a head for the penny, a tail for the dime
HH: heads for both coins

What is the probability of observing the event exactly one head from the two coins?

This probability can be obtained easily if we can assume that all four outcomes are equally likely. In this case, that seems quite reasonable. There are $N = 4$ possible outcomes, and $N_e = 2$ of these are favorable for the event of interest, observing exactly one head. Hence, by the classical interpretation of probability,

$$P(\text{exactly 1 head}) = \frac{2}{4} = \frac{1}{2}$$

Because the event of interest has a relative frequency interpretation, we could also obtain this same result empirically, using the relative frequency concept. To demonstrate how relative frequency can be used to obtain the probability of an event, we will use the ideas of simulation. Simulation is a technique that produces outcomes having the same probability of occurrence as the real situation events. The computer is a convenient tool for generating these outcomes. Suppose we wanted to simulate 1,000 tosses of the two coins. We can use a computer program such as SAS or Minitab to simulate the tossing of a pair of coins. The program has a random number generator. We will designate an even number as *H* and an odd number as *T*. Since there are five even and five odd single-digit numbers, the probability of obtaining an even number is 5/10 = .5, which is the same as the probability of obtaining an odd number. Thus, we can request 500 pairs of single-digit numbers. This set of 500 pairs of numbers will represent 500 tosses of the two coins, with the first digit representing the outcome of tossing the penny and the second digit representing the outcome of tossing the dime. For example, the pair (3, 6) would represent a tail for the penny and a head for the dime. Using version 14 of Minitab, the following steps will generate 1,000 randomly selected numbers from 0 to 9:

1. Select **Calc** from the toolbar
2. Select **Random Data** from list
3. Select **Integer** from list
4. Generate **20** rows of data
5. Store in column(s): **c1–c50**
6. Minimum value: **0**
7. Maximum value: **9**

The preceding steps will produce 1,000 random single-digit numbers that can then be paired to yield 500 pairs of single-digit numbers. (Most computer packages contain a random number generator that can be used to produce similar results.) Table 4.1(a) contains the results of the simulation of 500 pairs/tosses, while Table 4.1(b) summarizes the results.

Note that this approach yields simulated probabilities that are nearly in agreement with our intuition; that is, intuitively we might expect these outcomes to be equally likely. Thus, each of the four outcomes should occur with a probability equal to 1/4, or .25. This assumption was made for the classical interpretation. We will show in Chapter 10 that in order to be 95% certain that the simulated probabilities are within .01 of the true probabilities, the number of tosses should be at least 7,500 and not 500 as we used previously.

TABLE 4.1(a) Simulation of tossing a penny and a dime 500 times

25	32	70	15	96	87	80	43	15	77	89	51	08	36	29	55	42	86	45	93	68	72	49	99	37
82	81	58	50	85	27	99	41	10	31	42	35	50	02	68	33	50	93	73	62	15	15	90	97	24
46	86	89	82	20	23	63	59	50	40	32	72	59	62	58	53	01	85	49	27	31	48	53	07	78
15	81	39	83	79	21	88	57	35	33	49	37	85	42	28	38	50	43	82	47	01	55	42	02	52
66	44	15	40	29	73	11	06	79	81	49	64	32	06	07	31	07	78	73	07	26	36	39	20	14
48	20	27	73	53	21	44	16	00	33	43	95	21	08	19	60	68	30	99	27	22	74	65	22	05
26	79	54	64	94	01	21	47	86	94	24	41	06	81	16	07	30	34	99	54	68	37	38	71	79
86	12	83	09	27	60	49	54	21	92	64	57	07	39	04	66	73	76	74	93	50	56	23	41	23
18	87	21	48	75	63	09	97	96	86	85	68	65	35	92	40	57	87	82	71	04	16	01	03	45
52	79	14	12	94	51	39	40	42	17	32	94	42	34	68	17	39	32	38	03	75	56	79	79	57
07	40	96	46	22	04	12	90	80	71	46	11	18	81	54	95	47	72	06	07	66	05	59	34	81
66	79	83	82	62	20	75	71	73	79	48	86	83	74	04	13	36	87	96	11	39	81	59	41	70
21	47	34	02	05	73	71	57	64	58	05	16	57	27	66	92	97	68	18	52	09	45	34	80	57
87	22	18	65	66	18	84	31	09	38	05	67	10	45	03	48	52	48	33	36	00	49	39	55	35
70	84	50	37	58	41	08	62	42	64	02	29	33	68	87	58	52	39	98	78	72	13	13	15	96
57	32	98	05	83	39	13	39	37	08	17	01	35	13	98	66	89	40	29	47	37	65	86	73	42
85	65	78	05	24	65	24	92	03	46	67	48	90	60	02	61	21	12	80	70	35	15	40	52	76
29	11	45	22	38	33	32	52	17	20	03	26	34	18	85	46	52	66	63	30	84	53	76	47	21
42	97	56	38	41	87	14	43	30	35	99	06	76	67	00	47	83	32	52	42	48	51	69	15	18
08	30	37	89	17	89	23	58	13	93	17	44	09	08	61	05	35	44	91	89	35	15	06	39	27

TABLE 4.1(b)
Summary of the simulation

Event	Outcome of Simulation	Frequency	Relative Frequency
TT	(Odd, Odd)	129	129/500 = .258
TH	(Odd, Even)	117	117/500 = .234
HT	(Even, Odd)	125	125/500 = .250
HH	(Even, Even)	129	129/500 = .258

If we wish to find the probability of tossing two coins and observing exactly one head, we have, from Table 4.1(b),

$$P(\text{exactly 1 head}) \cong \frac{117 + 125}{500} = .484$$

This is very close to the theoretical probability, which we have shown to be .5.

Note that we could easily modify our example to accommodate the tossing of an unfair coin. Suppose we are tossing a penny that is weighted so that the probability of a head occurring in a toss is .70 and the probability of a tail is .30. We could designate an H outcome whenever one of the random digits 0, 1, 2, 3, 4, 5, or 6 occurs and a T outcome whenever one of the digits 7, 8, or 9 occurs. The same simulation program can be run as before, but we would interpret the output differently.

4.3 Basic Event Relations and Probability Laws

The probability of an event, say event A, will always satisfy the property

$$0 \leq P(A) \leq 1$$

that is, the probability of an event lies anywhere in the interval from 0 (the occurrence of the event is impossible) to 1 (the occurrence of an event is a "sure thing").

either *A* or *B* occurs

Suppose A and B represent two experimental events and you are interested in a new event, the event that **either *A* or *B* occurs.** For example, suppose that we toss a pair of dice and define the following events:

A: A total of 7 shows
B: A total of 11 shows

Then the event "either A or B occurs" is the event that you toss a total of either 7 or 11 with the pair of dice.

mutually exclusive

Note that, for this example, the events A and B are **mutually exclusive;** that is, if you observe event A (a total of 7), you could not at the same time observe event B (a total of 11). Thus, if A occurs, B cannot occur (and vice versa).

DEFINITION 4.1

Two events A and B are said to be **mutually exclusive** if (when the experiment is performed a single time) the occurrence of one of the events excludes the possibility of the occurrence of the other event.

The concept of mutually exclusive events is used to specify a second property that the probabilities of events must satisfy. When two events are mutually exclusive, then the probability that either one of the events will occur is the sum of the event probabilities.

DEFINITION 4.2

If two events, A and B, are mutually exclusive, the **probability** that either event occurs is $P(\text{either } A \text{ or } B) = P(A) + P(B)$.

Definition 4.2 is a special case of the union of two events, which we will soon define.

The definition of additivity of probabilities for mutually exclusive events can be extended beyond two events. For example, when we toss a pair of dice, the sum S of the numbers appearing on the dice can assume any one of the values $S = 2, 3, 4, \ldots, 11, 12$. On a single toss of the dice, we can observe only one of these values. Therefore, the values $2, 3, \ldots, 12$ represent mutually exclusive events. If we want to find the probability of tossing a sum less than or equal to 4, this probability is

$$P(S \leq 4) = P(2) + P(3) + P(4)$$

For this particular experiment, the dice can fall in 36 different equally likely ways. We can observe a 1 on die 1 and a 1 on die 2, denoted by the symbol (1, 1). We can observe a 1 on die 1 and a 2 on die 2, denoted by (1, 2). In other words, for this experiment, the possible outcomes are

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

As you can see, only one of these events, (1, 1), will result in a sum equal to 2. Therefore, we would expect a 2 to occur with a relative frequency of 1/36 in a long series of repetitions of the experiment, and we let $P(2) = 1/36$. The sum $S = 3$ will occur if we observe either of the outcomes (1, 2) or (2, 1). Therefore, $P(3) = 2/36 = 1/18$. Similarly, we find $P(4) = 3/36 = 1/12$. It follows that

$$P(S \leq 4) = P(2) + P(3) + P(4) = \frac{1}{36} + \frac{1}{18} + \frac{1}{12} = \frac{1}{6}$$

complement

A third property of event probabilities concerns an event and its **complement.**

DEFINITION 4.3

The **complement** of an event A is the event that A *does not* occur. The complement of A is denoted by the symbol $\overline{A}$.

Thus, if we define the complement of an event A as a new event—namely, "A does not occur"—it follows that

$$P(A) + P(\overline{A}) = 1$$

For an example, refer again to the two-coin-toss experiment. If, in many repetitions of the experiment, the proportion of times you observe event A, "two heads show," is 1/4, then it follows that the proportion of times you observe the event $\overline{A}$, "two heads do not show," is 3/4. Thus, $P(A)$ and $P(\overline{A})$ will always sum to 1.

We can summarize the three properties that the probabilities of events must satisfy as follows:

Properties of Probabilities

If A and B are any two mutually exclusive events associated with an experiment, then $P(A)$ and $P(B)$ must satisfy the following properties:

1. $0 \leq P(A) \leq 1$ and $0 \leq P(B) \leq 1$
2. $P(\text{either } A \text{ or } B) = P(A) + P(B)$
3. $P(A) + P(\overline{A}) = 1$ and $P(B) + P(\overline{B}) = 1$

union
intersection

We can now define two additional event relations: the **union** and the **intersection** of two events.

DEFINITION 4.4

The **union** of two events A and B is the set of all outcomes that are included in either A or B (or both). The union is denoted as $A \cup B$.

DEFINITION 4.5

The **intersection** of two events A and B is the set of all outcomes that are included in both A and B. The intersection is denoted as $A \cap B$.

These definitions along with the definition of the complement of an event formalize some simple concepts. The event $\overline{A}$ occurs when A *does not;* $A \cup B$ occurs when either A or B occurs; $A \cap B$ occurs when A *and* B occur.

The additivity of probabilities for mutually exclusive events, called the *addition law for mutually exclusive events,* can be extended to give the general addition law.

DEFINITION 4.6

Consider two events A and B; the **probability of the union** of A and B is

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

EXAMPLE 4.1

Events and event probabilities are shown in the Venn diagram in Figure 4.1. Use this diagram to determine the following probabilities:

a. $P(A)$, $P(\overline{A})$
b. $P(B)$, $P(\overline{B})$
c. $P(A \cap B)$
d. $P(A \cup B)$

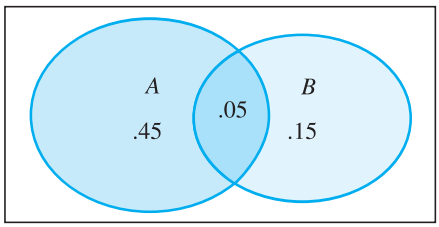


FIGURE 4.1
Probabilities for events A and B

Solution From the Venn diagram, we are able to determine the following probabilities:

a. $P(A) = .5$, therefore $P(\overline{A}) = 1 - .5 = .5$
b. $P(B) = .2$, therefore $P(\overline{B}) = 1 - .2 = .8$
c. $P(A \cap B) = .05$
d. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = .5 + .2 - .05 = .65$

4.4 Conditional Probability and Independence

Consider the following situation: The examination of a large number of insurance claims, categorized according to type of insurance and whether the claim was fraudulent, produced the results shown in Table 4.2. Suppose you are responsible for checking insurance claims—in particular, for detecting fraudulent claims—and you examine the next claim that is processed. What is the probability of the event F, "the claim is fraudulent"? To answer the question, you examine Table 4.2 and note that 10% of all claims are fraudulent. Thus, assuming that the percentages given in the table are reasonable approximations to the true probabilities of receiving specific types of claims, it follows that $P(F) = .10$. Would you say that the risk that you face a fraudulent claim has probability .10? We think not, because you have additional information that may affect the assessment of $P(F)$. This additional information concerns the type of policy you were examining (fire, auto, or other).

TABLE 4.2
Categorization of insurance claims

	Type of Policy (%)			
Category	**Fire**	**Auto**	**Other**	**Total %**
Fraudulent	6	1	3	10
Nonfraudulent	14	29	47	90
Total	20	30	50	100

Suppose that you have the additional information that the claim was associated with a fire policy. Checking Table 4.2, we see that 20% (or .20) of all claims are associated with a fire policy and that 6% (or .06) of all claims are fraudulent fire policy claims. Therefore, it follows that the probability that the claim is fraudulent, given that you know the policy is a fire policy, is

$$P(F \mid \text{fire policy}) = \frac{\text{proportion of claims that are fraudulent fire policy claims}}{\text{proportion of claims that are against fire policies}}$$

$$= \frac{.06}{.20} = .30$$

conditional probability

This probability, $P(F \mid \text{fire policy})$, is called a **conditional probability** of the event F—that is, the probability of event F given the fact that the event "fire policy" has already occurred. This tells you that 30% of all fire policy claims are fraudulent. The vertical bar in the expression $P(F \mid \text{fire policy})$ represents the phrase "given that," or simply "given." Thus, the expression is read, "the probability of the event F given the event fire policy."

unconditional probability

The probability $P(F) = .10$, called the **unconditional** or **marginal probability** of the event F, gives the proportion of times a claim is fraudulent—that is, the proportion of times event F occurs in a very large (infinitely large) number of repetitions of the experiment (receiving an insurance claim and determining whether the claim is fraudulent). In contrast, the conditional probability of F, given that the claim is for a fire policy, $P(F \mid \text{fire policy})$, gives the proportion of fire policy claims that are fraudulent. Clearly, the conditional probabilities of F, given the types of policies, will be of much greater assistance in measuring the risk of fraud than the unconditional probability of F.

DEFINITION 4.7

Consider two events A and B with nonzero probabilities, $P(A)$ and $P(B)$. The **conditional probability** of event A given event B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The conditional probability of event B given event A is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

This definition for conditional probabilities gives rise to what is referred to as the *multiplication law*.

DEFINITION 4.8

The **probability of the intersection** of two events A and B is

$$P(A \cap B) = P(A)P(B \mid A)$$
$$= P(B)P(A \mid B)$$

The only difference between Definitions 4.7 and 4.8, both of which involve conditional probabilities, relates to what probabilities are known and what needs to be calculated. When the intersection probability $P(A \cap B)$ and the individual probability $P(A)$ are known, we can compute $P(B|A)$. When we know $P(A)$ and $P(B|A)$, we can compute $P(A \cap B)$.

EXAMPLE 4.2

A corporation is proposing to select two of its current regional managers as vice presidents. In the history of the company, there has never been a female vice president. The corporation has six male regional managers and four female regional managers. Make the assumption that the 10 regional managers are equally qualified and hence all possible groups of two managers should have the same chance of being selected as the vice presidents. Now find the probability that both vice presidents are male.

Solution Let A be the event that the first vice president selected is male and let B be the event that the second vice president selected is also male. The event that represents both selected vice presidents are male is the event (A and B)—that is, the event $A \cap B$. Therefore, we want to calculate $P(A \cap B) = P(B|A)P(A)$, using Definition 4.8.

For this example,

$$P(A) = P(\text{first selection is male}) = \frac{\#\text{ of male managers}}{\#\text{ of managers}} = \frac{6}{10}$$

and

$$P(B|A) = P(\text{second selection is male given first selection was male}) = \frac{\#\text{ of male managers after one male manager was selected}}{\#\text{ of managers after one male manager was selected}} = \frac{5}{9}$$

Thus,

$$P(A \cap B) = P(A)P(B|A) = \frac{6}{10}\left(\frac{5}{9}\right) = \frac{30}{90} = \frac{1}{3}$$

Thus, the probability that both vice presidents are male is 1/3, under the condition that all candidates are equally qualified and that each group of two managers has the same chance of being selected. Thus, there is a relatively large probability of selecting two males as the vice presidents under the condition that all candidates are equally likely to be selected.

Suppose that the probability of event A is the same whether event B has or has not occurred; that is, suppose

$$P(A|B) = P(A|\overline{B}) = P(A)$$

Then we say that the occurrence of event A is not dependent on the occurrence of event B, or, simply, that A and B are **independent events.** When $P(A|B) \neq P(A)$, the occurrence of A depends on the occurrence of B, and events A and B are said to be **dependent events.**

independent events

dependent events

DEFINITION 4.9

Two events A and B are **independent events** if

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B)$$

(*Note:* You can show that if $P(A|B) = P(A)$, then $P(B|A) = P(B)$, and vice versa.)

Definition 4.9 leads to a special case of $P(A \cap B)$. When events A and B are independent, it follows that

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B)$$

The concept of independence is of particular importance in sampling. Later in the text, we will discuss drawing samples from two (or more) populations to compare the population means, variances, or some other population parameters. For most of these applications, we will select samples in such a way that the observed values in one sample are independent of the values that appear in another sample. We call these **independent samples.**

independent samples

4.5 Bayes' Formula

In this section, we will show how Bayes' Formula can be used to update conditional probabilities by using sample data when available. These "updated" conditional probabilities are useful in decision making. A particular application of these techniques involves the evaluation of diagnostic tests. Suppose a meat inspector must decide whether a randomly selected meat sample contains *E. coli* bacteria. The inspector conducts a diagnostic test. Ideally, a positive result (Pos) would mean that the meat sample actually has *E. coli,* and a negative result (Neg) would imply that the meat sample is free of *E. coli.* However, the diagnostic test is occasionally in error. The results of the test may be a **false positive,** for which the test's indication of *E. coli* presence is incorrect, or a **false negative,** for which the test's conclusion of *E. coli* absence is incorrect. Large-scale screening tests are conducted to evaluate the accuracy of a given diagnostic test. For example, *E. coli* (E) is placed in 10,000 meat samples, and the diagnostic test yields a positive result for 9,500 samples and a negative result for 500 samples; that is, there are 500 false negatives out of the 10,000 tests. Another 10,000 samples have all traces of *E. coli* (NE) removed, and the diagnostic test yields a positive result for 100 samples and a negative result for 9,900 samples. There are 100 false positives out of the 10,000 tests. We can summarize the results in Table 4.3.

false positive
false negative

Evaluation of test results is as follows:

$$\text{True positive rate} = P(\text{Pos}|\text{E}) = \frac{9{,}500}{10{,}000} = .95$$

$$\text{False positive rate} = P(\text{Pos}|\text{NE}) = \frac{100}{10{,}000} = .01$$

$$\text{True negative rate} = P(\text{Neg}|\text{NE}) = \frac{9{,}900}{10{,}000} = .99$$

$$\text{False negative rate} = P(\text{Neg}|\text{NE}) = \frac{500}{10{,}000} = .05$$

TABLE 4.3
E. coli test data

Diagnostic Test Result	Meat Sample Status	
	E	NE
Positive	9,500	100
Negative	500	9,900
Total	10,000	10,000

sensitivity
specificity

The **sensitivity** of the diagnostic test is the true positive rate—that is, $P(\text{test is positive}|\text{disease is present})$. The **specificity** of the diagnostic test is the true negative rate—that is, $P(\text{test is negative}|\text{disease is not present})$.

The primary question facing the inspector is to evaluate the probability of *E. coli* being present in the meat sample when the test yields a positive result—that is, the inspector needs to know $P(\text{E}|\text{Pos})$. Bayes' Formula provides us with a method to obtain this probability.

Bayes' Formula

If A and B are any events whose probabilities are not 0 or 1, then

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

The above formula was developed by Thomas Bayes in a book published in 1763. We will illustrate the application of Bayes' Formula by returning to the meat inspection example. We can use Bayes' Formula to compute $P(\text{E}|\text{Pos})$ for the meat inspection example. To make this calculation, we need to know the *rate* of *E. coli* in the type of meat being inspected. For this example, suppose that *E. coli* is present in 4.5% of all meat samples; that is, *E. coli* has prevalence $P(\text{E}) = .045$. We can then compute $P(\text{E}|\text{Pos})$ as follows:

$$P(\text{E}|\text{Pos}) = \frac{P(\text{Pos}|\text{E})P(\text{E})}{P(\text{Pos}|\text{E})P(\text{E}) + P(\text{Pos}|\text{NE})P(\text{NE})}$$

$$= \frac{(.95)(.045)}{(.95)(.045) + (.01)(1 - .045)} = .817$$

Thus, *E. coli* is truly present in 81.7% of the tested samples in which a positive test result occurs. Also, we can conclude that 18.3% of the tested samples indicated *E. coli* was present when in fact there was no *E. coli* in the meat sample.

EXAMPLE 4.3

A book club classifies members as heavy, medium, or light purchasers, and separate mailings are prepared for each of these groups. Overall, 20% of the members are heavy purchasers, 30% medium, and 50% light. A member is not classified into a group until 18 months after joining the club, but a test is made of the feasibility of using the first 3 months' purchases to classify members. The following percentages are obtained from existing records of individuals classified as heavy, medium, or light purchasers (Table 4.4):

TABLE 4.4
Book club membership classifications

First 3 Months' Purchases	Group (%)		
	Heavy	Medium	Light
0	5	15	60
1	10	30	20
2	30	40	15
3+	55	15	5

If a member purchases no books in the first 3 months, what is the probability that the member is a light purchaser? (*Note:* This table contains "conditional" percentages for each column.)

Solution Using the conditional probabilities in the table, the underlying purchase probabilities, and Bayes' Formula, we can compute this conditional probability.

$$
\begin{aligned}
&P(\text{light}\,|\,0) \\
&\quad = \frac{P(0\,|\,\text{light})P(\text{light})}{P(0\,|\,\text{light})P(\text{light}) + P(0\,|\,\text{medium})P(\text{medium}) + P(0\,|\,\text{heavy})P(\text{heavy})} \\
&\quad = \frac{(.60)(.50)}{(.60)(.50) + (.15)(.30) + (.05)(.20)} \\
&\quad = .845
\end{aligned}
$$

states of nature
prior probabilities
observable events

likelihoods
posterior probabilities

These examples indicate the basic idea of Bayes' Formula. There is some number k of possible, mutually exclusive, underlying events $A_1, \ldots, A_k$, which are sometimes called the **states of nature.** Unconditional probabilities $P(A_1), \ldots, P(A_k)$, often called **prior probabilities,** are specified. There are m possible, mutually exclusive, **observable events** $B_1, \ldots, B_m$. The conditional probabilities of each observable event given each state of nature, $P(B_i\,|\,A_i)$, are also specified, and these probabilities are called **likelihoods.** The problem is to find the **posterior probabilities** $P(A_i\,|\,B_i)$. *Prior* and *posterior* refer to probabilities before and after observing an event B_i.

Bayes' Formula

If $A_1, \ldots, A_k$ are mutually exclusive states of nature, and if $B_1, \ldots, B_m$ are m possible mutually exclusive observable events, then

$$
\begin{aligned}
P(A_i\,|\,B_j) &= \frac{P(B_j\,|\,A_i)P(A_i)}{P(B_j\,|\,A_1)P(A_1) + P(B_j\,|\,A_2)P(A_2) + \cdots + P(B_j\,|\,A_k)P(A_k)} \\
&= \frac{P(B_j\,|\,A_i)P(A_i)}{\Sigma_i\, P(B_j\,|\,A_i)P(A_i)}
\end{aligned}
$$

EXAMPLE 4.4

In the manufacture of circuit boards, there are three major types of defective boards. The types of defects, along with the percentage of all circuit boards having these defects, are (1) improper electrode coverage (D_1), 2.8%; (2) plating separation (D_2), 1.2%; and (3) etching problems (D_3), 3.2%. A circuit board will contain at most one of the three defects. Defects can be detected with certainty using destructive testing of the finished circuit boards; however, this is not a very practical method for inspecting a large percentage of the circuit boards. A nondestructive inspection procedure has been developed, which has the following outcomes: A_1, which indicates the board has only defect D_1; A_2, which indicates the board has only defect D_2; A_3, which indicates the board has only defect D_3; and A_4, which indicates the board has no defects. The respective likelihoods for the four outcomes of the nondestructive test determined by evaluating a large number of boards known to have exactly one of the three types of defects are given in Table 4.5.

TABLE 4.5
Circuit board defect data

Test Outcome	Type of Defect			
	D_1	D_2	D_3	None
A_1	.90	.06	.02	.02
A_2	.05	.80	.06	.01
A_3	.03	.05	.82	.02
A_4 (no defects)	.02	.09	.10	.95

If a circuit board is tested using the nondestructive test and the outcome indicates no defects (A_4), what are the probabilities that the board has no defect or a D_1, D_2, or D_3 type of defect?

Let D_4 represent the situation in which the circuit board has no defects.

$$P(D_1|A_4) = \frac{P(A_4|D_1)P(D_1)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.02)(.028)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.00056}{.88644} = .00063$$

$$P(D_2|A_4) = \frac{P(A_4|D_2)P(D_2)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.09)(.012)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.00108}{.88644} = .00122$$

$$P(D_3|A_4) = \frac{P(A_4|D_3)P(D_3)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.10)(.032)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.0032}{.88644} = .0036$$

$$P(D_4|A_4) = \frac{P(A_4|D_4)P(D_4)}{P(A_4|D_1)P(D_1) + P(A_4|D_2)P(D_2) + P(A_4|D_3)P(D_3) + P(A_4|D_4)P(D_4)}$$
$$= \frac{(.95)(.928)}{(.02)(.028) + (.09)(.012) + (.10)(.032) + (.95)(.928)} = \frac{.8816}{.88644} = .9945$$

Thus, if the new test indicates that none of the three types of defects is present in the circuit board, there is a very high probability, .9945, that the circuit board in fact is free of defects. In Exercise 4.31, we will ask you to assess the sensitivity of the test for determining the three types of defects.

4.6 Variables: Discrete and Continuous

The basic language of probability developed in this chapter deals with many different kinds of events. We are interested in calculating the probabilities associated with both quantitative and qualitative events. For example, we developed techniques that could be used to determine the probability that a machinist selected at random from the workers in a large automotive plant would suffer an accident during an 8-hour shift. These same techniques are also applicable to finding the probability that a machinist selected at random would work more than 80 hours without suffering an accident.

These qualitative and quantitative events can be classified as events (or outcomes) associated with qualitative and quantitative variables. For example, in the automotive accident study, the randomly selected machinist's accident report

would consist of checking one of the following: No Accident, Minor Accident, or Major Accident. Thus, the data on 100 machinists in the study would be observations on a qualitative variable, because the possible responses are the different categories of accident and are not different in any measurable, numerical amount. Because we cannot predict with certainty what type of accident a particular machinist will suffer, the variable is classified as a **qualitative random variable.** Other examples of qualitative random variables that are commonly measured are political party affiliation, socioeconomic status, the species of insect discovered on an apple leaf, and the brand preferences of customers. There are a finite (and typically quite small) number of possible outcomes associated with any qualitative variable. Using the methods of this chapter, it is possible to calculate the probabilities associated with these events.

qualitative random variable

Many times the events of interest in an experiment are quantitative outcomes associated with a **quantitative random variable,** since the possible responses vary in numerical magnitude. For example, in the automotive accident study, the number of consecutive 8-hour shifts between accidents for a randomly selected machinist is an observation on a quantitative random variable. Events of interest, such as the number of 8-hour shifts between accidents for a randomly selected machinist, are observations on a quantitative random variable. Other examples of quantitative random variables are the change in earnings per share of a stock over the next quarter, the length of time a patient is in remission after a cancer treatment, the yield per acre of a new variety of wheat, and the number of persons voting for the incumbent in an upcoming election. The methods of this chapter can be applied to calculate the probability associated with any particular event.

quantitative random variable

There are major advantages to dealing with quantitative random variables. The numerical yardstick underlying a quantitative variable makes the mean and standard deviation (for instance) sensible. With qualitative random variables the methods of this chapter can be used to calculate the probabilities of various events, and that's about all. With quantitative random variables, we can do much more: we can average the resulting quantities, find standard deviations, and assess probable errors, among other things. Hereafter, we use the term **random variable** to mean quantitative random variable.

random variable

Most events of interest result in numerical observations or measurements. If a quantitative variable measured (or observed) in an experiment is denoted by the symbol y, we are interested in the values that y can assume. These values are called *numerical outcomes.* The number of different plant species per acre in a coal strip mine after a reclamation project is a numerical outcome. The percentage of registered voters who cast ballots in a given election is also a numerical outcome. The quantitative variable y is called a *random variable* because the value that y assumes in a given experiment is a chance or random outcome.

DEFINITION 4.10

When observations on a quantitative random variable can assume only a countable number of values, the variable is called a **discrete random variable.**

Examples of discrete variables are these:

1. Number of bushels of apples per tree of a genetically altered apple variety
2. Change in the number of accidents per month at an intersection after a new signaling device has been installed
3. Number of "dead persons" voting in the last mayoral election in a major midwest city

Note that it is possible to count the number of values that each of these random variables can assume.

DEFINITION 4.11

When observations on a quantitative random variable can assume any one of the uncountable number of values in a line interval, the variable is called a **continuous random variable.**

For example, the daily maximum temperature in Rochester, New York, can assume any of the infinitely many values on a line interval. It can be 89.6, 89.799, or 89.7611114. Typical continuous random variables are temperature, pressure, height, weight, and distance.

The distinction between **discrete** and **continuous random** variables is pertinent when we are seeking the probabilities associated with specific values of a random variable. The need for the distinction will be apparent when probability distributions are discussed in later sections of this chapter.

4.7 Probability Distributions for Discrete Random Variables

probability distribution

As previously stated, we need to know the probability of observing a particular sample outcome in order to make an inference about the population from which the sample was drawn. To do this, we need to know the probability associated with each value of the variable y. Viewed as relative frequencies, these probabilities generate a distribution of theoretical relative frequencies called the **probability distribution** of y. Probability distributions differ for discrete and continuous random variables. For discrete random variables, we will compute the probability of specific individual values occurring. For continuous random variables, the probability of an interval of values is the event of interest.

The *probability distribution for a discrete random variable* displays the probability $P(y)$ associated with each value of y. This display can be presented as a table, a graph, or a formula. To illustrate, consider the tossing of two coins in Section 4.2 and let y be the number of heads observed. Then y can take the values 0, 1, or 2. From the data of Table 4.1, we can determine the approximate probability for each value of y, as given in Table 4.6. We point out that the relative frequencies in the table are very close to the theoretical relative frequencies (probabilities), which can be shown to be .25, .50, and .25 using the classical interpretation of probability. If we had employed 2,000,000 tosses of the coins instead of 500, the relative frequencies for $y = 0, 1$, and 2 would be indistinguishable from the theoretical probabilities.

The probability distribution for y, the number of heads in the toss of two coins, is shown in Table 4.7 and is presented graphically in Figure 4.2.

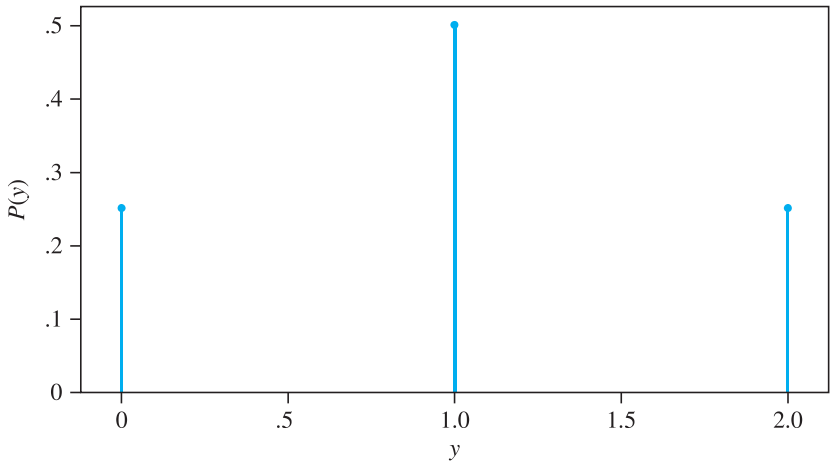
TABLE 4.6
Empirical sampling results for y: the number of heads in 500 tosses of two coins

y	Frequency	Relative Frequency
0	129	.258
1	242	.484
2	129	.258

TABLE 4.7
Probability distribution for the number of heads when two coins are tossed

y	$P(y)$
0	.25
1	.50
2	.25

FIGURE 4.2
Probability distribution for the number of heads when two coins are tossed



The probability distribution for this simple discrete random variable illustrates three important properties of discrete random variables.

Properties of Discrete Random Variables

1. The probability associated with every value of y lies between 0 and 1.
2. The sum of the probabilities for all values of y is equal to 1.
3. The probabilities for a discrete random variable are additive. Hence, the probability that $y = 1$ or 2 is equal to $P(1) + P(2)$.

The relevance of the probability distribution to statistical inference will be emphasized when we discuss the probability distribution for the binomial random variable.

4.8 Two Discrete Random Variables: The Binomial and the Poisson

Many populations of interest to business persons and scientists can be viewed as large sets of 0s and 1s. For example, consider the set of responses of all adults in the United States to the question, "Do you favor the development of nuclear energy?" If we disallow "no opinion," the responses will constitute a set of "yes" responses and "no" responses. If we assign a 1 to each yes and a 0 to each no, the population will consist of a set of 0s and 1s, and the sum of the 1s will equal the total number of persons favoring the development. The sum of the 1s divided by the number of adults in the United States will equal the proportion of people who favor the development.

Gallup and Harris polls are examples of the sampling of 0, 1 populations. People are surveyed, and their opinions are recorded. Based on the sample responses, Gallup and Harris estimate the proportions of people in the population who favor some particular issue or possess some particular characteristic.

Similar surveys are conducted in the biological sciences, engineering, and business, but they may be called experiments rather than polls. For example, experiments are conducted to determine the effect of new drugs on small animals, such as rats or mice, before progressing to larger animals and, eventually, to human participants. Many of these experiments bear a marked resemblance to a poll in that the

experimenter records only whether the drug was effective. Thus, if 300 rats are injected with a drug and 230 show a favorable response, the experimenter has conducted a "poll"—a poll of rat reaction to the drug, 230 "in favor" and 70 "opposed."

Similar "polls" are conducted by most manufacturers to determine the fraction of a product that is of good quality. Samples of industrial products are collected before shipment and each item in the sample is judged "defective" or "acceptable" according to criteria established by the company's quality control department. Based on the number of defectives in the sample, the company can decide whether the product is suitable for shipment. Note that this example, as well as those preceding, has the practical objective of making an inference about a population based on information contained in a sample.

The public opinion poll, the consumer preference poll, the drug-testing experiment, and the industrial sampling for defectives are all examples of a common, frequently conducted sampling situation known as a *binomial experiment.* The binomial experiment is conducted in all areas of science and business and only differs from one situation to another in the nature of objects being sampled (people, rats, electric lightbulbs, oranges). Thus, it is useful to define its characteristics. We can then apply our knowledge of this one kind of experiment to a variety of sampling experiments.

For all practical purposes the binomial experiment is identical to the coin-tossing example of previous sections. Here, n different coins are tossed (or a single coin is tossed n times), and we are interested in the number of heads observed. We assume that the probability of tossing a head on a single trial is π (π may equal .50, as it would for a balanced coin, but in many practical situations π will take some other value between 0 and 1). We also assume that the outcome for any one toss is unaffected by the results of any preceding tosses. These characteristics can be summarized as shown here.

DEFINITION 4.12

A **binomial experiment** is one that has the following properties:

1. The experiment consists of n identical trials.
2. Each trial results in one of two outcomes. We will label one outcome a success and the other a failure.
3. The probability of success on a single trial is equal to π and π remains the same from trial to trial.*
4. The trials are independent; that is, the outcome of one trial does not influence the outcome of any other trial.
5. The random variable y is the number of successes observed during the n trials.

EXAMPLE 4.5

An article in the March 5, 1998, issue of *The New England Journal of Medicine* discussed a large outbreak of tuberculosis. One person, called the index patient, was diagnosed with tuberculosis in 1995. The 232 co-workers of the index patient were given a tuberculin screening test. The number of co-workers recording a positive reading on the test was the random variable of interest. Did this study satisfy the properties of a binomial experiment?

---

*Some textbooks and computer programs use the letter p rather than π. We have chosen π to avoid confusion with p-values, discussed in Chapter 5.

Solution To answer the question, we check each of the five characteristics of the binomial experiment to determine whether they were satisfied.

1. Were there n identical trials? Yes. There were $n = 232$ workers who had approximately equal contact with the index patient.
2. Did each trial result in one of two outcomes? Yes. Each co-worker recorded either a positive or negative reading on the test.
3. Was the probability of success the same from trial to trial? Yes, if the co-workers had equivalent risk factors and equal exposures to the index patient.
4. Were the trials independent? Yes. The outcome of one screening test was unaffected by the outcome of the other screening tests.
5. Was the random variable of interest to the experimenter the number of successes y in the 232 screening tests? Yes. The number of co-workers who obtained a positive reading on the screening test was the variable of interest.

All five characteristics were satisfied, so the tuberculin screening test represented a binomial experiment.

EXAMPLE 4.6

A large power utility company uses gas turbines to generate electricity. The engineers employed at the company monitor the reliability of the turbines—that is, the probability that the turbine will perform properly under standard operating conditions over a specified period of time. The engineers want to estimate the probability a turbine will operate successfully for 30 days after being put into service. The engineers randomly selected 75 of the 100 turbines currently in use and examined the maintenance records. They recorded the number of turbines that did not need repairs during the 30-day time period. Is this a binomial experiment?

Solution Check this experiment against the five characteristics of a binomial experiment.

1. Are there identical trials? The 75 trials could be assumed identical only if the 100 turbines are of the same type of turbine, are the same age, and are operated under the same conditions.
2. Does each trial result in one of two outcomes? Yes. Each turbine either does or does not need repairs in the 30-day time period.
3. Is the probability of success the same from trial to trial? No. If we let success denote a turbine "did not need repairs," then the probability of success can change considerably from trial to trial. For example, suppose that 15 of the 100 turbines needed repairs during the 30-day inspection period. Then π, the probability of success for the first turbine examined, would be $85/100 = .85$. If the first trial is a failure (turbine needed repairs), the probability that the second turbine examined did not need repairs is $85/99 = .859$. Suppose that after 60 turbines have been examined, 50 did not need repairs and 10 needed repairs. The probability of success of the next (61st) turbine would be $35/40 = .875$.
4. Were the trials independent? Yes, provided that the failure of one turbine does not affect the performance of any other turbine. However,

the trials may be dependent in certain situations. For example, suppose that a major storm occurs that results in several turbines being damaged. Then the common event, a storm, may result in a common result, the simultaneous failure of several turbines.

5. Was the random variable of interest to the engineers the number of successes in the 75 trials? Yes. The number of turbines not needing repairs during the 30-day period was the random variable of interest.

This example shows how the probability of success can change substantially from trial to trial in situations in which the sample size is a relatively large portion of the total population size. This experiment does not satisfy the properties of a binomial experiment.

Note that very few real-life situations satisfy perfectly the requirements stated in Definition 4.12, but for many the lack of agreement is so small that the binomial experiment still provides a very good model for reality.

Having defined the binomial experiment and suggested several practical applications, we now examine the probability distribution for the binomial random variable y, the number of successes observed in n trials. Although it is possible to approximate $P(y)$, the probability associated with a value of y in a binomial experiment, by using a relative frequency approach, it is easier to use a general formula for binomial probabilities.

Formula for Computing $P(y)$ in a Binomial Experiment

The probability of observing y successes in n trials of a binomial experiment is

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$$

where

n = number of trials
π = probability of success on a single trial
$1-\pi$ = probability of failure on a single trial
y = number of successes in n trials
$n! = n(n-1)(n-2)\cdots(3)(2)(1)$

As indicated in the box, the notation $n!$ (referred to as n factorial) is used for the product

$$n! = n(n-1)(n-2)\cdots(3)(2)(1)$$

For $n = 3$,

$$n! = 3! = (3)(3-1)(3-2) = (3)(2)(1) = 6$$

Similarly, for $n = 4$,

$$4! = (4)(3)(2)(1) = 24$$

We also note that 0! is defined to be equal to 1.

To see how the formula for binomial probabilities can be used to calculate the probability for a specific value of y, consider the following examples.

EXAMPLE 4.7

A new variety of turf grass has been developed for use on golf courses, with the goal of obtaining a germination rate of 85%. To evaluate the grass, 20 seeds are planted in a greenhouse so that each seed will be exposed to identical conditions. If the 85% germination rate is correct, what is the probability that 18 or more of the 20 seeds will germinate?

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}$$

and substituting for $n = 20$, $\pi = .85$, $y = 18$, 19, and 20, we obtain

$$P(y = 18) = \frac{20!}{18!(20-18)!}(.85)^{18}(1-.85)^{20-18} = 190(.85)^{18}(.15)^2 = .229$$

$$P(y = 19) = \frac{20!}{19!(20-19)!}(.85)^{19}(1-.85)^{20-19} = 20(.85)^{19}(.15)^1 = .137$$

$$P(y = 20) = \frac{20!}{20!(20-20)!}(.85)^{20}(1-.85)^{20-20} = (.85)^{20} = .0388$$

$$P(y \geq 18) = P(y = 18) + P(y = 19) + P(y = 20) = .405$$

The calculations in Example 4.7 entail a considerable amount of effort even though n was only 20. For those situations involving a large value of n, we can use computer software to make the exact calculations. An approach that yields fairly accurate results in many situations and does not require the use of a computer will be discussed later in this chapter.

EXAMPLE 4.8

Suppose that a sample of households is randomly selected from all the households in the city in order to estimate the percentage in which the head of the household is unemployed. To illustrate the computation of a binomial probability, suppose that the unknown percentage is actually 10% and that a sample of $n = 5$ (we select a small sample to make the calculation manageable) is selected from the population. What is the probability that all five heads of the households are employed?

Solution We must carefully define which outcome we wish to call a success. For this example, we define a success as being employed. Then the probability of success when one person is selected from the population is $\pi = .9$ (because the proportion unemployed is .1). We wish to find the probability that $y = 5$ (all five are employed) in five trials.

$$P(y = 5) = \frac{5!}{5!(5-5)!}(.9)^5(.1)^0$$

$$= \frac{5!}{5!10!}(.9)^5(.1)^0$$

$$= (.9)^5 = .590$$

The binomial probability distribution for $n = 5$, $\pi = .9$ is shown in Figure 4.3. The probability of observing five employed in a sample of five is shown to be 0.59 in Figure 4.3.

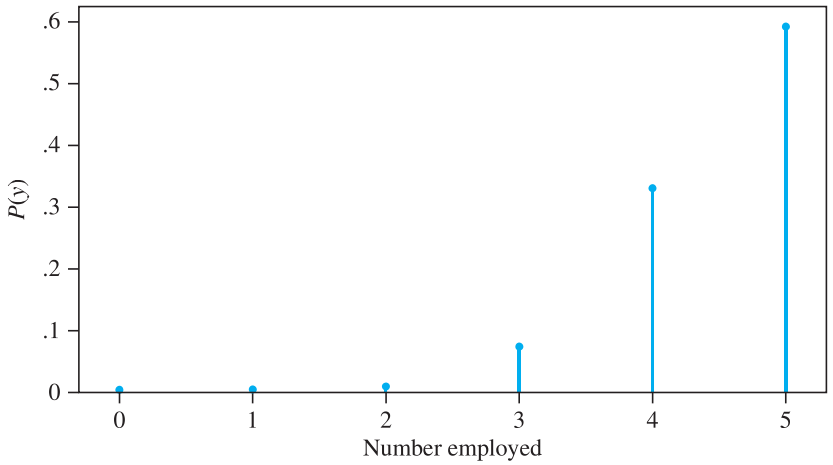


FIGURE 4.3
The binomial probability distribution for $n = 5$, $\pi = .9$

EXAMPLE 4.9

Refer to Example 4.8 and calculate the probability that exactly one person in the sample of five households is unemployed. What is the probability of one or fewer being unemployed?

Solution Since y is the number of employed in the sample of five, one unemployed person would correspond to four employed ($y = 4$). Then

$$\begin{aligned} P(4) &= \frac{5!}{4!(5-4)!}(.9)^4(.1)^1 \\ &= \frac{(5)(4)(3)(2)(1)}{(4)(3)(2)(1)(1)}(.9)^4(.1) \\ &= 5(.9)^4(.1) \\ &= .328 \end{aligned}$$

Thus, the probability of selecting four employed heads of households in a sample of five is .328, or, roughly, one chance in three.

The outcome "one or fewer unemployed" is the same as the outcome "4 or 5 employed." Since y represents the number employed, we seek the probability that $y = 4$ or 5. Because the values associated with a random variable represent mutually exclusive events, the probabilities for discrete random variables are additive. Thus, we have

$$\begin{aligned} P(y = 4 \text{ or } 5) &= P(4) + P(5) \\ &= .328 + .590 \\ &= .918 \end{aligned}$$

Thus, the probability that a random sample of five households will yield either four or five employed heads of households is .918. This high probability is consistent with our intuition: we could expect the number of employed in the sample to be large if 90% of all heads of households in the city are employed.

Like any relative frequency histogram, a binomial probability distribution possesses a mean, μ, and a standard deviation, σ. Although we omit the derivations, we give the formulas for these parameters.

Mean and Standard Deviation of the Binomial Probability Distribution

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1 - \pi)}$$

where π is the probability of success in a given trial and n is the number of trials in the binomial experiment.

If we know π and the sample size, n, we can calculate μ and σ to locate the center and describe the variability for a particular binomial probability distribution. Thus, we can quickly determine those values of y that are probable and those that are improbable.

EXAMPLE 4.10

We will consider the turf grass seed example to illustrate the calculation of the mean and standard deviation. Suppose the company producing the turf grass takes a sample of 20 seeds on a regular basis to monitor the quality of the seeds. If the germination rate of the seeds stays constant at 85%, then the average number of seeds that will germinate in the sample of 20 seeds is

$$\mu = n\pi = 20(.85) = 17$$

with a standard deviation of

$$\sigma = \sqrt{n\pi(1 - \pi)} = \sqrt{20(.85)(1 - .85)} = 1.60$$

Suppose we examine the germination records of a large number of samples of 20 seeds each. If the germination rate has remained constant at 85%, then the average number of seeds that germinate should be close to 17 per sample. If in a particular sample of 20 seeds we determine that only 12 had germinated, would the germination rate of 85% seem consistent with our results? Using a computer software program, we can generate the probability distribution for the number of seeds that germinate in the sample of 20 seeds, as shown in Figures 4.4(a) and 4.4(b).

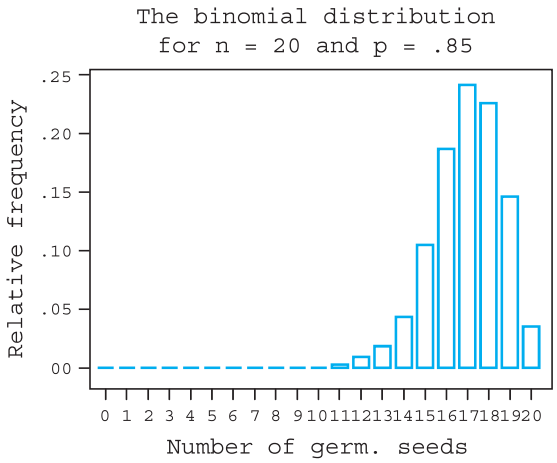
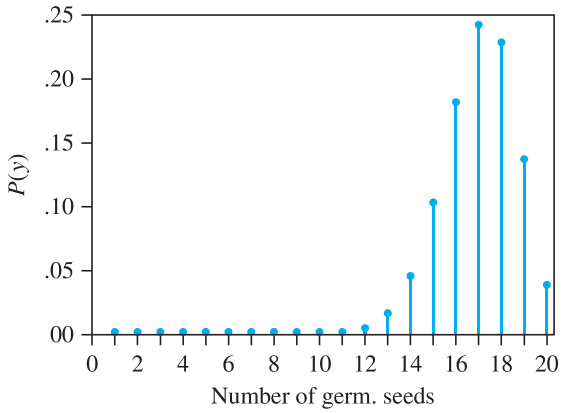


FIGURE 4.4(a)
The binomial distribution for $n = 20$ and $p = .85$

FIGURE 4.4(b)
The binomial distribution for $n = 20$ and $p = .85$



A software program was used to generate Figure 4.4(a). Many such packages place rectangles centered at each of the possible integer values of the binomial random variable as shown in Figure 4.4(a) even though there is zero probability for any value but the integers to occur. This results in a distorted representation of the binomial distribution. A more appropriate display of the distribution is given in Figure 4.4(b).

Although the distribution is tending toward left skewness (see Figure 4.4(b), the Empirical Rule should work well for this relatively mound-shaped distribution. Thus, $y = 12$ seeds is more than 3 standard deviations less than the mean number of seeds, $\mu = 17$; it is highly improbable that in 20 seeds we would obtain only 12 germinated seeds if π really is equal to .85. The germination rate is most likely a value considerably less than .85.

EXAMPLE 4.11

A cable TV company is investigating the feasibility of offering a new service in a large midwestern city. In order for the proposed new service to be economically viable, it is necessary that at least 50% of their current subscribers add the new service. A survey of 1,218 customers reveals that 516 would add the new service. Do you think the company should expend the capital to offer the new service in this city?

Solution In order to be economically viable, the company needs at least 50% of its current customers to subscribe to the new service. Is $y = 516$ out of 1,218 too small a value of y to imply a value of π (the proportion of current customers who would add new service) equal to .50 or larger? If $\pi = .5$,

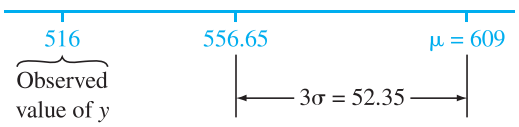
$$\mu = n\pi = 1{,}218(.5) = 609$$

$$\sigma = \sqrt{n\pi(1 - \pi)} = \sqrt{1{,}218(.5)(1 - .5)} = 17.45$$

and $3\sigma = 52.35$.

You can see from Figure 4.5 that $y = 516$ is more than 3σ, or 52.35, less than $\mu = 609$, the value of μ if π really equalled .5. Thus the observed number of

FIGURE 4.5
Location of the observed value of y ($y = 516$) relative to μ



customers in the sample who would add the new service is much too small if the number of current customers who would not add the service, in fact, is 50% or more of all customers. Consequently, the company concluded that offering the new service was not a good idea.

The purpose of this section is to present the binomial probability distribution so you can see how binomial probabilities are calculated and so you can calculate them for small values of n, if you wish. In practice, n is usually large (in national surveys, sample sizes as large as 1,500 are common), and the computation of the binomial probabilities is tedious. Later in this chapter, we will present a simple procedure for obtaining approximate values to the probabilities we need in making inferences. In order to obtain very accurate calculations when n is large, we recommend using a computer software program.

Poisson Distribution

In 1837, S. D. Poisson developed a discrete probability distribution, suitably called the **Poisson Distribution,** which has as one of its important applications the modeling of events of a particular time over a unit of time or space—for example, the number of automobiles arriving at a toll booth during a given 5-minute period of time. The event of interest would be an arriving automobile, and the unit of time would be 5 minutes. A second example would be the situation in which an environmentalist measures the number of PCB particles discovered in a liter of water sampled from a stream contaminated by an electronics production plant. The event would be a PCB particle is discovered. The unit of space would be 1 liter of sampled water.

Let y be the number of events occurring during a fixed time interval of length t or a fixed region R of area or volume $m(R)$. Then the probability distribution of y is Poisson, provided certain conditions are satisfied:

1. Events occur one at a time; two or more events do not occur precisely at the same time or same space.
2. The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a nonoverlapping time period or region of space; that is, the occurrence (or nonoccurrence) of an event during one period or one region does not affect the probability of an event occurring at some other time or in some other region.
3. The expected number of events during one period or region, μ, is the same as the expected number of events in any other period or region.

Although these assumptions seem somewhat restrictive, many situations appear to satisfy these conditions. For example, the number of arrivals of customers at a checkout counter, parking lot toll booth, inspection station, or garage repair shop during a specified time interval can often be modeled by a Poisson distribution. Similarly, the number of clumps of algae of a particular species observed in a unit volume of lake water could be approximated by a Poisson probability distribution.

Assuming that the above conditions hold, the Poisson probability of observing y events in a unit of time or space is given by the formula

$$P(y) = \frac{\mu^y e^{-\mu}}{y!}$$

where e is a naturally occurring constant approximately equal to 2.71828 (in fact, $e = 2 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$), $y! = y(y-1)(y-2)\cdots(1)$, and μ is the average value of y. Table 15 in the Appendix gives Poisson probabilities for various values of the parameter μ.

EXAMPLE 4.12

A large industrial plant is being planned in a rural area. As a part of the environmental impact statement, a team of wildlife scientists is surveying the number and types of small mammals in the region. Let y denote the number of field mice captured in a trap over a 24-hour period. Suppose that y has a Poisson distribution with $\mu = 2.3$; that is, the average number of field mice captured per trap is 2.3. What is the probability of finding exactly four field mice in a randomly selected trap? What is the probability of finding at most four field mice in a randomly selected trap? What is the probability of finding more than four field mice in a randomly selected trap?

Solution The probability that a trap contains exactly four field mice is computed to be

$$P(y = 4) = \frac{e^{-2.3}(2.3)^4}{4!} = \frac{(.1002588)(27.9841)}{24} = .1169$$

Alternatively, we could use Table 15 in the Appendix. We read from the table with $\mu = 2.3$ and $y = 4$ that $P(y = 4) = .1169$.

The probability of finding at most four field mice in a randomly selected trap is, using the values from Table 15, with $\mu = 2.3$

$$\begin{aligned} P(y \leq 4) &= P(y = 0) + P(y = 1) + P(y = 2) + P(y = 3) + P(y = 4) \\ &= .1003 + .2306 + .2652 + .2033 + .1169 = .9163. \end{aligned}$$

The probability of finding more than four field mice in a randomly selected trap is using the idea of complementary events

$$P(y > 4) = 1 - P(y \leq 4) = 1 - .9163 = .0837$$

Thus, it is a very unlikely event to find five or more field mice in a trap.

When n is large and π is small in a binomial experiment, the Poisson distribution provides a good approximation to the binomial distribution. As a general rule, the Poisson distribution provides an adequate approximation to the binomial distribution when $n \geq 100, \pi \leq .01$, and $n\pi \leq 20$. In applying the Poisson approximation to the binomial distribution, take

$$\mu = n\pi$$

EXAMPLE 4.13

In observing patients administered a new drug product in a properly conducted clinical trial, the number of persons experiencing a particular side effect might be quite small. Suppose π (the probability a person experiences a side effect to the drug) is .001 and 1,000 patients in the clinical trial received the drug. Compute the probability that none of a random sample of $n = 1{,}000$ patients administered the drug experiences a particular side effect (such as damage to a heart valve) when $\pi = .001$.

Solution The number of patients, y, experiencing the side effect would have a binomial distribution with $n = 1{,}000$ and $\pi = .001$. The mean of the binomial

distribution is $\mu = n\pi = 1{,}000(.001) = 1$. Applying the Poisson probability distribution with $\mu = 1$, we have

$$P(y = 0) = \frac{(1)^0 e^{-1}}{0!} = e^{-1} = \frac{1}{2.71828} = .367879$$

(Note also from Table 15 in the Appendix that the entry corresponding to $y = 0$ and $\mu = 1$ is .3679.)

For the calculation in Example 4.13 it is easy to compute the exact binomial probability and then compare the results to the Poisson approximation. With $n = 1{,}000$ and $\pi = .001$, we obtain the following.

$$P(y = 0) = \frac{1{,}000!}{0!(1{,}000 - 0)!}(.001)^0(1 - .001)^{1{,}000} = (.999)^{1{,}000} = .367695$$

The Poisson approximation was accurate to the third decimal place.

EXAMPLE 4.14

Suppose that after a clinical trial of a new medication involving 1,000 patients, no patient experienced a side effect to the drug. Would it be reasonable to infer that less than .1% of the entire population would experience this side effect while taking the drug?

Solution Certainly not. We computed the probability of observing $y = 0$ in $n = 1{,}000$ trials assuming $\pi = .001$ (i.e., assuming .1% of the population would experience the side effect) to be .368. Because this probability is quite large, it would not be wise to infer that $\pi < .001$. Rather, we would conclude that there is not sufficient evidence to contradict the assumption that π is .001 or larger.

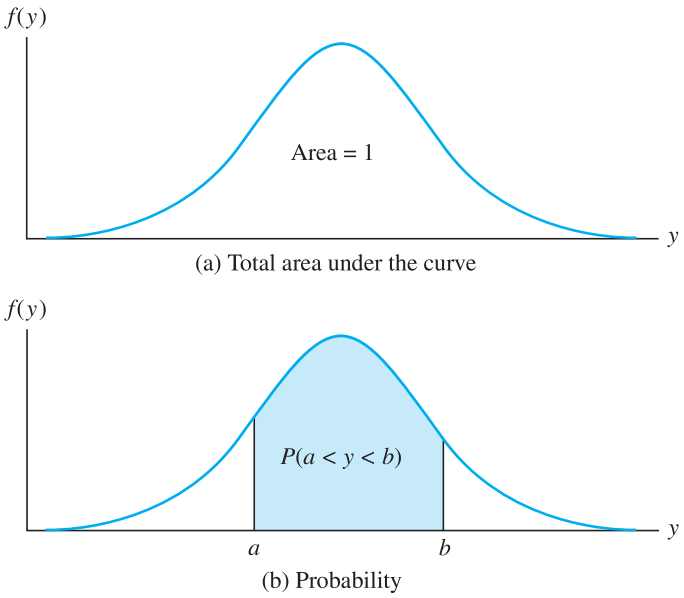
4.9 Probability Distributions for Continuous Random Variables

Discrete random variables (such as the binomial) have possible values that are distinct and separate, such as 0 or 1 or 2 or 3. Other random variables are most usefully considered to be *continuous:* their possible values form a whole interval (or range, or continuum). For instance, the 1-year return per dollar invested in a common stock could range from 0 to some quite large value. In practice, virtually all random variables assume a discrete set of values; the return per dollar of a million-dollar common-stock investment could be \$1.06219423 or \$1.06219424 or \$1.06219425 or . . . . However, when there are many possible values for a random variable, it is sometimes mathematically useful to treat the random variable as continuous.

Theoretically, then, a continuous random variable is one that can assume values associated with infinitely many points in a line interval. We state, without elaboration, that it is impossible to assign a small amount of probability to each value of y (as was done for a discrete random variable) and retain the property that the probabilities sum to 1.

To overcome this difficulty, we revert to the concept of the relative frequency histogram of Chapter 3, where we talked about the probability of y falling in a given interval. Recall that the relative frequency histogram for a population containing a large number of measurements will almost be a smooth curve because the number

FIGURE 4.6
Probability distribution for a continuous random variable

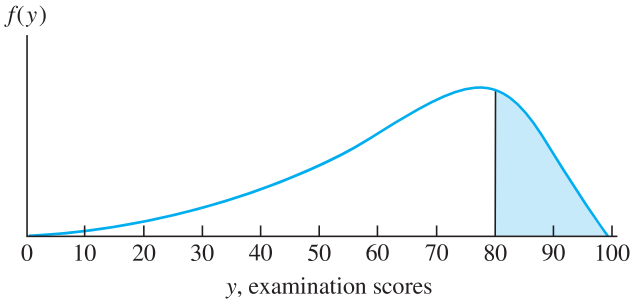


of class intervals can be made large and the width of the intervals can be decreased. Thus, we envision a smooth curve that provides a model for the population relative frequency distribution generated by repeated observation of a continuous random variable. This will be similar to the curve shown in Figure 4.6.

Recall that the histogram relative frequencies are proportional to areas over the class intervals and that these areas possess a probabilistic interpretation. Thus, if a measurement is randomly selected from the set, the probability that it will fall in an interval is proportional to the histogram area above the interval. Since a population is the whole (100%, or 1), we want the total area under the probability curve to equal 1. If we let the total area under the curve equal 1, then areas over intervals are exactly equal to the corresponding probabilities.

The graph for the probability distribution for a continuous random variable is shown in Figure 4.7. The ordinate (height of the curve) for a given value of y is denoted by the symbol $f(y)$. Many people are tempted to say that $f(y)$, like $P(y)$ for the binomial random variable, designates the probability associated with the continuous random variable y. However, as we mentioned before, it is impossible to assign a probability to each of the infinitely many possible values of a continuous random variable. Thus, all we can say is that $f(y)$ represents the height of the probability distribution for a given value of y.

FIGURE 4.7
Hypothetical probability distribution for student examination scores

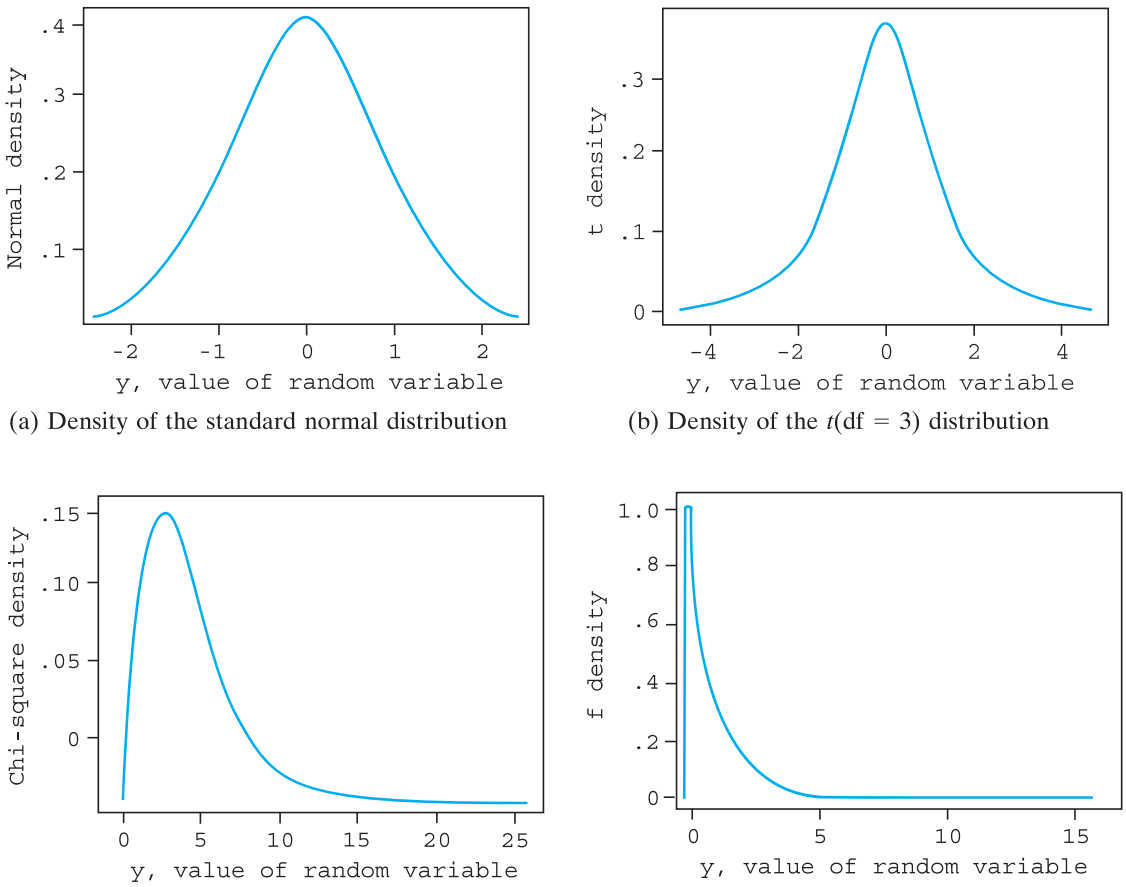


The probability that a continuous random variable falls in an interval, say, between two points a and b, follows directly from the probabilistic interpretation given to the area over an interval for the relative frequency histogram (Section 3.3) and is equal to the area under the curve over the interval a to b, as shown in Figure 4.6. This probability is written $P(a < y < b)$.

There are curves of many shapes that can be used to represent the population relative frequency distribution for measurements associated with a continuous random variable. Fortunately, the areas for many of these curves have been tabulated and are ready for use. Thus, if we know that student examination scores possess a particular probability distribution, as in Figure 4.7, and if areas under the curve have been tabulated, we can find the probability that a particular student will score more than 80% by looking up the tabulated area, which is shaded in Figure 4.7.

Figure 4.8 depicts four important probability distributions that will be used extensively in the following chapters. Which probability distribution we use in a particular situation is very important because probability statements are determined by the area under the curve. As can be seen in Figure 4.8, we would obtain very different answers depending on which distribution is selected. For example, the probability the random variable takes on a value less than 5.0 is essentially 1.0 for the probability distributions in Figures 4.8(a) and (b) but is .584 and .947 for the probability distributions in Figures 4.8(c) and (d), respectively. In some situations,

FIGURE 4.8



(a) Density of the standard normal distribution

(b) Density of the t(df = 3) distribution

(c) Density of the chi-square (df = 5) distribution

(d) Density of the F(df = 2, 6) distribution

we will not know exactly the distribution for the random variable in a particular study. In these situations, we can use the observed values for the random variable to construct a relative frequency histogram, which is a sample estimate of the true probability frequency distribution. As far as statistical inferences are concerned, the selection of the *exact* shape of the probability distribution for a continuous random variable is not crucial in many cases, because most of our inference procedures are insensitive to the exact specification of the shape.

We will find that data collected on continuous variables often possess a nearly bell-shaped frequency distribution, such as depicted in Figure 4.8(a). A continuous variable (the normal) and its probability distribution (bell-shaped curve) provide a good model for these types of data. The normally distributed variable is also very important in statistical inference. We will study the normal distribution in detail in the next section.

4.10 A Continuous Probability Distribution: The Normal Distribution

normal curve

Many variables of interest, including several statistics to be discussed in later sections and chapters, have mound-shaped frequency distributions that can be approximated by using a **normal curve.** For example, the distribution of total scores on the Brief Psychiatric Rating Scale for outpatients having a current history of repeated aggressive acts is mound-shaped. Other practical examples of mound-shaped distributions are social perceptiveness scores of preschool children selected from a particular socioeconomic background, psychomotor retardation scores for patients with circular-type manic-depressive illness, milk yields for cattle of a particular breed, and perceived anxiety scores for residents of a community. Each of these mound-shaped distributions can be approximated with a normal curve.

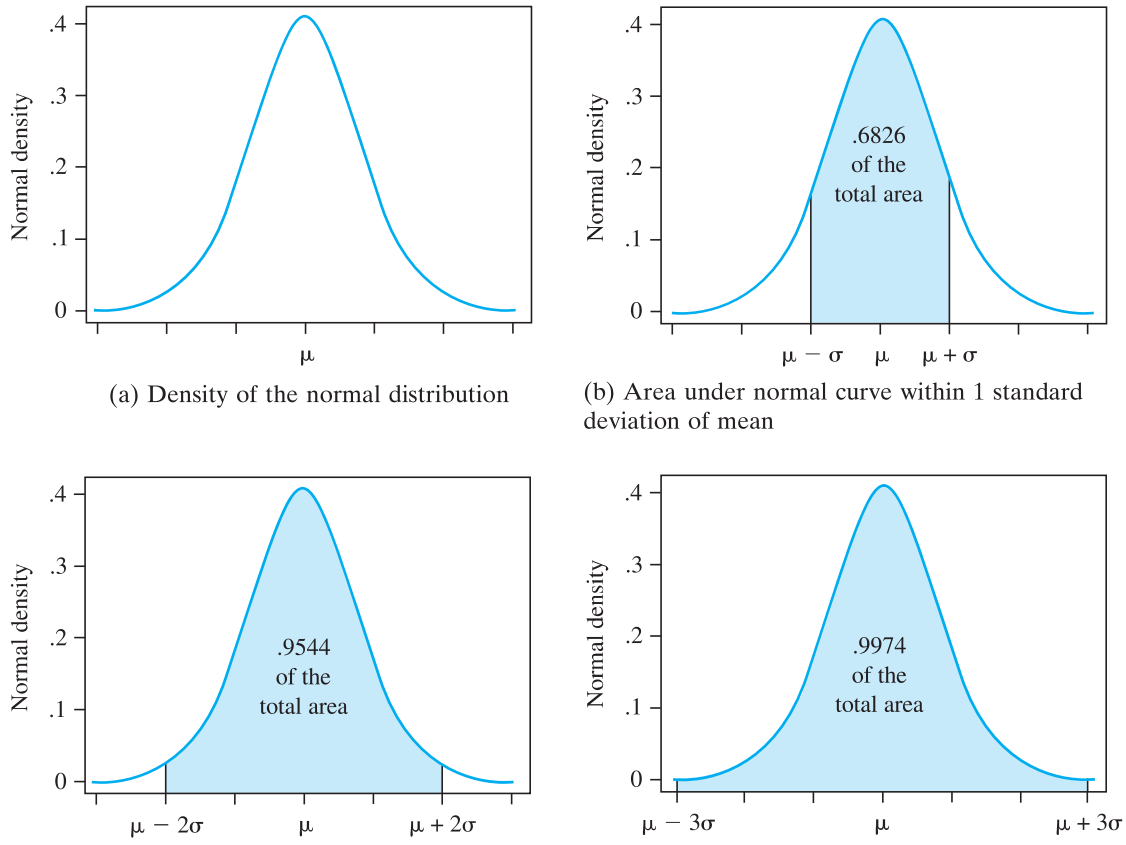
Since the normal distribution has been well tabulated, areas under a normal curve—which correspond to probabilities—can be used to approximate probabilities associated with the variables of interest in our experimentation. Thus, the normal random variable and its associated distribution play an important role in statistical inference.

The relative frequency histogram for the normal random variable, called the *normal curve* or *normal probability distribution,* is a smooth bell-shaped curve. Figure 4.9(a) shows a normal curve. If we let y represent the normal random variable, then the height of the probability distribution for a specific value of y is represented by $f(y)$.* The probabilities associated with a normal curve form the basis for the Empirical Rule.

As we see from Figure 4.9(a), the normal probability distribution is bell shaped and symmetrical about the mean μ. Although the normal random variable y may theoretically assume values from $-\infty$ to $+\infty$, we know from the Empirical Rule that approximately all the measurements are within 3 standard deviations (3σ) of μ. From the Empirical Rule, we also know that if we select a measurement at random from a population of measurements that possesses a mound-shaped distribution, the probability is approximately .68 that the measurement will lie within 1 standard deviation of its mean (see Figure 4.9(b)). Similarly, we know that the probability

*For the normal distribution, $f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$, where μ and σ are the mean and standard deviation, respectively, of the population of y-values.

FIGURE 4.9



(a) Density of the normal distribution

(b) Area under normal curve within 1 standard deviation of mean

(c) Area under normal curve within 2 standard deviations of mean

(d) Area under normal curve within 3 standard deviations of mean

is approximately .954 that a value will lie in the interval $\mu \pm 2\sigma$ and .997 in the interval $\mu \pm 3\sigma$ (see Figures 4.9(c) and (d)). What we do not know, however, is the probability that the measurement will be within 1.65 standard deviations of its mean, or within 2.58 standard deviations of its mean. The procedure we are going to discuss in this section will enable us to calculate the probability that a measurement falls within any distance of the mean μ for a normal curve.

Because there are many different normal curves (depending on the parameters μ and σ), it might seem to be an impossible task to tabulate areas (probabilities) for all normal curves, especially if each curve requires a separate table. Fortunately, this is not the case. By specifying the probability that a variable y lies within a certain number of standard deviations of its mean (just as we did in using the Empirical Rule), we need only one table of probabilities.

area under a normal curve

Table 1 in the Appendix gives the **area under a normal curve** to the left of a value y that is z standard deviations ($z\sigma$) away from the mean (see Figure 4.10). The area shown by the shading in Figure 4.10 is the probability listed in Table 1 in the Appendix. Values of z to the nearest tenth are listed along the left-hand column of the table, with z to the nearest hundredth along the top of the table. To find the probability that a normal random variable will lie to the left of a point 1.65 standard deviations above the mean, we look up the table entry corresponding to $z = 1.65$. This probability is .9505 (see Figure 4.11).

FIGURE 4.10
Area under a normal curve as given in Appendix Table 1

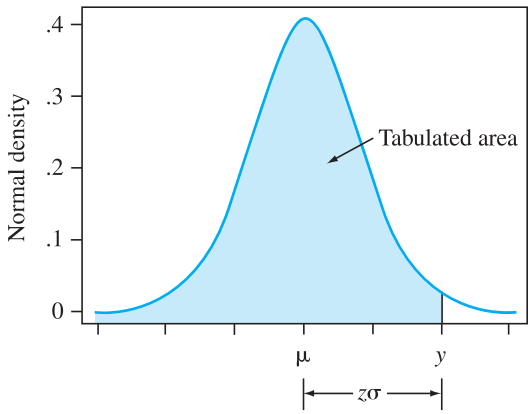
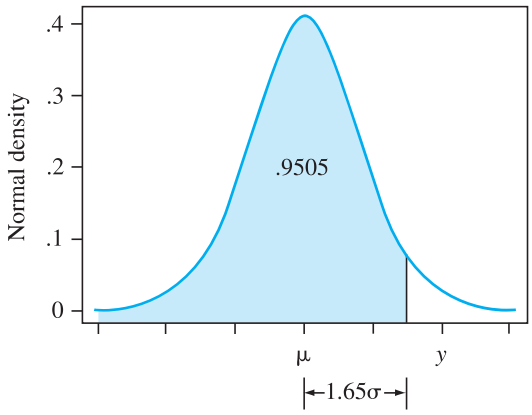


FIGURE 4.11
Area under a normal curve from μ to a point 1.65 standard deviations above the mean



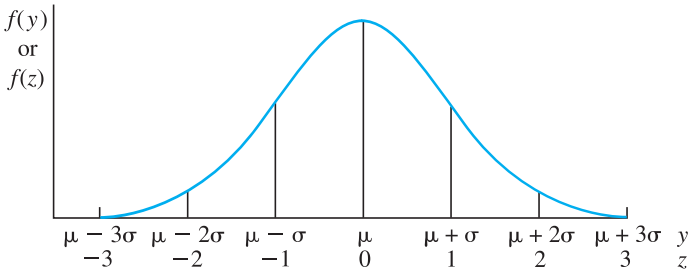
To determine the probability that a measurement will be less than some value y, we first calculate the number of standard deviations that y lies away from the mean by using the formula

$$z = \frac{y - \mu}{\sigma}$$

***z*-score**

The value of z computed using this formula is sometimes referred to as the ***z*-score** associated with the y-value. Using the computed value of z, we determine the appropriate probability by using Table 1 in the Appendix. Note that we are merely coding the value y by subtracting μ and dividing by σ. (In other words, $y = z\sigma + \mu$.) Figure 4.12 illustrates the values of z corresponding to specific values of y. Thus, a value of y that is 2 standard deviations below (to the left of) μ corresponds to $z = -2$.

FIGURE 4.12
Relationship between specific values of y and $z = (y - \mu)/\sigma$



EXAMPLE 4.15

Consider a normal distribution with $\mu = 20$ and $\sigma = 2$. Determine the probability that a measurement will be less than 23.

Solution When first working problems of this type, it might be a good idea to draw a picture so that you can see the area in question, as we have in Figure 4.13.

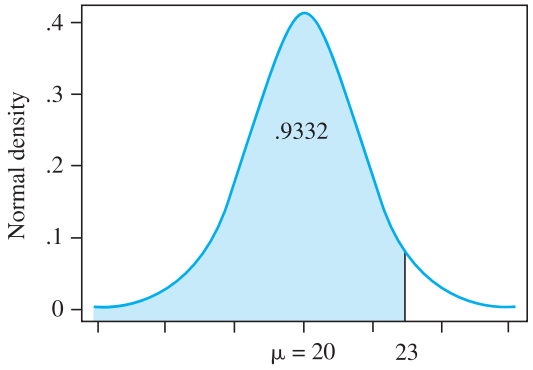


FIGURE 4.13
Area less than $y = 23$ under normal curve, with $\mu = 20$, $\sigma = 2$

To determine the area under the curve to the left of the value $y = 23$, we first calculate the number of standard deviations $y = 23$ lies away from the mean.

$$z = \frac{y - \mu}{\sigma} = \frac{23 - 20}{2} = 1.5$$

Thus, $y = 23$ lies 1.5 standard deviations above $\mu = 20$. Referring to Table 1 in the Appendix, we find the area corresponding to $z = 1.5$ to be .9332. This is the probability that a measurement is less than 23.

EXAMPLE 4.16

For the normal distribution of Example 4.15 with $\mu = 20$ and $\sigma = 2$, find the probability that y will be less than 16.

Solution In determining the area to the left of 16, we use

$$z = \frac{y - \mu}{\sigma} = \frac{16 - 20}{2} = -2$$

We find the appropriate area from Table 1 to be .0228; thus, .0228 is the probability that a measurement is less than 16. The area is shown in Figure 4.14.

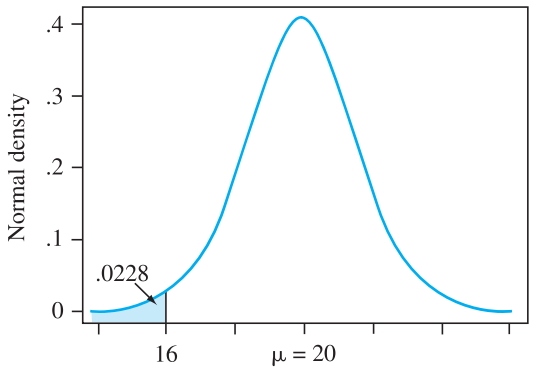


FIGURE 4.14
Area less than $y = 16$ under normal curve, with $\mu = 20$, $\sigma = 2$

EXAMPLE 4.17

A high accumulation of ozone gas in the lower atmosphere at ground level is air pollution and can be harmful to people, animals, crops, and various materials. Elevated levels above the national standard may cause lung and respiratory disorders. Nitrogen oxides and hydrocarbons are known as the chief "precursors" of ozone. These compounds react in the presence of sunlight to produce ozone. The sources of these precursor pollutants include cars, trucks, power plants, and factories. Large industrial areas and cities with heavy summer traffic are the main contributors to ozone formation. The United States Environmental Protection Agency (EPA) has developed procedures for measuring vehicle emission levels of nitrogen oxide. Let P denote the amount of this pollutant in a randomly selected automobile in Houston, Texas. Suppose the distribution of P can be adequately modelled by a normal distribution with a mean level of $\mu = 70$ ppb (parts per billion) and standard deviation of $\sigma = 13$ ppb.

a. What is the probability that a randomly selected vehicle will have emission levels less than 60 ppb?
b. What is the probability that a randomly selected vehicle will have emission levels greater than 90 ppb?
c. What is the probability that a randomly selected vehicle will have emission levels between 60 and 90 ppb?

Solution We begin by drawing pictures of the areas we are looking for (Figures 4.15 (a)–(c)). To answer part (a) we must compute the z-values corresponding to the value of 60. The value $y = 60$ corresponds to a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{60 - 70}{13} = -.77$$

From Table 1, the area to the left of 60 is .2206 (see Figure 4.15(a)).

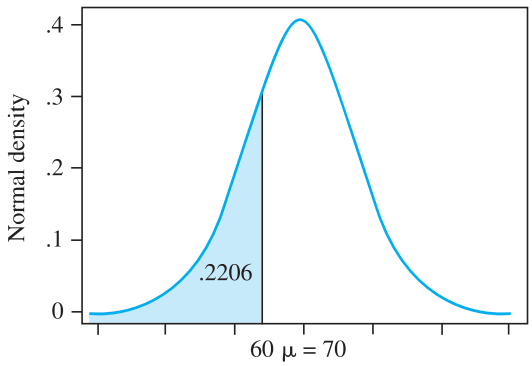


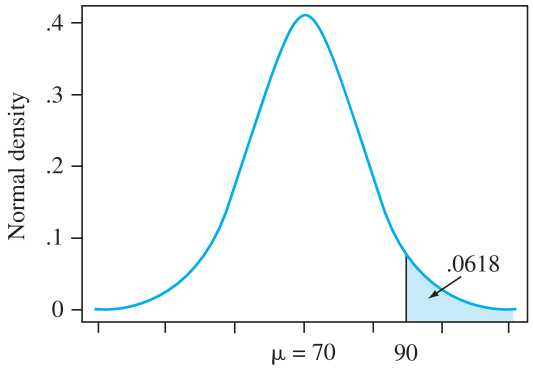
FIGURE 4.15(a)
Area less than $y = 60$ under normal curve, with $\mu = 70$, $\sigma = 13$

To answer part (b), the value $y = 90$ corresponds to a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{90 - 70}{13} = 1.54$$

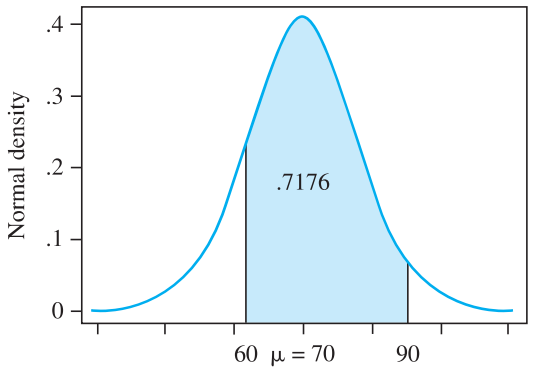
so from Table 1 we obtain .9382, the tabulated area less than 90. Thus, the area greater than 90 must be $1 - .9382 = .0618$, since the total area under the curve is 1 (see Figure 4.15(b)).

FIGURE 4.15(b)
Area greater than $y = 90$ under normal curve, with $\mu = 70, \sigma = 13$



To answer part (c), we can use our results from (a) and (b). The area between two values y_1 and y_2 is determined by finding the difference between the areas to the left of the two values, (see Figure 4.15(c)). We have the area less than 60 is .2206, and the area less than 90 is .9382. Hence, the area between 60 and 90 is .9382 − .2206 = .7176. We can thus conclude that 22.06% of inspected vehicles will have nitrogen oxide levels less than 60 ppb, 6.18% of inspected vehicles will have nitrogen oxide levels greater than 90 ppb, and 71.76% of inspected vehicles will have nitrogen oxide levels between 60 ppb and 90 ppb.

FIGURE 4.15(c)
Area between 60 and 90 under normal curve, with $\mu = 70$, $\sigma = 13$



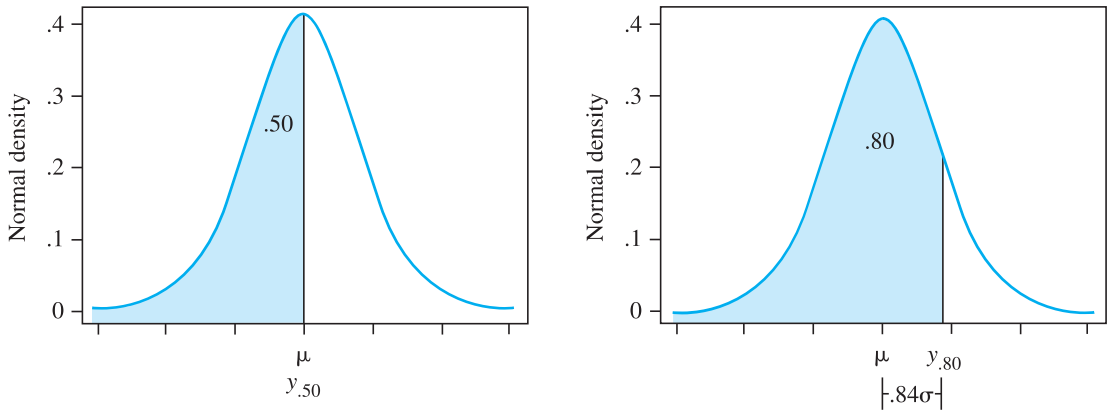
100*p*th percentile

An important aspect of the normal distribution is that we can easily find the percentiles of the distribution. The **100*p*th percentile** of a distribution is that value, y_p, such that $100p\%$ of the population values fall below y_p and $100(1 - p)\%$ are above y_p. For example, the median of a population is the 50th percentile, $y_{.50}$, and the quartiles are the 25th and 75th percentiles. The normal distribution is symmetric, so the median and the mean are the same value, $y_{.50} = \mu$ (see Figure 4.16(a)).

To find the percentiles of the standard normal distribution, we reverse our use of Table 1. To find the 100*p*th percentile, z_p, we find the probability p in Table 1 and then read out its corresponding number, z_p, along the margins of the table. For example, to find the 80th percentile, $z_{.80}$, we locate the probability $p = .8000$ in Table 1. The value nearest to .8000 is .7995, which corresponds to a z-value of 0.84. Thus, $z_{.80} = 0.84$ (see Figure 4.16 (b)). Now, to find the 100*p*th percentile, y_p, of a normal distribution with mean μ and standard deviation σ, we need to apply the reverse of our standardization formula,

$$y_p = \mu + z_p\sigma$$

FIGURE 4.16



(a) For the normal curve, the mean and median agree

(b) The 80th percentile for the normal curve

Suppose we wanted to determine the 80th percentile of a population having a normal distribution with $\mu = 55$ and $\sigma = 3$. We have determined that $z_{.80} = 0.84$; thus, the 80th percentile for the population would be $y_{.80} = 55 + (.84)(3) = 57.52$.

EXAMPLE 4.18

A State of Texas environmental agency, using the vehicle inspection process described in Example 4.17, is going to offer a reduced vehicle license fee to those vehicles having very low emission levels. As a preliminary pilot project, they will offer this incentive to the group of vehicle owners having the best 10% of emission levels. What emission level should the agency use in order to identify the best 10% of all emission levels?

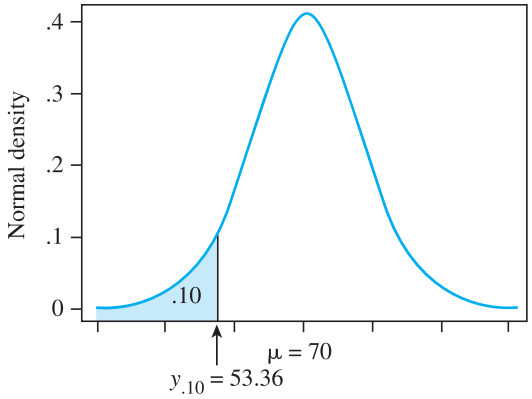
Solution The best 10% of all emission levels would be the 10% having the lowest emission levels, as depicted in Figure 4.17.

To find the tenth percentile (see Figure 4.17), we first find $z_{.10}$ in Table 1. Since .1003 is the value nearest .1000 and its corresponding z-value is -1.28, we take $z_{.10} = -1.28$. We then compute

$$y_{.10} = \mu + z_{.10}\sigma = 70 + (-1.28)(13) = 70 - 16.64 = 53.36$$

Thus, 10% of the vehicles have emissions less than 53.36 ppb.

FIGURE 4.17
The tenth percentile for a normal curve, with $\mu = 70$, $\sigma = 13$



EXAMPLE 4.19

An analysis of income tax returns from the previous year indicates that for a given income classification, the amount of money owed to the government over and above the amount paid in the estimated tax vouchers for the first three payments is approximately normally distributed with a mean of $530 and a standard deviation of $205. Find the 75th percentile for this distribution of measurements. The government wants to target that group of returns having the largest 25% of amounts owed.

Solution We need to determine the 75th percentile, $y_{.75}$, (Figure 4.18). From Table 1, we find $z_{.75} = .67$ because the probability nearest .7500 is .7486, which corresponds to a z-score of .67. We then compute

$$y_{.75} = \mu + z_{.75}\sigma = 530 + (.67)(205) = 667.35$$

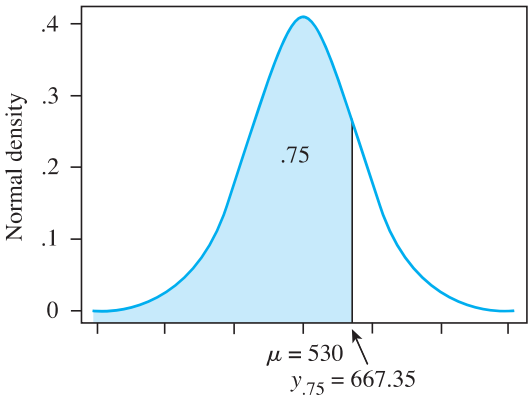


FIGURE 4.18
The 75th percentile for a normal curve, with $\mu = 530$, $\sigma = 205$

Thus, 25% of the tax returns in this classification exceed $667.35 in the amount owed the government.

4.11 Random Sampling

Thus far in the text, we have discussed random samples and introduced various sampling schemes in Chapter 2. What is the importance of random sampling? We must know how the sample was selected so we can determine probabilities associated with various sample outcomes. The probabilities of samples selected *in a random manner* can be determined, and we can use these probabilities to make inferences about the population from which the sample was drawn.

Sample data selected in a nonrandom fashion are frequently distorted by a *selection bias.* A selection bias exists whenever there is a systematic tendency to overrepresent or underrepresent some part of the population. For example, a survey of households conducted during the week entirely between the hours of 9 A.M. and 5 P.M. would be severely biased toward households with at least one member at home. Hence, any inferences made from the sample data would be biased toward the attributes or opinions of those families with at least one member at home and may not be truly representative of the population of households in the region.

random sample

Now we turn to a definition of a **random sample** of n measurements selected from a population containing N measurements ($N > n$). (*Note:* This is a simple random sample as discussed in Chapter 2. Since most of the random samples discussed in this text will be simple random samples, we'll drop the adjective unless needed for clarification.)

DEFINITION 4.13

A sample of n measurements selected from a population is said to be a **random sample** if every different sample of size n from the population has an equal probability of being selected.

EXAMPLE 4.20

A study of crimes related to handguns is being planned for the ten largest cities in the United States. The study will randomly select two of the ten largest cities for an in-depth study following the preliminary findings. The population of interest is the ten largest cities $\{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}\}$. List all possible different samples consisting of two cities that could be selected from the population of ten cities. Give the probability associated with each sample in a random sample of $n = 2$ cities selected from the population.

Solution All possible samples are listed in Table 4.8.

TABLE 4.8
Samples of size 2

Sample	Cities	Sample	Cities	Sample	Cities
1	C_1, C_2	16	C_2, C_9	31	C_5, C_6
2	C_1, C_3	17	C_2, C_{10}	32	C_5, C_7
3	C_1, C_4	18	C_3, C_4	33	C_5, C_8
4	C_1, C_5	19	C_3, C_5	34	C_5, C_9
5	C_1, C_6	20	C_3, C_6	35	C_5, C_{10}
6	C_1, C_7	21	C_3, C_7	36	C_6, C_7
7	C_1, C_8	22	C_3, C_8	37	C_6, C_8
8	C_1, C_9	23	C_3, C_9	38	C_6, C_9
9	C_1, C_{10}	24	C_3, C_{10}	39	C_6, C_{10}
10	C_2, C_3	25	C_4, C_5	40	C_7, C_8
11	C_2, C_4	26	C_4, C_6	41	C_7, C_9
12	C_2, C_5	27	C_4, C_7	42	C_7, C_{10}
13	C_2, C_6	28	C_4, C_8	43	C_8, C_9
14	C_2, C_7	29	C_4, C_9	44	C_8, C_{10}
15	C_2, C_8	30	C_4, C_{10}	45	C_9, C_{10}

Now, let us suppose that we select a random sample of $n = 2$ cities from the 45 possible samples. The sample selected is called a *random* sample if every sample has an equal probability, 1/45, of being selected.

random number table

One of the simplest and most reliable ways to select a random sample of n measurements from a population is to use a table of random numbers (see Table 13 in the Appendix). **Random number tables** are constructed in such a way that, no matter where you start in the table and no matter in which direction you move, the digits occur randomly and with equal probability. Thus, if we wished to choose a random sample of $n = 10$ measurements from a population containing 100 measurements, we could label the measurements in the population from 0 to 99 (or 1 to 100). Then by referring to Table 13 in the Appendix and choosing a random starting point, the next 10 two-digit numbers going across the page would indicate the labels of the particular measurements to be included in the random sample. Similarly, by moving up or down the page, we would also obtain a random sample.

This listing of all possible samples is feasible only when both the sample size n and the population size N are small. We can determine the number, M, of distinct

samples of size n that can be selected from a population of N measurements using the following formula:

$$M = \frac{N!}{n!(N-n)!}$$

In Example 4.20, we had $N = 10$ and $n = 2$. Thus,

$$M = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = 45$$

The value of M becomes very large even when N is fairly small. For example, if $N = 50$ and $n = 5$, then $M = 2{,}118{,}760$. Thus, it would be very impractical to list all 2,118,760 possible samples consisting of $n = 5$ measurements from a population of $N = 50$ measurements and then randomly select one of the samples. In practice, we construct a list of elements in the population by assigning a number from 1 to N to each element in the population, called the *sampling frame.* We then randomly select n integers from the integers $(1, 2, \ldots, N)$ by using a table of random numbers (see Table 13 in the Appendix) or by using a computer program. Most statistical software programs contain routines for randomly selecting n integers from the integers $(1, 2, \ldots, N)$, where $N > n$. Exercise 4.76 contains the necessary commands for using Minitab to generate the random sample.

EXAMPLE 4.21

The school board in a large school district has decided to test for illegal drug use among those high school students participating in extracurricular activities. Because these tests are very expensive, they have decided to institute a random testing procedure. Every week, 20 students will be randomly selected from the 850 high school students participating in extracurricular activities and a drug test will be performed. Refer to Table 13 in the Appendix or use a computer software program to determine which students should be tested.

Solution Using the list of all 850 students participating in extracurricular activities, we label the students from 0 to 849 (or, equivalently, from 1 to 850). Then, referring to Table 13 in the Appendix, we select a starting point (close your eyes and pick a point in the table). Suppose we selected line 1, column 3. Going down the page in Table 13, we select the first 20 three-digit numbers between 000 and 849. We would obtain the following 20 numbers:

015	110	482	333
255	564	526	463
225	054	710	337
062	636	518	224
818	533	524	055

These 20 numbers identify the 20 students that are to be included in the first week of drug testing. We would repeat the process in subsequent weeks using a new starting point.

A telephone directory is often used in selecting people to participate in surveys or pools, especially in surveys related to economics or politics. In the 1936 presidential campaign, Franklin Roosevelt was running as the Democratic candidate against the Republican candidate, Governor Alfred Landon of Kansas. This was a difficult time for the nation; the country had not yet recovered from the Great Depression of the early 1930s, and there were still 9 million people unemployed.

The *Literary Digest* set out to sample the voting public and predict the winner of the election. Using names and addresses taken from telephone books and club memberships, the *Literary Digest* sent out 10 million questionnaires and got 2.4 million back. Based on the responses to the questionnaire, the *Digest* predicted a Landon victory by 57% to 43%.

At this time, George Gallup was starting his survey business. He conducted two surveys. The first one, based on 3,000 people, predicted what the results of the *Digest* survey would be long before the *Digest* results were published; the second survey, based on 50,000, was used to forecast *correctly* the Roosevelt victory.

How did Gallup correctly predict what the *Literary Digest* survey would predict and then, with another survey, correctly predict the outcome of the election? Where did the *Literary Digest* go wrong? The first problem was a severe selection bias. By taking the names and addresses from telephone directories and club memberships, its survey systematically excluded the poor. Unfortunately for the *Digest,* the vote was split along economic lines; the poor gave Roosevelt a large majority, whereas the rich tended to vote for Landon. A second reason for the error could be due to a *nonresponse bias.* Because only 20% of the 10 million people returned their surveys, and approximately half of those responding favored Landon, one might suspect that maybe the nonrespondents had different preferences than did the respondents. This was, in fact, true.

How, then does one achieve a random sample? Careful planning and a certain amount of ingenuity are required to have even a decent chance to approximate random sampling. This is especially true when the universe of interest involves people. People can be difficult to work with; they have a tendency to discard mail questionnaires and refuse to participate in personal interviews. Unless we are very careful, the data we obtain may be full of biases having unknown effects on the inferences we are attempting to make.

We do not have sufficient time to explore the topic of random sampling further in this text; entire courses at the undergraduate and graduate levels can be devoted to sample-survey research methodology. The important point to remember is that data from a random sample will provide the foundation for making statistical inferences in later chapters. Random samples are not easy to obtain, but with care we can avoid many potential biases that could affect the inferences we make. References providing detailed discussions on how to properly conduct a survey were given in Chapter 2.

4.12 Sampling Distributions

We discussed several different measures of central tendency and variability in Chapter 3 and distinguished between numerical descriptive measures of a population (parameters) and numerical descriptive measures of a sample (statistics). Thus, μ and σ are parameters, whereas $\bar{y}$ and s are statistics.

The numerical value of a sample statistic cannot be predicted exactly in advance. Even if we knew that a population mean μ was \$216.37 and that the population standard deviation σ was \$32.90—even if we knew the complete population distribution—we could not say that the sample mean $\bar{y}$ would be exactly equal to \$216.37. A sample statistic is a random variable; it is subject to random variation because it is based on a random sample of measurements selected from the population of interest. Also, like any other random variable, a sample statistic has a probability distribution. We call the probability distribution of a sample statistic the *sampling*

distribution of that statistic. Stated differently, the sampling distribution of a statistic is the population of all possible values for that statistic.

The actual mathematical derivation of sampling distributions is one of the basic problems of mathematical statistics. We will illustrate how the sampling distribution for $\bar{y}$ can be obtained for a simplified population. Later in the chapter, we will present several general results.

EXAMPLE 4.22

The sample $\bar{y}$ is to be calculated from a random sample of size 2 taken from a population consisting of 10 values (2, 3, 4, 5, 6, 7, 8, 9, 10, 11). Find the sampling distribution of $\bar{y}$, based on a random sample of size 2.

Solution One way to find the sampling distribution is by counting. There are 45 possible samples of 2 items selected from the 10 items. These are shown in Table 4.9.

TABLE 4.9
List of values for the sample mean, $\bar{y}$

Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$	Sample	Value of $\bar{y}$
2, 3	2.5	3, 10	6.5	6, 7	6.5
2, 4	3	3, 11	7	6, 8	7
2, 5	3.5	4, 5	4.5	6, 9	7.5
2, 6	4	4, 6	5	6, 10	8
2, 7	4.5	4, 7	5.5	6, 11	8.5
2, 8	5	4, 8	6	7, 8	7.5
2, 9	5.5	4, 9	6.5	7, 9	8
2, 10	6	4, 10	7	7, 10	8.5
2, 11	6.5	4, 11	7.5	7, 11	9
3, 4	3.5	5, 6	5.5	8, 9	8.5
3, 5	4	5, 7	6	8, 10	9
3, 6	4.5	5, 8	6.5	8, 11	9.5
3, 7	5	5, 9	7	9, 10	9.5
3, 8	5.5	5, 10	7.5	9, 11	10
3, 9	6	5, 11	8	10, 11	10.5

Assuming each sample of size 2 is equally likely, it follows that the sampling distribution for $\bar{y}$ based on $n = 2$ observations selected from the population {2, 3, 4, 5, 6, 7, 8, 9, 10, 11} is as indicated in Table 4.10.

TABLE 4.10
Sampling distribution for $\bar{y}$

$\bar{y}$	$P(\bar{y})$	$\bar{y}$	$P(\bar{y})$
2.5	1/45	7	4/45
3	1/45	7.5	4/45
3.5	2/45	8	3/45
4	2/45	8.5	3/45
4.5	3/45	9	2/45
5	3/45	9.5	2/45
5.5	4/45	10	1/45
6	4/45	10.5	1/45
6.5	5/45		

The sampling distribution is shown as a graph in Figure 4.19. Note that the distribution is symmetric, with a mean of 6.5 and a standard deviation of approximately 2.0 (the range divided by 4).

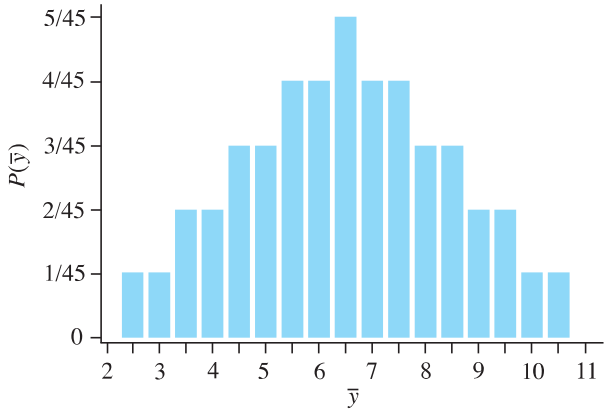


FIGURE 4.19
Sampling distribution for $\bar{y}$

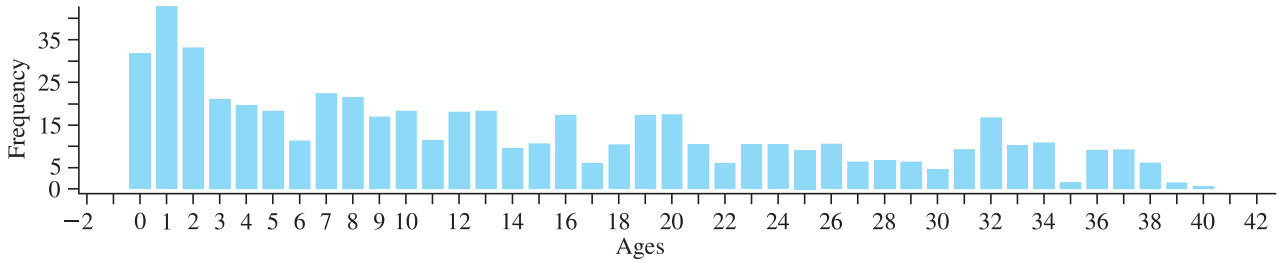
Example 4.22 illustrates for a very small population that we could in fact enumerate every possible sample of size 2 selected from the population and then compute all possible values of the sample mean. The next example will illustrate the properties of the sample mean, $\bar{y}$, when sampling from a larger population. This example will illustrate that the behavior of $\bar{y}$ as an estimator of μ depends on the sample size, n. Later in this chapter, we will illustrate the effect of the shape of the population distribution on the sampling distribution of $\bar{y}$.

EXAMPLE 4.23

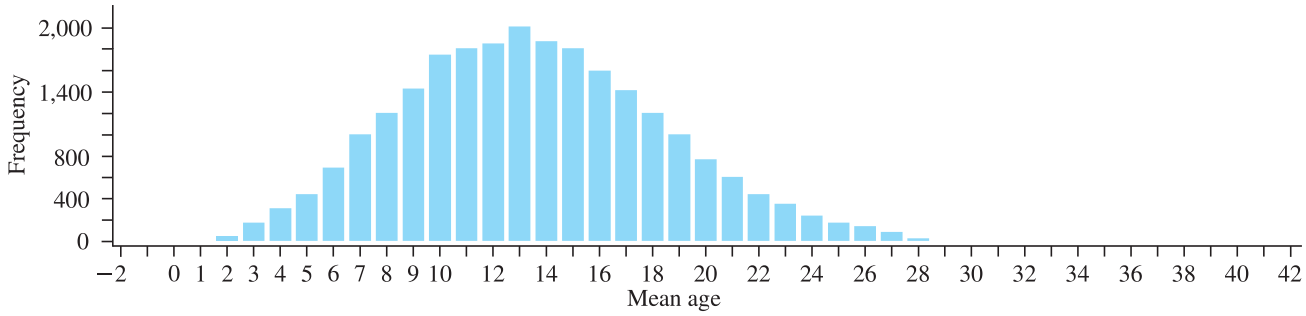
In this example, the population values are known and, hence, we can compute the exact values of the population mean, μ, and population standard deviation, σ. We will then examine the behavior of $\bar{y}$ based on samples of size $n = 5$, 10, and 25 selected from the population. The population consists of 500 pennies from which we compute the age of each penny: Age = 2008 − Date on penny. The histogram of the 500 ages is displayed in Figure 4.20(a). The shape is skewed to the right with a very long right tail. The mean and standard deviation are computed to be $\mu = 13.468$ years and $\sigma = 11.164$ years. In order to generate the sampling distribution of $\bar{y}$ for $n = 5$, we would need to generate all possible samples of size $n = 5$ and then compute the $\bar{y}$ from each of these samples. This would be an enormous task since there are 255,244,687,600 possible samples of size 5 that could be selected from a population of 500 elements. The number of possible samples of size 10 or 25 is so large it makes even the national debt look small. Thus, we will use a computer program to select 25,000 samples of size 5 from the population of 500 pennies. For example, the first sample consists of pennies with ages 4, 12, 26, 16, and 9. The sample mean $\bar{y} = (4 + 12 + 26 + 16 + 9)/5 = 13.4$. We repeat 25,000 times the process of selecting 5 pennies, recording their ages, y_1, y_2, y_3, y_4, y_5, and then computing $\bar{y} = (y_1 + y_2 + y_3 + y_4 + y_5)/5$. The 25,000 values for $\bar{y}$ are then plotted in a frequency histogram, called the *sampling distribution* of $\bar{y}$ for $n = 5$. A similar procedure is followed for samples of size $n = 10$ and $n = 25$. The sampling distributions obtained are displayed in Figures 4.20(b)–(d).

Note that all three sampling distributions have nearly the same central value, approximately 13.5. (See Table 4.11.) The mean values of $\bar{y}$ for the three samples are nearly the same as the population mean, $\mu = 13.468$. In fact, if we had generated all possible samples for all three values of n, the mean of the possible values of $\bar{y}$ would agree exactly with μ.

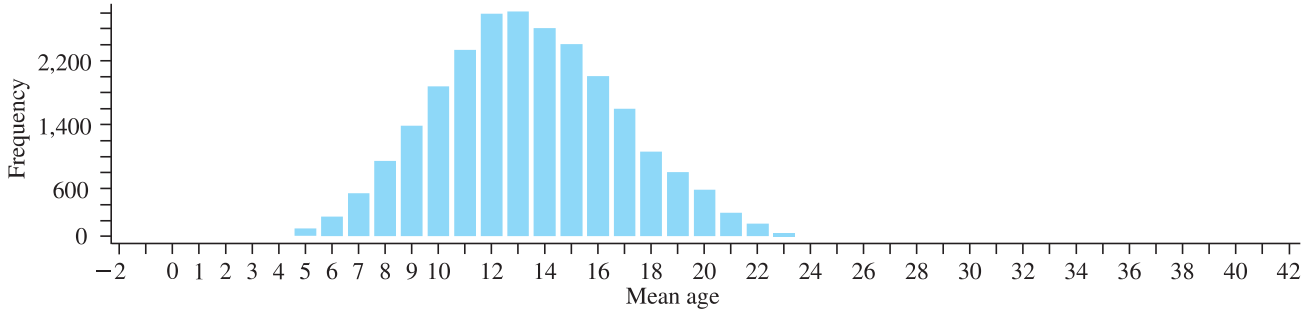
FIGURE 4.20



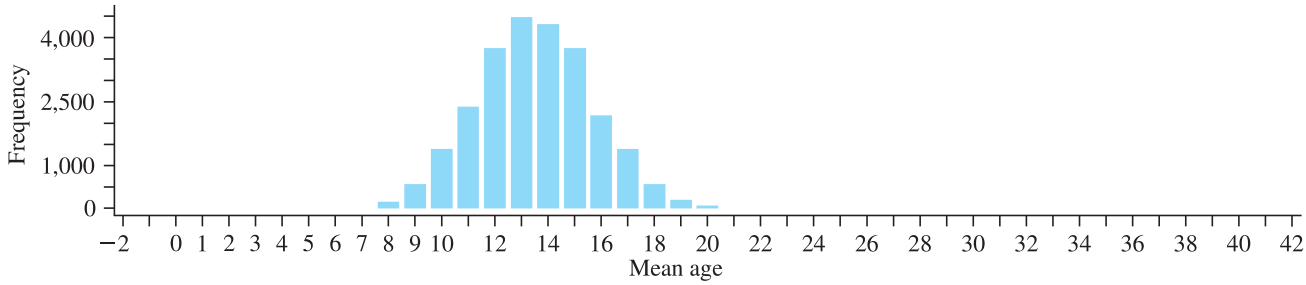
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of $\bar{y}$ for $n = 5$



(c) Sampling distribution of $\bar{y}$ for $n = 10$



(d) Sampling distribution of $\bar{y}$ for $n = 25$

TABLE 4.11
Means and standard deviations for the sampling distributions of $\bar{y}$

Sample Size	Mean of $\bar{y}$	Standard Deviation of $\bar{y}$	$11.1638/\sqrt{n}$
1 (Population)	13.468 (μ)	11.1638 (σ)	11.1638
5	13.485	4.9608	4.9926
10	13.438	3.4926	3.5303
25	13.473	2.1766	2.2328

The next characteristic to notice about the three histograms is their shape. All three are somewhat symmetric in shape, achieving a nearly normal distribution shape when $n = 25$. However, the histogram for $\bar{y}$ based on samples of size $n = 5$ is more spread out than the histogram based on $n = 10$, which, in turn, is more spread out than the histogram based on $n = 25$. When n is small, we are much more likely to obtain a value of $\bar{y}$ far from μ than when n is larger. What causes this increased dispersion in the values of $\bar{y}$? A single extreme y, either large or small relative to μ, in the sample has a greater influence on the size of $\bar{y}$ when n is small than when n is large. Thus, sample means based on small n are less accurate in their estimation of μ than their large-sample counterparts.

Table 4.11 contains summary statistics for the sampling distribution of $\bar{y}$. The sampling distribution of $\bar{y}$ has mean $\mu_{\bar{y}}$ and standard deviation $\sigma_{\bar{y}}$, which are related to the population mean, μ, and standard deviation, σ, by the following relationship:

$$\mu_{\bar{y}} = \mu \qquad \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

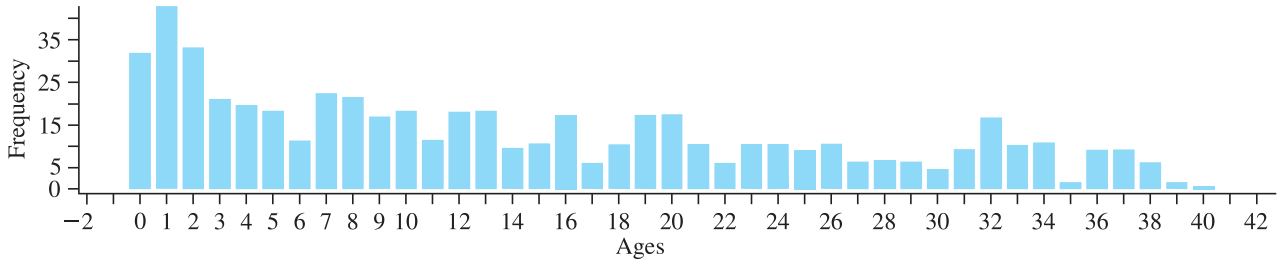
From Table 4.11, we note that the three sampling deviations have means that are approximately equal to the population mean. Also, the three sampling deviations have standard deviations that are approximately equal to $\sigma/\sqrt{n}$. If we had generated all possible values of $\bar{y}$, then the standard deviation of $\bar{y}$ would equal $\sigma/\sqrt{n}$ exactly. This quantity, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, is called the **standard error of $\bar{y}$.**

standard error of $\bar{y}$

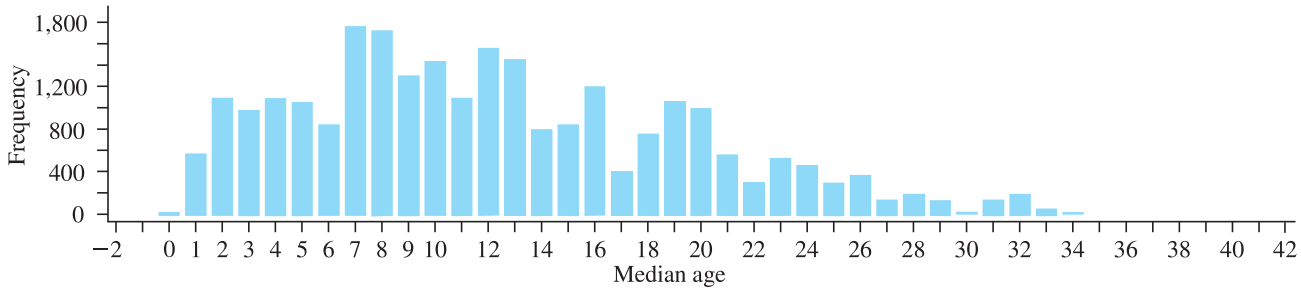
Quite a few of the more common sample statistics, such as the sample median and the sample standard deviation, have sampling distributions that are nearly normal for moderately sized values of n. We can observe this behavior by computing the sample median and sample standard deviation from each of the three sets of 25,000 sample ($n = 5, 10, 25$) selected from the population of 500 pennies. The resulting sampling distributions are displayed in Figures 4.21(a)–(d), for the sample median, and Figures 4.22(a)–(d), for the sample standard deviation. The sampling distribution of both the median and the standard deviation are more highly skewed in comparison to the sampling distribution of the sample mean. In fact, the value of n at which the sampling distributions of the sample median and standard deviation have a nearly normal shape is much larger than the value required for the sample mean. A series of theorems in mathematical statistics called the **Central Limit Theorems** provide theoretical justification for our approximating the true sampling distribution of many sample statistics with the normal distribution. We will discuss one such theorem for the sample mean. Similar theorems exist for the sample median, sample standard deviation, and the sample proportion.

Central Limit Theorems

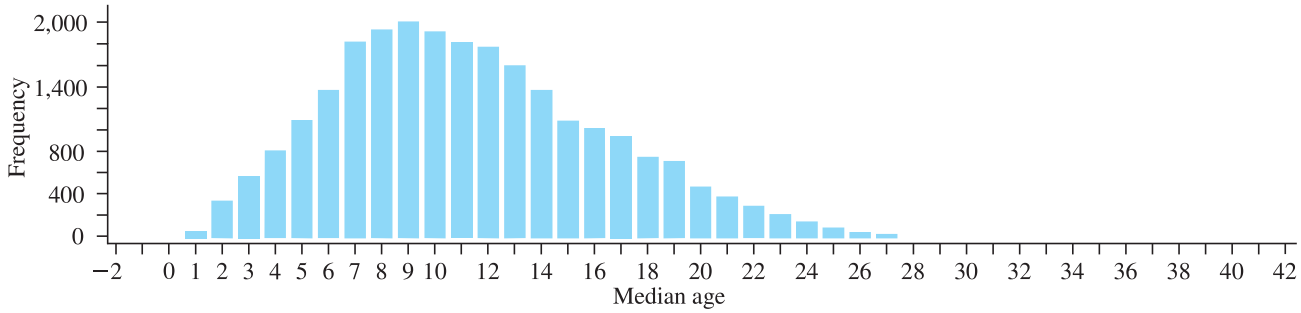
FIGURE 4.21



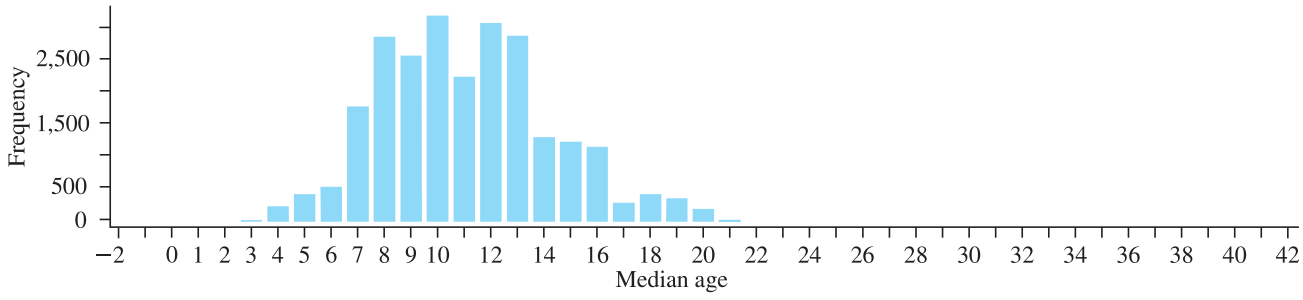
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of median for $n = 5$

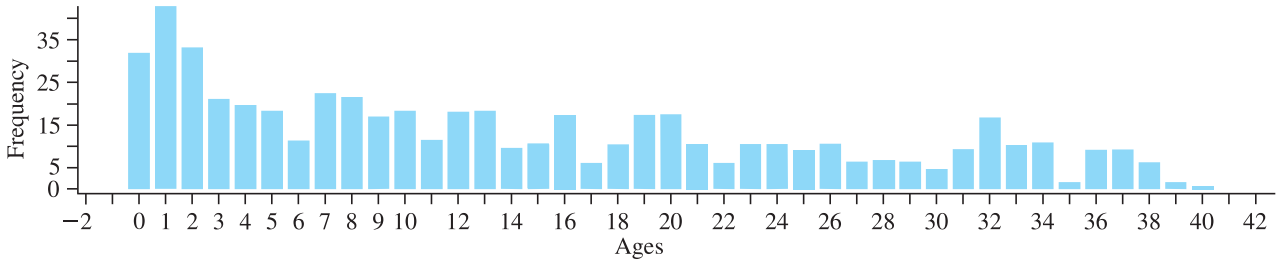


(c) Sampling distribution of median for $n = 10$

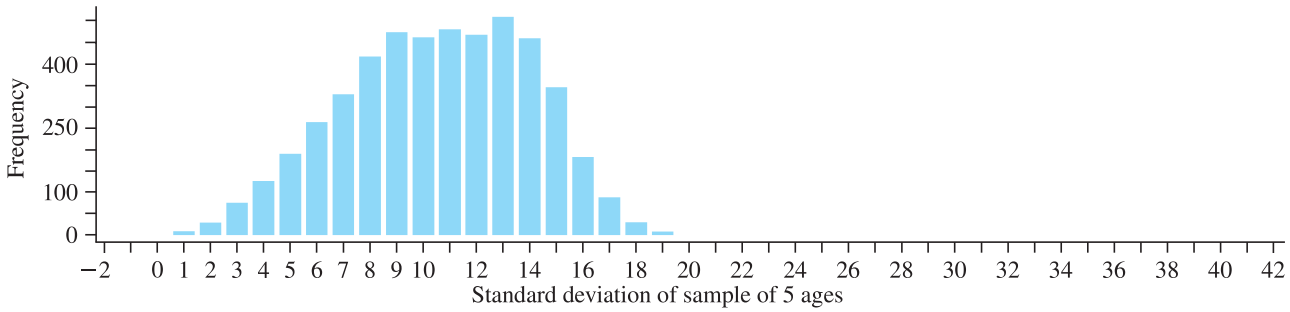


(d) Sampling distribution of median for $n = 25$

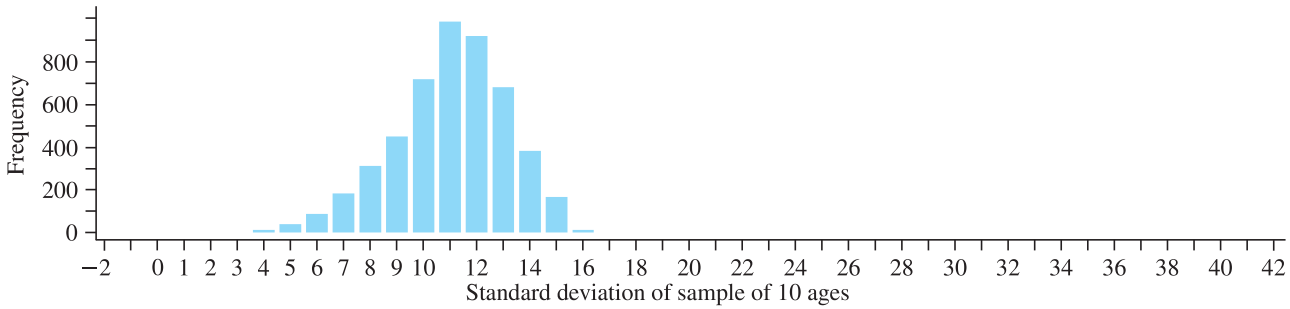
FIGURE 4.22



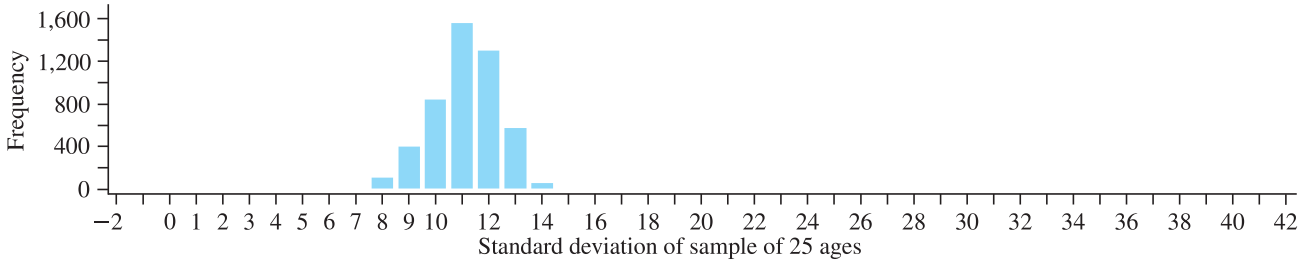
(a) Histogram of ages for 500 pennies



(b) Sampling distribution of standard deviation for $n = 5$



(c) Sampling distribution of standard deviation for $n = 10$



(d) Sampling distribution of standard deviation for $n = 25$

THEOREM 4.1

Central Limit Theorem for $\bar{y}$

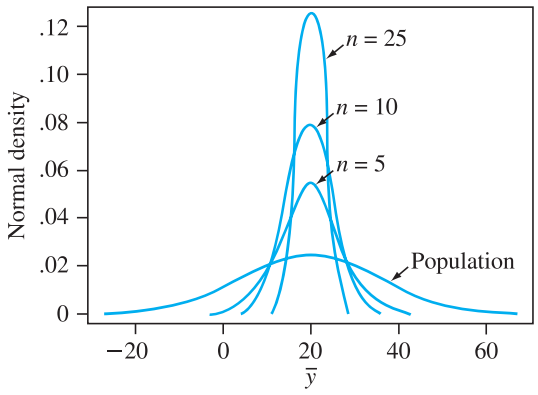
Let $\bar{y}$ denote the sample mean computed from a random sample of n measurements from a population having a mean, μ, and finite standard deviation σ. Let $\mu_{\bar{y}}$ and $\sigma_{\bar{y}}$ denote the mean and standard deviation of the sampling distribution of $\bar{y}$, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\bar{y}} = \mu$
2. $\sigma_{\bar{y}} = \sigma/\sqrt{n}$
3. When n is large, the sampling distribution of $\bar{y}$ will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of $\bar{y}$ is exactly normal for any sample size n.

Figure 4.20 illustrates the Central Limit Theorem. Figure 4.20(a) displays the distribution of the measurements y in the population from which the samples are to be drawn. No specific shape was required for these measurements for the Central Limit Theorem to be validated. Figures 4.20(b)–(d) illustrate the sampling distribution for the sample mean $\bar{y}$ when n is 5, 10, and 25, respectively. We note that even for a very small sample size, $n = 10$, the shape of the sampling distribution of $\bar{y}$ is very similar to that of a normal distribution. This is not true in general. If the population distribution had many extreme values or several modes, the sampling distribution of $\bar{y}$ would require n to be considerably larger in order to achieve a symmetric bell shape.

We have seen that the sample size n has an effect on the shape of the sampling distribution of $\bar{y}$. The shape of the distribution of the population measurements also will affect the shape of the sampling distribution of $\bar{y}$. Figures 4.23 and 4.24 illustrate the effect of the population shape on the shape of the sampling distribution of $\bar{y}$. In Figure 4.23, the population measurements have a normal distribution. The sampling distribution of $\bar{y}$ is *exactly* a normal distribution for all values of n, as is illustrated for $n = 5$, 10, and 25 in Figure 4.23. When the population distribution is nonnormal, as depicted in Figure 4.24, the sampling distribution of $\bar{y}$ will not have a normal shape for small n (see Figure 4.24 with $n = 5$). However, for $n = 10$ and 25, the sampling distributions are nearly normal in shape, as can be seen in Figure 4.24.

FIGURE 4.23
Sampling distribution of $\bar{y}$ for $n = 5, 10, 25$ when sampling from a normal distribution



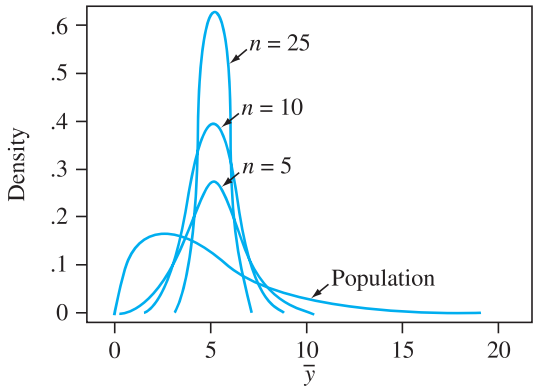


FIGURE 4.24
Sampling distribution of $\bar{y}$ for $n = 5, 10, 25$ when sampling from a skewed distribution

It is very unlikely that the exact shape of the population distribution will be known. Thus, the exact shape of the sampling distribution of $\bar{y}$ will not be known either. The important point to remember is that the sampling distribution of $\bar{y}$ will be approximately normally distributed with a mean $\mu_{\bar{y}} = \mu$, the population mean, and a standard deviation $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. The approximation will be more precise as n, the sample size for each sample, increases and as the shape of the population distribution becomes more like the shape of a normal distribution.

An obvious question is, How large should the sample size be for the Central Limit Theorem to hold? Numerous simulation studies have been conducted over the years and the results of these studies suggest that, in general, the Central Limit Theorem holds for $n > 30$. However, one should not apply this rule blindly. If the population is heavily skewed, the sampling distribution for $\bar{y}$ will still be skewed even for $n > 30$. On the other hand, if the population is symmetric, the Central Limit Theorem holds for $n < 30$.

Therefore, take a look at the data. If the sample histogram is clearly skewed, then the population will also probably be skewed. Consequently, a value of n much higher than 30 may be required to have the sampling distribution of $\bar{y}$ be approximately normal. Any inference based on the normality of $\bar{y}$ for $n \leq 30$ under this condition should be examined carefully.

EXAMPLE 4.24

A person visits her doctor with concerns about her blood pressure. If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure and medication may be prescribed. A patient's blood pressure readings often have a considerable variation during a given day. Suppose a patient's systolic blood pressure readings during a given day have a normal distribution with a mean $\mu = 160$ mm mercury and a standard deviation $\sigma = 20$ mm.

a. What is the probability that a single blood pressure measurement will fail to detect that the patient has high blood pressure?
b. If five blood pressure measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 150 and hence fail to indicate that the patient has high blood pressure?
c. How many measurements would be required in a given day so that there is at most 1% probability of failing to detect that the patient has high blood pressure?

Solution Let y be the blood pressure measurement of the patient. y has a normal distribution with $\mu = 160$ and $\sigma = 20$.

a. $P(\text{measurement fails to detect high pressure}) = P(y \le 150) = P(z \le \frac{150 - 160}{20}) = P(z \le -0.5) = .3085$. Thus, there is over a 30% chance of failing to detect that the patient has high blood pressure if only a single measurement is taken.

b. Let $\bar{y}$ be the average blood pressure of the five measurements. Then, $\bar{y}$ has a normal distribution with $\mu = 160$ and $\sigma = 20/\sqrt{5} = 8.944$.

$$P(\bar{y} \le 150) = P\left(z \le \frac{150 - 160}{8.944}\right) = P(z \le -1.12) = .1314$$

Therefore, by using the average of five measurements, the chance of failing to detect the patient has high blood pressure has been reduced from over 30% to about 13%.

c. We need to determine the sample size n such that $P(\bar{y} < 150) \le .01$. Now, $P(\bar{y} < 150) = P(z \le \frac{150 - 160}{20/\sqrt{n}})$. From the normal tables, we have $P(z \le -2.326) = .01$, therefore, $\frac{150 - 160}{20/\sqrt{n}} = -2.326$. Solving for n, yields $n = 21.64$. It would require at least 22 measurements in order to achieve the goal of at most a 1% chance of failing to detect high blood pressure.

As demonstrated in Figures 4.21 and 4.22, the Central Limit Theorem can be extended to many different sample statistics. The form of the Central Limit Theorem for the sample median and sample standard deviation is somewhat more complex than for the sample mean. Many of the statistics that we will encounter in later chapters will be either averages or sums of variables. The Central Limit Theorem for sums can be easily obtained from the Central Limit Theorem for the sample mean. Suppose we have a random sample of n measurements, $y_1, \ldots, y_n$, from a population and let $\Sigma y = y_1 + \cdots + y_n$.

THEOREM 4.2

Central Limit Theorem for Σy

Let Σy denote the sum of a random sample of n measurements from a population having a mean μ and finite standard deviation σ. Let $\mu_{\Sigma y}$ and $\sigma_{\Sigma y}$ denote the mean and standard deviation of the sampling distribution of Σy, respectively. Based on repeated random samples of size n from the population, we can conclude the following:

1. $\mu_{\Sigma y} = n\mu$
2. $\sigma_{\Sigma y} = \sqrt{n}\sigma$
3. When n is large, the sampling distribution of Σy will be approximately normal (with the approximation becoming more precise as n increases).
4. When the population distribution is normal, the sampling distribution of Σy is exactly normal for any sample size n.

Usually, a sample statistic is used as an estimate of a population parameter. For example, a sample mean $\bar{y}$ can be used to estimate the population mean μ from which the sample was selected. Similarly, a sample median and sample standard deviation estimate the corresponding population median and standard deviation.

The sampling distribution of a sample statistic is then used to determine how accurate the estimate is likely to be. In Example 4.22, the population mean μ is known to be 6.5. Obviously, we do not know μ in any practical study or experiment. However, we can use the sampling distribution of $\bar{y}$ to determine the probability that the value of $\bar{y}$ for a random sample of $n = 2$ measurements from the population will be more than three units from μ. Using the data in Example 4.22, this probability is

$$P(2.5) + P(3) + P(10) + P(10.5) = \frac{4}{45}$$

In general, we would use the normal approximation from the Central Limit Theorem in making this calculation because the sampling distribution of a sample statistic is seldom known. This type of calculation will be developed in Chapter 5. Since a sample statistic is used to make inferences about a population parameter, the sampling distribution of the statistic is crucial in determining the accuracy of the inference.

interpretations of a sampling distribution

Sampling distributions can be **interpreted** in at least two ways. One way uses the long-run relative frequency approach. Imagine taking repeated samples of a fixed size from a given population and calculating the value of the sample statistic for each sample. In the long run, the relative frequencies for the possible values of the sample statistic will approach the corresponding sampling distribution probabilities. For example, if one took a large number of samples from the population distribution corresponding to the probabilities of Example 4.22 and, for each sample, computed the sample mean, approximately 9% would have $\bar{y} = 5.5$.

The other way to interpret a sampling distribution makes use of the classical interpretation of probability. Imagine listing all possible samples that could be drawn from a given population. The probability that a sample statistic will have a particular value (say, that $\bar{y} = 5.5$) is then the proportion of all possible samples that yield that value. In Example 4.22, $P(\bar{y} = 5.5) = 4/45$ corresponds to the fact that 4 of the 45 samples have a sample mean equal to 5.5. Both the repeated-sampling and the classical approach to finding probabilities for a sample statistic are legitimate.

In practice, though, a sample is taken only once, and only one value of the sample statistic is calculated. A sampling distribution is not something you can see in practice; it is not an empirically observed distribution. Rather, it is a theoretical concept, a set of probabilities derived from assumptions about the population and about the sampling method.

There's an unfortunate similarity between the phrase "sampling distribution," meaning the theoretically derived probability distribution of a statistic, and the phrase "sample distribution," which refers to the histogram of individual values actually observed in a particular sample. The two phrases mean very different things.

sample histogram

To avoid confusion, we will refer to the distribution of sample values as the **sample histogram** rather than as the sample distribution.

4.13 Normal Approximation to the Binomial

A binomial random variable y was defined earlier to be the number of successes observed in n independent trials of a random experiment in which each trial resulted in either a success (S) or a failure (F) and $P(S) = \pi$ for all n trials. We will now demonstrate how the Central Limit Theorem for sums enables us to calculate probabilities for a binomial random variable by using an appropriate normal curve as an approximation to the binomial distribution. We said in Section 4.8 that probabilities associated with values of y can be computed for a binomial experiment for