

Chapter 7

Random Variables and Probability Distributions

© Walter Bibikow/Getty Images



This chapter is the first of two chapters that together link the basic ideas of probability explored in Chapter 6 with the techniques of statistical inference. Chapter 6 used probability to describe the long-run relative frequency of occurrence of various types of outcomes. In this chapter we introduce probability models that can be used to describe the distribution of values of a variable. In Chapter 8, we will see how these same probability models can be used to describe the behavior of sample statistics. Such models are essential if we are to reach conclusions based on a sample from the population of interest.

In a chance experiment, we often focus on some numerical aspect of the outcome. An environmental scientist who obtains an air sample from a specified location might be especially concerned with the concentration of ozone (a major constituent of smog). A quality control inspector who must decide whether to accept a large shipment of components may base the decision on the number of defective components in a group of 20 components randomly selected from the shipment.

Before selection of the air sample, the value of the ozone concentration is uncertain. Similarly, the number of defective components among the 20 selected might be any whole number between 0 and 20. Because the value of a variable quantity such as ozone concentration or number of defective components is subject to uncertainty, such variables are called *random variables*.

Improve your understanding and save time! Visit www.cengage.com/login where you will find:

- Step-by-step instructions for MINITAB, Excel, TI-83, SPSS, and JMP
- Video solutions to selected exercises
- Data sets available for selected examples and exercises
- Exam-prep pre-tests that build a Personalized Learning Plan based on your results so that you know exactly what to study
- Help from a live statistics tutor 24 hours a day

In this chapter we begin by distinguishing between discrete and continuous numerical variables. We show how variation in both discrete and continuous numerical variables can be described by a probability distribution; this distribution can then be used to make probability statements about values of the random variable. Special emphasis is given to three commonly encountered probability distributions: the binomial, geometric, and normal distributions.

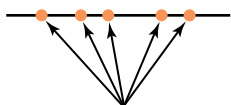
7.1 Random Variables

In most chance experiments, an investigator focuses attention on one or more variable quantities. For example, consider a management consultant who is studying the operation of a supermarket. The chance experiment might involve randomly selecting a customer leaving the store. One interesting numerical variable might be the number of items x purchased by the customer. Possible values of this variable are 0 (a frustrated customer), 1, 2, 3, and so on. Until a customer is selected and the number of items counted, the value of x is uncertain. Another variable of potential interest might be the time y (minutes) spent in a checkout line. One possible value of y is 3.0 min and another is 4.0 min, but *any* other number between 3.0 and 4.0 is also a possibility. Whereas possible values of x are isolated points on the number line, possible y values form an entire interval (a continuum) on the number line.

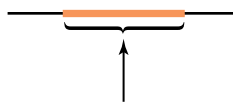
DEFINITION

A numerical variable whose value depends on the outcome of a chance experiment is called a **random variable**. A random variable associates a numerical value with each outcome of a chance experiment.

A random variable is **discrete** if its set of possible values is a collection of isolated points on the number line. The variable is **continuous** if its set of possible values includes an entire interval on the number line.



Possible values of a discrete random variable



Possible values of a continuous random variable

Figure 7.1 Two different types of random variables.

We use lowercase letters, such as x and y , to represent random variables.*

Figure 7.1 shows a set of possible values for each type of random variable. In practice, a discrete random variable almost always arises in connection with counting (e.g., the number of items purchased, the number of gas pumps in use, or the number of broken eggs in a carton). A continuous random variable is one whose value is typically obtained by measurement (temperature in a freezer compartment, weight of a pineapple, amount of time spent in the store, etc.). Because there is a limit to the accuracy of any measuring instrument, such as a watch or a scale, it may seem that any variable should be regarded as discrete. However, when there is a large number of closely spaced values, the variable's behavior is most easily studied by conceptualizing it as continuous. (Doing so allows the use of calculus to solve some types of probability problems.)

*In some books, uppercase letters are used to name random variables, with lowercase letters representing a particular value that the variable might assume. We have opted to use a simpler and less formal notation.

Example 7.1 Car Sales

Consider an experiment in which the type of car, new (N) or used (U), chosen by each of three successive customers at a discount car dealership is noted. Define a random variable x by

$$x = \text{number of customers purchasing a new car}$$

The experimental outcome in which the first and third customers purchase a new car and the second customer purchases a used car can be abbreviated NUN. The associated x value is 2, because two of the three customers selected a new car. Similarly, the x value for the outcome NNN (all three purchase a new car) is 3. We display each of the eight possible experimental outcomes and the corresponding value of x in the following table:

| | | | | | | | | |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|
| Outcome | UUU | NUU | UNU | UUN | NNU | NUN | UNN | NNN |
| x value | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 |

There are only four possible x values—0, 1, 2, and 3—and these are isolated points on the number line. Thus, x is a discrete random variable. ■

In some situations, the random variable of interest is discrete, but the number of possible values is not finite. This is illustrated in Example 7.2.

Example 7.2 This Could Be a Long Game . . .

Two friends agree to play a game that consists of a sequence of trials. The game continues until one player wins two trials in a row. One random variable of interest might be

$$x = \text{number of trials required to complete the game}$$

Let A denote a win for Player 1 and B denote a win for Player 2. The simplest possible experimental outcomes are AA (the case in which Player 1 wins the first two trials and the game ends) and BB (the case in which Player 2 wins the first two trials). With either of these two outcomes, $x = 2$. There are also two outcomes for which $x = 3$: ABB and BAA. Some other possible outcomes and associated x values are

| Outcomes | x value |
|------------------------|-----------|
| AA, BB | 2 |
| BAA, ABB | 3 |
| ABAA, BABB | 4 |
| ABABB, BABAA | 5 |
| ⋮ | ⋮ |
| ABABABABAA, BABABABABB | 10 |

and so on.

Any positive integer that is at least 2 is a possible value. Because the values 2, 3, 4, . . . are isolated points on the number line (x is determined by counting), x is a discrete random variable even though there is no upper limit to the number of possible values.

Example 7.3 Stress

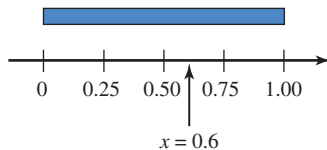


Figure 7.2 The bar for Example 7.3 and the outcome $x = 0.6$.

In an engineering stress test, pressure is applied to a thin 1-ft-long bar until the bar snaps. The precise location where the bar will snap is uncertain. Let x be the distance from the left end of the bar to the break. Then $x = 0.25$ is one possibility, $x = 0.9$ is another, and in fact any number between 0 and 1 is a possible value of x . (Figure 7.2 shows the case of the outcome $x = 0.6$.) This set of possible values is an entire interval on the number line, so x is a continuous random variable.

Even though in practice we may be able to measure the distance only to the nearest tenth of an inch or hundredth of an inch, the *actual* distance could be any number between 0 and 1. So, even though the recorded values might be rounded because of the accuracy of the measuring instrument, the variable is still continuous.

In data analysis, random variables often arise in the context of summarizing sample data when a sample is selected from some population. This is illustrated in Example 7.4.

Example 7.4 College Plans

Suppose that a counselor plans to select a random sample of 50 seniors at a large high school and to ask each student in the sample whether he or she plans to attend college after graduation. The process of sampling is a chance experiment. The sample space for this experiment consists of all the different possible random samples of size 50 that might result (there is a very large number of these), and for simple random sampling each of these outcomes is equally likely. Let

$$x = \text{number of successes in the sample}$$

where a success in this instance is defined as a student who plans to attend college. Then x is a random variable, because it associates a numerical value with each of the possible outcomes (random samples) that might occur. Possible values of x are 0, 1, 2, . . . , 50, and x is a discrete random variable.

Exercises 7.1–7.7

7.1 State whether each of the following random variables is discrete or continuous:

- The number of defective tires on a car
- The body temperature of a hospital patient

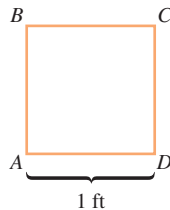
- The number of pages in a book
- The number of draws (with replacement) from a deck of cards until a heart is selected
- The lifetime of a lightbulb

7.2 Classify each of the following random variables as either discrete or continuous:

- The fuel efficiency (mpg) of an automobile
- The amount of rainfall at a particular location during the next year
- The distance that a person throws a baseball
- The number of questions asked during a 1-hr lecture
- The tension (in pounds per square inch) at which a tennis racket is strung
- The amount of water used by a household during a given month
- The number of traffic citations issued by the highway patrol in a particular county on a given day

7.3 Starting at a particular time, each car entering an intersection is observed to see whether it turns left (L) or right (R) or goes straight ahead (S). The experiment terminates as soon as a car is observed to go straight. Let y denote the number of cars observed. What are possible y values? List five different outcomes and their associated y values.

7.4 A point is randomly selected from the interior of a square, as pictured:



Let x denote the distance from the lower left-hand corner A of the square to the selected point. What are possible values of x ? Is x a discrete or a continuous variable?

7.5 A point is randomly selected on the surface of a lake that has a maximum depth of 100 ft. Let y be the depth of the lake at the randomly chosen point. What are possible values of y ? Is y discrete or continuous?

7.6 A person stands at the corner marked A of the square pictured in Exercise 7.4 and tosses a coin. If it lands heads up, the person moves one corner clockwise, to B . If the coin lands tails up, the person moves one corner counterclockwise, to D . This process is then repeated until the person arrives back at A . Let y denote the number of coin tosses. What are possible values of y ? Is y discrete or continuous?

7.7 A box contains four slips of paper marked 1, 2, 3, and 4. Two slips are selected without replacement. List the possible values for each of the following random variables:

- x = sum of the two numbers
- y = difference between the first and second numbers
- z = number of slips selected that show an even number
- w = number of slips selected that show a 4

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.2

Probability Distributions for Discrete Random Variables

The probability distribution for a random variable is a model that describes the long-run behavior of the variable. For example, suppose that the Department of Animal Regulation in a particular county is interested in studying the variable x = number of licensed dogs or cats for a household. County regulations prohibit more than five dogs or cats per household. If we consider the chance experiment of randomly selecting a household in this county, then x is a discrete random variable because it associates a numerical value (0, 1, 2, 3, 4, or 5) with each of the possible outcomes (households) in the sample space. Although we know what the possible values for x are, it would also be useful to know how this variable behaves in repeated observation. What would be the most common value? What proportion of the time would $x = 5$ be observed? $x = 3$? A probability distribution provides this type of information about the long-run behavior of a random variable.

DEFINITION

The **probability distribution of a discrete random variable** x gives the probability associated with each possible x value. Each probability is the limiting relative frequency of occurrence of the corresponding x value when the chance experiment is repeatedly performed.

Common ways to display a probability distribution for a discrete random variable are a table, a probability histogram, or a formula.

If one possible value of x is 2, we often write $p(2)$ in place of $P(x = 2)$. Similarly, $p(5)$ denotes the probability that $x = 5$, and so on.

Example 7.5 Hot Tub Models

Suppose that each of four randomly selected customers purchasing a hot tub at a certain store chooses either an electric (E) or a gas (G) model. Assume that these customers make their choices independently of one another and that 40% of all customers select an electric model. This implies that for any particular one of the four customers, $P(E) = .4$ and $P(G) = .6$. One possible experimental outcome is EGGE, where the first and fourth customers select electric models and the other two choose gas models. Because the customers make their choices independently, the multiplication rule for independent events implies that

$$\begin{aligned} P(\text{EGGE}) &= P(\text{1st chooses E and 2nd chooses G and 3rd chooses G and 4th chooses E}) \\ &= P(E)P(G)P(G)P(E) \\ &= (.4)(.6)(.6)(.4) \\ &= .0576 \end{aligned}$$

Similarly,

$$\begin{aligned} P(\text{EGEG}) &= P(E)P(G)P(E)P(G) \\ &= (.4)(.6)(.4)(.6) \\ &= .0576 \quad (\text{identical to } P(\text{EGGE})) \end{aligned}$$

and

$$P(\text{GGGE}) = (.6)(.6)(.6)(.4) = .0864$$

The number among the four customers who purchase an electric hot tub is a random variable. Let

x = the number of electric hot tubs purchased by the four customers

Table 7.1 displays the 16 possible experimental outcomes, the probability of each outcome, and the value of the random variable x that is associated with each outcome.

The probability distribution of x is easily obtained from this information. Consider the smallest possible x value, 0. The only outcome for which $x = 0$ is GGGG, so

$$p(0) = P(x = 0) = P(\text{GGGG}) = .1296$$

Table 7.1 Outcomes and Probabilities for Example 7.5

| Outcome | Probability | x Value | Outcome | Probability | x Value |
|---------|-------------|---------|---------|-------------|---------|
| GGGG | .1296 | 0 | GEEG | .0576 | 2 |
| EGGG | .0864 | 1 | GEGE | .0576 | 2 |
| GEGG | .0864 | 1 | GGEE | .0576 | 2 |
| GGEG | .0864 | 1 | GEEE | .0384 | 3 |
| GGGE | .0864 | 1 | EGEE | .0384 | 3 |
| EEGG | .0576 | 2 | EEGE | .0384 | 3 |
| EGEG | .0576 | 2 | EEEG | .0384 | 3 |
| EGGE | .0576 | 2 | EEEE | .0256 | 4 |

There are four different outcomes for which $x = 1$, so $p(1)$ results from summing the four corresponding probabilities:

$$\begin{aligned}
 p(1) &= P(x = 1) = P(\text{EGGG or GEGG or GGEG or GGGE}) \\
 &= P(\text{EGGG}) + P(\text{GEGG}) + P(\text{GGEG}) + P(\text{GGGE}) \\
 &= .0864 + .0864 + .0864 + .0864 \\
 &= 4(.0864) \\
 &= .3456
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 p(2) &= P(\text{EEGG}) + \cdots + P(\text{GGEE}) = 6(.0576) = .3456 \\
 p(3) &= 4(.0384) = .1536 \\
 p(4) &= .0256
 \end{aligned}$$

The probability distribution of x is summarized in the following table:

| | | | | | |
|-------------------------------|-------|-------|-------|-------|-------|
| x Value | 0 | 1 | 2 | 3 | 4 |
| $p(x)$ = Probability of Value | .1296 | .3456 | .3456 | .1536 | .0256 |

To interpret $p(3) = .1536$, think of performing the chance experiment repeatedly, each time with a new group of four customers. In the long run, 15.36% of these groups will have exactly three customers purchasing an electric hot tub. The probability distribution can be used to determine probabilities of various events involving x . For example, the probability that at least two of the four customers choose electric models is

$$\begin{aligned}
 P(x \geq 2) &= P(x = 2 \text{ or } x = 3 \text{ or } x = 4) \\
 &= p(2) + p(3) + p(4) \\
 &= .5248
 \end{aligned}$$

Thus, in the long run, 52.48% of the time a group of four hot tub purchasers will include at least two who select electric models. ■

A probability distribution table for a discrete variable shows the possible x values and also $p(x)$ for each possible x value. Because $p(x)$ is a probability, it must be a number between 0 and 1, and because the probability distribution lists all possible x val-

ues, the sum of all the $p(x)$ values must equal 1. These properties of discrete probability distributions are summarized in the following box.

Properties of Discrete Probability Distributions

1. For every possible x value, $0 \leq p(x) \leq 1$.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

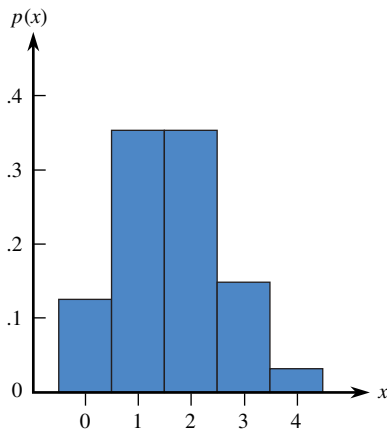


Figure 7.3 Probability histogram for the distribution of Example 7.5.

A pictorial representation of a discrete probability distribution is called a *probability histogram*. The picture has a rectangle centered above each possible value of x , and the area of each rectangle is the probability of the corresponding value. Figure 7.3 displays the probability histogram for the probability distribution of Example 7.5.

In Example 7.5, the probability distribution was derived by starting with a simple experimental situation and applying basic probability rules. When a derivation from fundamental probabilities is not possible because of the complexity of the experimental situation, an investigator often conjectures a probability distribution consistent with empirical evidence and prior knowledge. It must also be consistent with rules of probability. Specifically,

1. $p(x) \geq 0$ for every x value.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

Example 7.6 Automobile Defects

A consumer organization that evaluates new automobiles customarily reports the number of major defects on each car examined. Let x denote the number of major defects on a randomly selected car of a certain type. A large number of automobiles were evaluated, and a probability distribution consistent with these observations is:

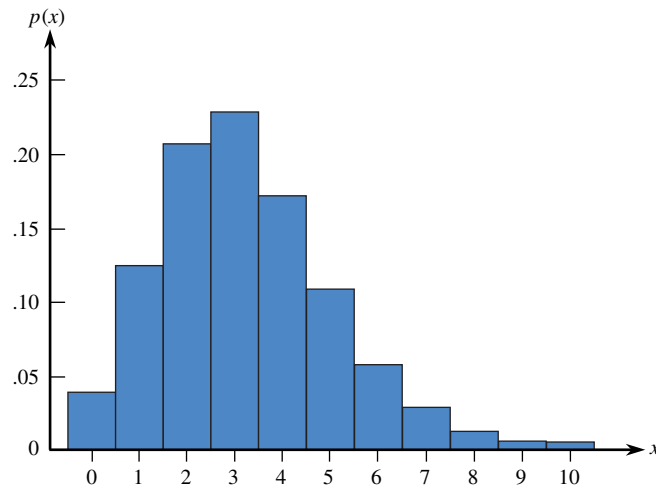
| | | | | | | | | | | | |
|--------|------|------|------|------|------|------|------|------|------|------|------|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $p(x)$ | .041 | .010 | .209 | .223 | .178 | .114 | .061 | .028 | .011 | .004 | .001 |

The corresponding probability histogram appears in Figure 7.4. The probabilities in this distribution reflect the organization's experience. For example, $p(3) = .223$ indicates that 22.3% of new automobiles had 3 major defects. The probability that the number of major defects is between 2 and 5 inclusive is

$$P(2 \leq x \leq 5) = p(2) + p(3) + p(4) + p(5) = .724$$

If car after car of this type were examined, in the long run, 72.4% would have 2, 3, 4, or 5 major defects.

Figure 7.4 Probability histogram for the distribution of the number of major defects on a randomly selected car.



We have seen examples in which the probability distribution of a discrete random variable has been given as a table or as a probability (relative frequency) histogram. It is also possible to give a formula that allows calculation of the probability for each possible value of the random variable. Examples of this approach are given in Section 7.5.

Exercises 7.8–7.19

7.8 Let x be the number of courses for which a randomly selected student at a certain university is registered. The probability distribution of x appears in the following table:

| | | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|-----|
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $p(x)$ | .02 | .03 | .09 | .25 | .40 | .16 | .05 |

- What is $P(x = 4)$?
- What is $P(x \leq 4)$?
- What is the probability that the selected student is taking at most five courses?
- What is the probability that the selected student is taking at least five courses? more than five courses?
- Calculate $P(3 \leq x \leq 6)$ and $P(3 < x < 6)$. Explain in words why these two probabilities are different.

7.9 ▼ Let y denote the number of broken eggs in a randomly selected carton of one dozen eggs. Suppose that the probability distribution of y is as follows:

| | | | | | |
|--------|-----|-----|-----|-----|---|
| y | 0 | 1 | 2 | 3 | 4 |
| $p(y)$ | .65 | .20 | .10 | .04 | ? |

- Only y values of 0, 1, 2, 3, and 4 have positive probabilities. What is $p(4)$?
- How would you interpret $p(1) = .20$?
- Calculate $P(y \leq 2)$, the probability that the carton contains at most two broken eggs, and interpret this probability.
- Calculate $P(y < 2)$, the probability that the carton contains *fewer than* two broken eggs. Why is this smaller than the probability in Part (c)?

- e. What is the probability that the carton contains exactly 10 unbroken eggs?
- f. What is the probability that at least 10 eggs are unbroken?

7.10 A restaurant has four bottles of a certain wine in stock. Unbeknownst to the wine steward, two of these bottles (Bottles 1 and 2) are bad. Suppose that two bottles are ordered, and let x be the number of good bottles among these two.

- a. One possible experimental outcome is (1,2) (Bottles 1 and 2 are the ones selected) and another is (2,4). List all possible outcomes.
- b. Assuming that the two bottles are randomly selected from among the four, what is the probability of each outcome in Part (a)?
- c. The value of x for the (1,2) outcome is 0 (neither selected bottle is good), and $x = 1$ for the outcome (2,4). Determine the x value for each possible outcome. Then use the probabilities in Part (b) to determine the probability distribution of x .

7.11 Airlines sometimes overbook flights. Suppose that for a plane with 100 seats, an airline takes 110 reservations. Define the variable x as the number of people who actually show up for a sold-out flight. From past experience, the probability distribution of x is given in the following table:

| | | | | | | | | |
|--------|-----|-----|-----|------|------|------|-------|-------|
| x | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 |
| $p(x)$ | .05 | .10 | .12 | .14 | .24 | .17 | .06 | .04 |
| x | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 |
| $p(x)$ | .03 | .02 | .01 | .005 | .005 | .005 | .0037 | .0013 |

- a. What is the probability that the airline can accommodate everyone who shows up for the flight?
- b. What is the probability that not all passengers can be accommodated?
- c. If you are trying to get a seat on such a flight and you are number 1 on the standby list, what is the probability that you will be able to take the flight? What if you are number 3?

7.12 Suppose that a computer manufacturer receives computer boards in lots of five. Two boards are selected from each lot for inspection. We can represent possible outcomes of the selection process by pairs. For example, the pair (1,2) represents the selection of Boards 1 and 2 for inspection.

- a. List the 10 different possible outcomes.
- b. Suppose that Boards 1 and 2 are the only defective boards in a lot of five. Two boards are to be chosen at ran-

dom. Define x to be the number of defective boards observed among those inspected. Find the probability distribution of x .

7.13 Simulate the chance experiment described in Exercise 7.12 using five slips of paper, with two marked *defective* and three marked *nondefective*. Place the slips in a box, mix them well, and draw out two. Record the number of defective boards. Replace the slips and repeat until you have 50 observations on the variable x . Construct a relative frequency distribution for the 50 observations, and compare this with the probability distribution obtained in Exercise 7.12.

7.14 Of all airline flight requests received by a certain discount ticket broker, 70% are for domestic travel (D) and 30% are for international flights (I). Let x be the number of requests among the next three requests received that are for domestic flights. Assuming independence of successive requests, determine the probability distribution of x . (Hint: One possible outcome is DID, with the probability $(.7)(.3)(.7) = .147$.)

7.15 Suppose that 20% of all homeowners in an earthquake-prone area of California are insured against earthquake damage. Four homeowners are selected at random; let x denote the number among the four who have earthquake insurance.

- a. Find the probability distribution of x . (Hint: Let S denote a homeowner who has insurance and F one who does not. Then one possible outcome is SFSS, with probability $(.2)(.8)(.2)(.2)$ and associated x value of 3. There are 15 other outcomes.)
- b. What is the most likely value of x ?
- c. What is the probability that at least two of the four selected homeowners have earthquake insurance?

7.16 A box contains five slips of paper, marked \$1, \$1, \$1, \$10, and \$25. The winner of a contest selects two slips of paper at random and then gets the larger of the dollar amounts on the two slips. Define a random variable w by $w =$ amount awarded. Determine the probability distribution of w . (Hint: Think of the slips as numbered 1, 2, 3, 4, and 5, so that an outcome of the experiment consists of two of these numbers.)

7.17 Components coming off an assembly line are either free of defects (S, for success) or defective (F, for failure). Suppose that 70% of all such components are defect-free.

Components are independently selected and tested one by one. Let y denote the number of components that must be tested until a defect-free component is obtained.

- What is the smallest possible y value, and what experimental outcome gives this y value? What is the second smallest y value, and what outcome gives rise to it?
- What is the set of all possible y values?
- Determine the probability of each of the five smallest y values. You should see a pattern that leads to a simple formula for $p(y)$, the probability distribution of y .

7.18 A contractor is required by a county planning department to submit anywhere from one to five forms (depending on the nature of the project) in applying for a building permit. Let y be the number of forms required of the next applicant. The probability that y forms are required is known to be proportional to y ; that is, $p(y) = ky$ for $y = 1, \dots, 5$.

- What is the value of k ? (Hint: $\sum p(y) = 1$.)
- What is the probability that at most three forms are required?

- What is the probability that between two and four forms (inclusive) are required?
- Could $p(y) = y^2/50$ for $y = 1, 2, 3, 4, 5$ be the probability distribution of y ? Explain.

7.19 A library subscribes to two different weekly news magazines, each of which is supposed to arrive in Wednesday's mail. In actuality, each one could arrive on Wednesday (W), Thursday (T), Friday (F), or Saturday (S). Suppose that the two magazines arrive independently of one another and that for each magazine $P(W) = .4$, $P(T) = .3$, $P(F) = .2$, and $P(S) = .1$. Define a random variable y by $y =$ the number of days beyond Wednesday that it takes for both magazines to arrive. For example, if the first magazine arrives on Friday and the second magazine arrives on Wednesday, then $y = 2$, whereas $y = 1$ if both magazines arrive on Thursday. Obtain the probability distribution of y . (Hint: Draw a tree diagram with two generations of branches, the first labeled with arrival days for Magazine 1 and the second for Magazine 2.)

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.3

Probability Distributions for Continuous Random Variables

A continuous random variable is one that has as its set of possible values an entire interval on the number line. An example is the weight x (in pounds) of a newborn child. Suppose for the moment that weight is recorded only to the nearest pound. Then possible x values are whole numbers, such as 4 or 9. The probability distribution can be pictured as a probability histogram in which the area of each rectangle is the probability of the corresponding weight value. The total area of all the rectangles is 1, and the probability that a weight (to the nearest pound) is between two values, such as 6 and 8, is the sum of the corresponding rectangular areas. Figure 7.5(a) illustrates this.

Now suppose that weight is measured to the nearest tenth of a pound. There are many more possible weight values than before, such as 5.0, 5.1, 5.7, 7.3, and 8.9. As shown in Figure 7.5(b), the rectangles in the probability histogram are much narrower, and this histogram has a much smoother appearance than the first one. Again, this histogram can be drawn so that the area of each rectangle equals the corresponding probability, and the total area of all the rectangles is 1.

Figure 7.5(c) shows what happens as weight is measured to a greater and greater degree of accuracy. The sequence of probability histograms approaches a smooth curve. The curve cannot go below the horizontal measurement scale, and the total area under the curve is 1 (because this is true of every probability histogram). The probability that

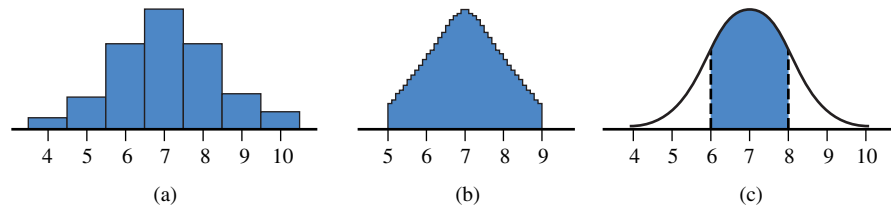


Figure 7.5 Probability distribution for birth weight: (a) weight measured to the nearest pound; (b) weight measured to the nearest tenth of a pound; (c) limiting curve as measurement accuracy increases; shaded area = $P(6 \leq \text{weight} \leq 8)$.

x falls in an interval such as $6 \leq x \leq 8$ is the area under the curve and above that interval.

DEFINITION

A **probability distribution for a continuous random variable x** is specified by a mathematical function denoted by $f(x)$ and called the **density function**. The graph of a density function is a smooth curve (the **density curve**). The following requirements must be met:

1. $f(x) \geq 0$ (so that the curve cannot dip below the horizontal axis).
2. The total area under the density curve is equal to 1.

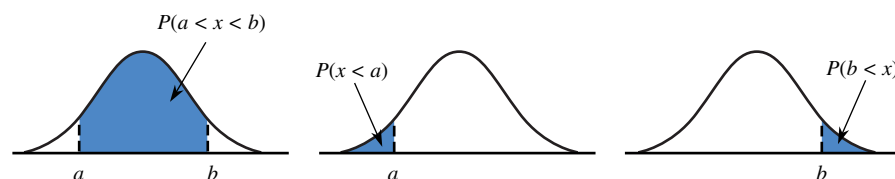
The probability that x falls in any particular interval is the area under the density curve and above the interval.

Many probability calculations for continuous random variables involve the following three events:

1. $a < x < b$, the event that the random variable x assumes a value between two given numbers, a and b
2. $x < a$, the event that the random variable x assumes a value less than a given number a
3. $b < x$, the event that the random variable x assumes a value greater than a given number b (this can also be written as $x > b$)

Figure 7.6 illustrates how the probabilities of these events are identified with areas under a density curve.

Figure 7.6 Probabilities as areas under a probability density curve.



Example 7.7 Application Processing Times

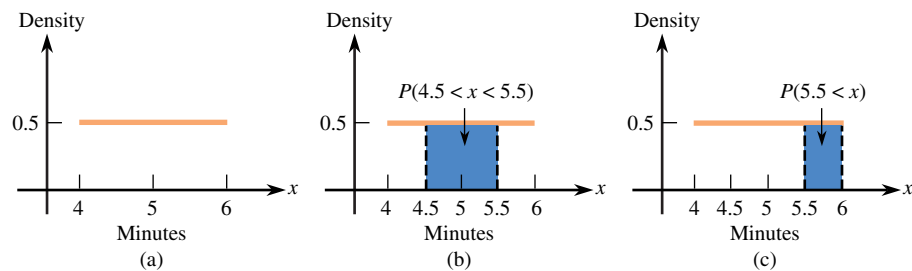
Define a continuous random variable x by $x =$ amount of time (in minutes) taken by a clerk to process a certain type of application form. Suppose that x has a probability distribution with density function

$$f(x) = \begin{cases} .5 & 4 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

The graph of $f(x)$, the density curve, is shown in Figure 7.7(a). It is especially easy to use this density curve to calculate probabilities, because it just requires finding the area of rectangles using the formula

$$\text{area} = (\text{base})(\text{height})$$

Figure 7.7 The uniform distribution for Example 7.7.



The curve has positive height, 0.5, only between $x = 4$ and $x = 6$. The total area under the curve is just the area of the rectangle with base extending from 4 to 6 and with height 0.5. This gives

$$\text{area} = (6 - 4)(0.5) = 1$$

as required.

When the density is constant over an interval (resulting in a horizontal density curve), the probability distribution is called a *uniform distribution*.

As illustrated in Figure 7.7(b), the probability that x is between 4.5 and 5.5 is

$$\begin{aligned} P(4.5 < x < 5.5) &= \text{area of shaded rectangle} \\ &= (\text{base width})(\text{height}) \\ &= (5.5 - 4.5)(.5) \\ &= .5 \end{aligned}$$

Similarly (see Figure 7.7(c)), because in this context $x > 5.5$ is equivalent to $5.5 \leq x \leq 6$, we have

$$P(5.5 < x) = (6 - 5.5)(.5) = .25$$

According to this model, in the long run, 25% of all forms that are processed will have processing times that exceed 5.5 min. ■

The probability that a *discrete* random variable x lies in the interval between two limits a and b depends on whether either limit is included in the interval. Suppose, for example, that x is the number of major defects on a new automobile. Then

$$P(3 \leq x \leq 7) = p(3) + p(4) + p(5) + p(6) + p(7)$$

whereas

$$P(3 < x < 7) = p(4) + p(5) + p(6)$$

However, if x is a *continuous* random variable, such as task completion time, then

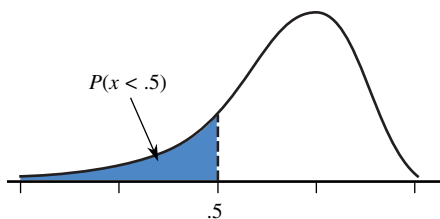
$$P(3 \leq x \leq 7) = P(3 < x < 7)$$

because the area under a density curve and above a single value such as 3 or 7 is 0. Geometrically, we can think of finding the area above a single point as finding the area of a rectangle with width = 0. The area above an interval of values therefore does not depend on whether either endpoint is included.

For any two numbers a and b with $a < b$,

$$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$

when x is a continuous random variable.



Probabilities for continuous random variables are often calculated using cumulative areas. A cumulative area is all of the area under the density curve to the left of a particular value. Figure 7.8 illustrates the cumulative area to the left of .5, which is $P(x < .5)$. The probability that x is in any particular interval, $P(a < x < b)$, is the difference between two cumulative areas.

Figure 7.8 A cumulative area under a density curve.

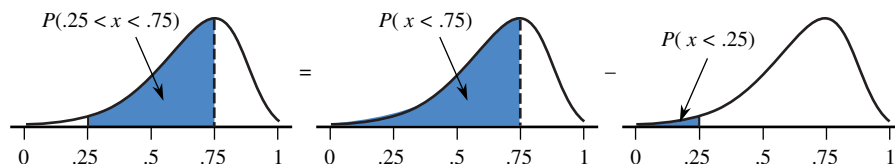
The probability that a continuous random variable x lies between a lower limit a and an upper limit b is

$$\begin{aligned} P(a < x < b) &= (\text{cumulative area to the left of } b) - (\text{cumulative area to the left of } a) \\ &= P(x < b) - P(x < a) \end{aligned}$$

The foregoing property is illustrated in Figure 7.9 for the case of $a = .25$ and $b = .75$. We will use this result extensively in Section 7.6 when we calculate probabilities using the normal distribution.

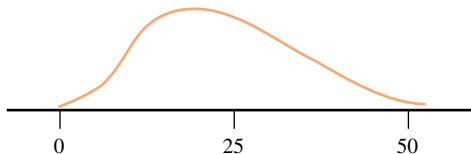
For some continuous distributions, cumulative areas can be calculated using methods from the branch of mathematics called integral calculus. However, because we are not assuming knowledge of calculus, we will rely on tables that have been constructed for the commonly encountered continuous probability distributions.

Figure 7.9 Calculation of $P(a < x < b)$ using cumulative areas.



Exercises 7.20–7.26

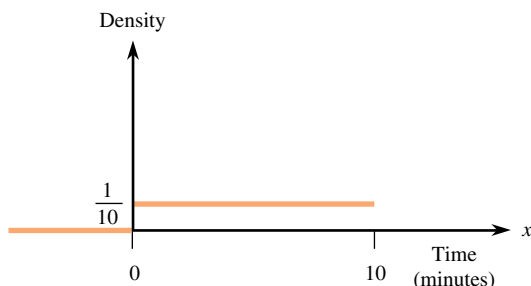
7.20 Let x denote the lifetime (in thousands of hours) of a certain type of fan used in diesel engines. The density curve of x is as pictured:



Shade the area under the curve corresponding to each of the following probabilities (draw a new curve for each part):

- $P(10 < x < 25)$
- $P(10 \leq x \leq 25)$
- $P(x < 30)$
- The probability that the lifetime is at least 25,000 hr
- The probability that the lifetime exceeds 25,000 hr

7.21 A particular professor never dismisses class early. Let x denote the amount of time past the hour (minutes) that elapses before the professor dismisses class. Suppose that x has a uniform distribution on the interval from 0 to 10 min. The density curve is shown in the following figure:



- What is the probability that at most 5 min elapse before dismissal?
- What is the probability that between 3 and 5 min elapse before dismissal?

7.22 Refer to the probability distribution given in Exercise 7.21. Put the following probabilities in order,

from smallest to largest:

$$P(2 < x < 3), P(2 \leq x \leq 3), P(x < 2), P(x > 7).$$

Explain your reasoning.

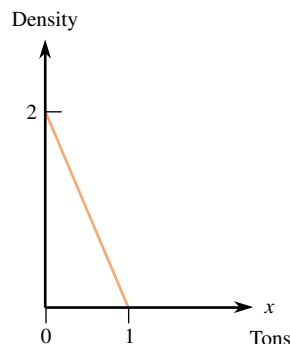
7.23 The article “Modeling Sediment and Water Column Interactions for Hydrophobic Pollutants” (*Water Research* [1984]: 1169–1174) suggests the uniform distribution on the interval from 7.5 to 20 as a model for x = depth (in centimeters) of the bioturbation layer in sediment for a certain region.

- Draw the density curve for x .
- What is the height of the density curve?
- What is the probability that x is at most 12?
- What is the probability that x is between 10 and 15? Between 12 and 17? Why are these two probabilities equal?

7.24 Let x denote the amount of gravel sold (in tons) during a randomly selected week at a particular sales facility. Suppose that the density curve has height $f(x)$ above the value x , where

$$f(x) = \begin{cases} 2(1 - x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The density curve (the graph of $f(x)$) is shown in the following figure:

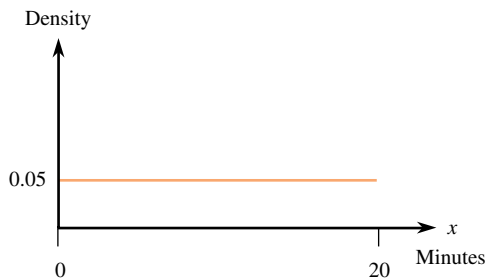


Use the fact that the area of a triangle = $\frac{1}{2}$ (base)(height) to calculate each of the following probabilities:

- $P\left(x < \frac{1}{2}\right)$

- b. $P\left(x \leq \frac{1}{2}\right)$
 c. $P\left(x < \frac{1}{4}\right)$
 d. $P\left(\frac{1}{4} < x < \frac{1}{2}\right)$ (Hint: Use the results of Parts (a)–(c).)
 e. The probability that gravel sold exceeds $\frac{1}{2}$ ton
 f. The probability that gravel sold is at least $\frac{1}{4}$ ton

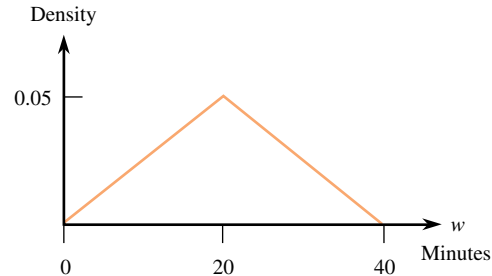
7.25 Let x be the amount of time (in minutes) that a particular San Francisco commuter must wait for a BART train. Suppose that the density curve is as pictured (a uniform distribution):



- a. What is the probability that x is less than 10 min? more than 15 min?

- b. What is the probability that x is between 7 and 12 min?
 c. Find the value c for which $P(x < c) = .9$.

7.26 Referring to Exercise 7.25, let x and y be waiting times on two independently selected days. Define a new random variable w by $w = x + y$, the sum of the two waiting times. The set of possible values for w is the interval from 0 to 40 (because both x and y can range from 0 to 20). It can be shown that the density curve of w is as pictured (this curve is called a triangular distribution, for obvious reasons!):



- a. Verify that the total area under the density curve is equal to 1. (Hint: The area of a triangle is $= \frac{1}{2}(\text{base})(\text{height})$.)
 b. What is the probability that w is less than 20? less than 10? greater than 30?
 c. What is the probability that w is between 10 and 30? (Hint: It might be easier first to find the probability that w is not between 10 and 30.)

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.4 Mean and Standard Deviation of a Random Variable

We study a random variable x , such as the number of insurance claims made by a homeowner (a discrete variable) or the birth weight of a baby (a continuous variable), to learn something about how its values are distributed along the measurement scale. The sample mean \bar{x} and sample standard deviation s summarize center and spread for the values in a sample. Similarly, the mean value and standard deviation of a random variable describe where the variable's probability distribution is centered and the extent to which it spreads out about the center.

The **mean value of a random variable x** , denoted by μ_x describes where the probability distribution of x is centered.

The **standard deviation of a random variable x** , denoted by σ_x describes variability in the probability distribution. When σ_x is small, observed values of x will tend to be close to the mean value (little variability). When the value of σ_x is large, there will be more variability in observed x values.

Figure 7.10(a) shows two discrete probability distributions with the same standard deviation (spread) but different means (center). One distribution has a mean of $\mu_x = 6$ and the other has $\mu_x = 10$. Which is which? Figure 7.10(b) shows two continuous probability distributions that have the same mean but different standard deviations. Which distribution—(i) or (ii)—has the larger standard deviation? Finally, Figure 7.10(c) shows three continuous distributions with different means and standard deviations. Which of the three distributions has the largest mean? Which has a mean of about 5? Which distribution has the smallest standard deviation? (The correct answers to our questions are the following: Figure 7.10(a)(ii) has a mean of 6, and Figure 7.10(a)(i) has a mean of 10; Figure 7.10(b)(ii) has the larger standard deviation; Figure 7.10(c)(iii) has

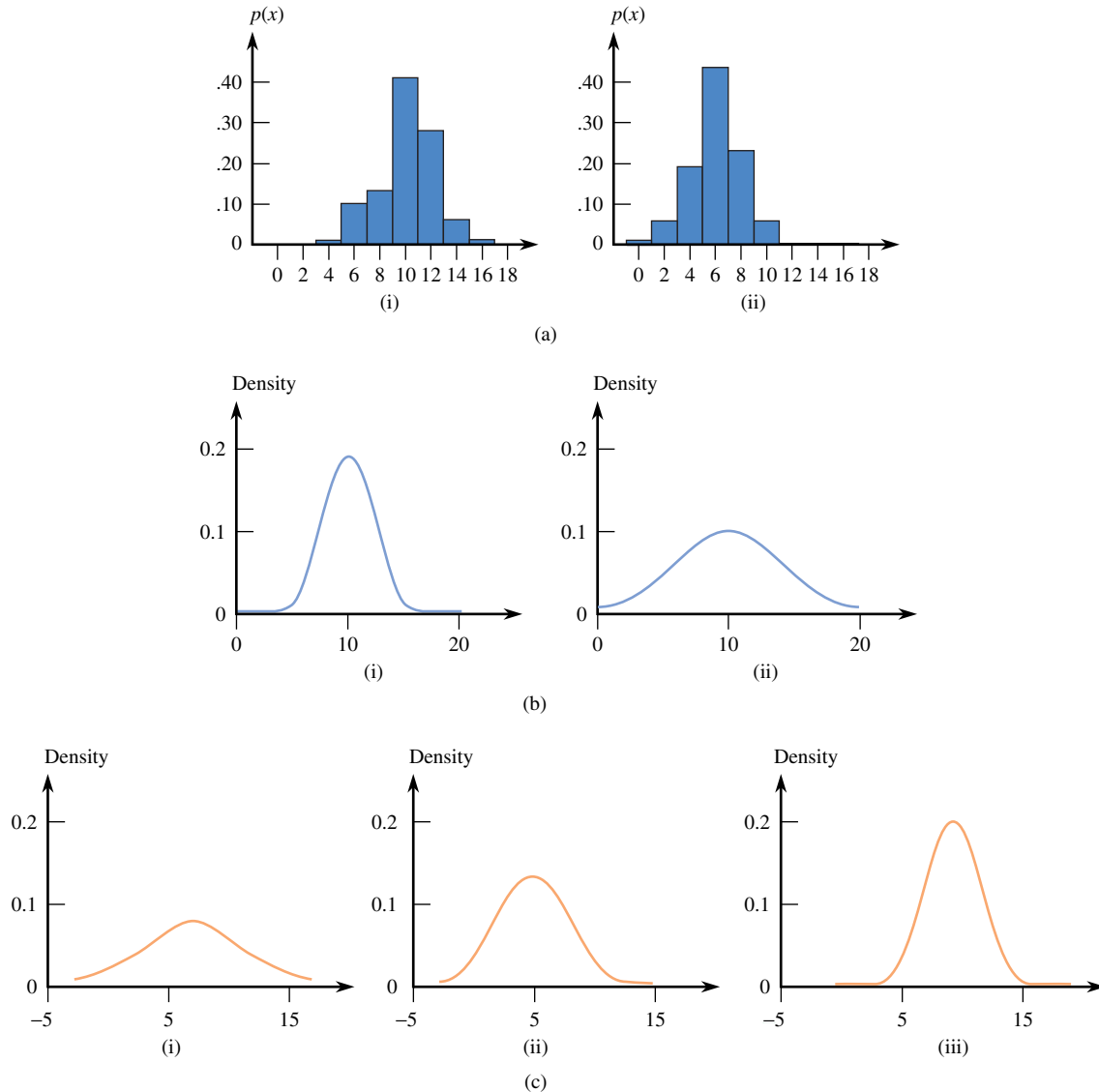


Figure 7.10 Some probability distributions: (a) different values of μ_x with the same value of σ_x ; (b) different values of σ_x with the same value of μ_x ; (c) different values of μ_x and σ_x .

the largest mean, Figure 7.10(c)(ii) has a mean of about 5, and Figure 7.10(c)(iii) has the smallest standard deviation.)

It is customary to use the terms *mean of the random variable x* and *mean of the probability distribution of x* interchangeably. Similarly, the standard deviation of the random variable x and the standard deviation of the probability distribution of x refer to the same thing. Although the mean and standard deviation are computed differently for discrete and continuous random variables, the interpretation is the same in both cases.

■ Mean Value of a Discrete Random Variable

Consider an experiment consisting of the random selection of an automobile licensed in a particular state. Let the discrete random variable x be the number of low-beam headlights on the selected car that need adjustment. Possible x values are 0, 1, and 2, and the probability distribution of x might be as follows:

| | | | |
|-------------|----|----|----|
| x value | 0 | 1 | 2 |
| Probability | .5 | .3 | .2 |

The corresponding probability histogram appears in Figure 7.11.

In a sample of 100 cars, the sample relative frequencies might differ somewhat from the given probabilities (which are the limiting relative frequencies). We might see:

| | | | |
|-----------|----|----|----|
| x value | 0 | 1 | 2 |
| Frequency | 46 | 33 | 21 |

The sample average value of x for these 100 observations is then the sum of 46 zeros, 33 ones, and 21 twos, all divided by 100:

$$\begin{aligned}\bar{x} &= \frac{(46)(0) + (33)(1) + (21)(2)}{100} \\ &= \left(\frac{46}{100}\right)(0) + \left(\frac{33}{100}\right)(1) + \left(\frac{21}{100}\right)(2) \\ &= (\text{rel. freq. of } 0)(0) + (\text{rel. freq. of } 1)(1) + (\text{rel. freq. of } 2)(2) \\ &= .75\end{aligned}$$

As the sample size increases, each relative frequency approaches the corresponding probability. In a very long sequence of experiments, the value of \bar{x} approaches

$$\begin{aligned}(\text{probability that } x = 0)(0) &+ (\text{probability that } x = 1)(1) + (\text{probability that } x = 2)(2) \\ &= (.5)(0) + (.3)(1) + (.2)(2) \\ &= .70 \\ &= \text{mean value of } x\end{aligned}$$

Notice that the expression for \bar{x} is a weighted average of possible x values; the weight of each value is the observed relative frequency. Similarly, the mean value of the random variable x is a weighted average, but now the weights are the probabilities from the probability distribution, as given in the definition in the following box.

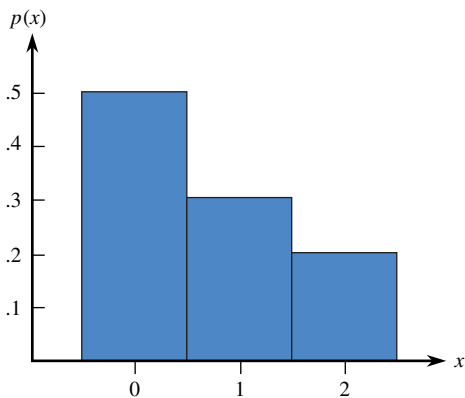


Figure 7.11 Probability histogram for the distribution of the number of headlights needing adjustments.

DEFINITION

The **mean value of a discrete random variable** x , denoted by μ_x , is computed by first multiplying each possible x value by the probability of observing that value and then adding the resulting quantities. Symbolically,

$$\mu_x = \sum_{\text{all possible } x \text{ values}} x \cdot p(x)$$

The term **expected value** is sometimes used in place of mean value, and $E(x)$ is alternative notation for μ_x .

Example 7.8 Exam Attempts

Individuals applying for a certain license are allowed up to four attempts to pass the licensing exam. Let x denote the number of attempts made by a randomly selected applicant. The probability distribution of x is as follows:

| | | | | |
|--------|-----|-----|-----|-----|
| x | 1 | 2 | 3 | 4 |
| $p(x)$ | .10 | .20 | .30 | .40 |

Then x has mean value

$$\begin{aligned} \mu_x &= \sum_{x=1,2,3,4} x \cdot p(x) \\ &= (1)p(1) + (2)p(2) + (3)p(3) + (4)p(4) \\ &= (1)(.10) + (2)(.20) + (3)(.30) + (4)(.40) \\ &= .10 + .40 + .90 + 1.60 \\ &= 3.00 \end{aligned}$$

It is no accident that the symbol μ_x for the mean value is the same symbol used previously for a population mean. When the probability distribution describes how x values are distributed among the members of a population (and therefore the probabilities are population relative frequencies), the mean value of x is exactly the average value of x in the population.

Example 7.9 Apgar Scores

At 1 min after birth and again at 5 min, each newborn child is given a numerical rating called an Apgar score. Possible values of this score are 0, 1, 2, . . . , 9, 10. A child's score is determined by five factors: muscle tone, skin color, respiratory effort, strength of heartbeat, and reflex, with a high score indicating a healthy infant. Let the random variable x denote the Apgar score (at 1 min) of a randomly selected newborn infant at a particular hospital, and suppose that x has the following probability distribution:

| | | | | | | | | | | | |
|--------|------|------|------|------|-----|-----|-----|-----|-----|-----|-----|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $p(x)$ | .002 | .001 | .002 | .005 | .02 | .04 | .17 | .38 | .25 | .12 | .01 |

The mean value of x is

$$\begin{aligned} \mu_x &= (0)p(0) + (1)p(1) + \cdots + (9)p(9) + (10)p(10) \\ &= (0)(.002) + (1)(.001) + \cdots + (9)(.12) + (10)(.01) \\ &= 7.16 \end{aligned}$$

The average Apgar score for a *sample* of newborn children born at this hospital may be $\bar{x} = 7.05$, $\bar{x} = 8.30$, or any one of a number of other possible values between 0 and 10. However, as child after child is born and rated, the average score will approach the value 7.16. This value can be interpreted as the mean Apgar score for the population of all babies born at this hospital.



■ **Standard Deviation of a Discrete Random Variable**

The mean value μ_x provides only a partial summary of a probability distribution. Two different distributions can have the same value of μ_x , yet a long sequence of sample values from one distribution might exhibit considerably more variability than a long sequence of values from the other distribution.

Example 7.10 Defective Components

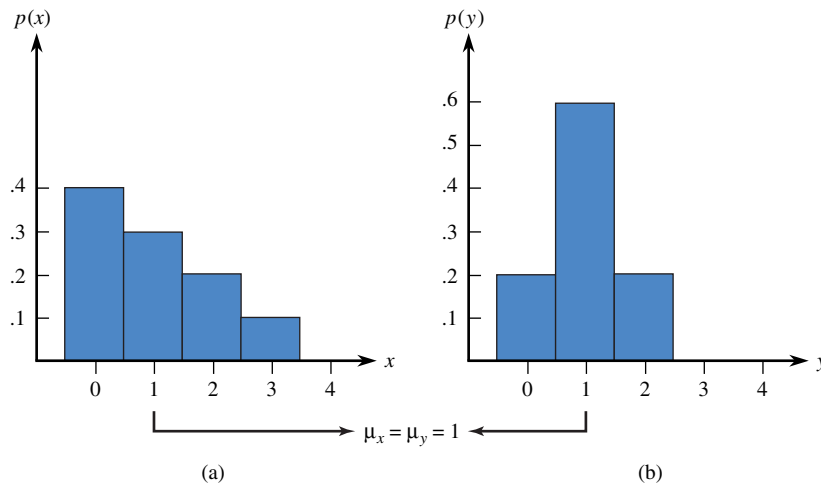
A television manufacturer receives certain components in lots of four from two different suppliers. Let x and y denote the number of defective components in randomly selected lots from the first and second suppliers, respectively. The probability distributions for x and y are as follows:

| | | | | | | | | | | | |
|--------|----|----|----|----|---|--------|----|----|----|---|---|
| x | 0 | 1 | 2 | 3 | 4 | y | 0 | 1 | 2 | 3 | 4 |
| $p(x)$ | .4 | .3 | .2 | .1 | 0 | $p(y)$ | .2 | .6 | .2 | 0 | 0 |

Probability histograms for x and y are given in Figure 7.12.

It is easy to verify that the mean values of both x and y are 1, so for either supplier the long-run average number of defective components per lot is 1. However, the two probability histograms show that the probability distribution for the second supplier is concentrated closer to the mean value than is the first supplier's distribution.

Figure 7.12 Probability distribution for the number of defective components in Example 7.10: (a) Supplier 1; (b) Supplier 2.



The greater spread of the first distribution implies that there will be more variability in a long sequence of observed x values than in an observed sequence of y values. For example, the y sequence will contain no 3's, whereas in the long run, 10% of the observed x values will be 3.

As with s^2 and s , the variance and standard deviation of x involve squared deviations from the mean. A value far from the mean results in a large squared deviation. However, such a value contributes substantially to variability in x only if the probability associated with that value is not too small. For example, if $\mu_x = 1$ and $x = 25$ is a possible value, then the squared deviation is $(25 - 1)^2 = 576$. If, however, $P(x = 25) = .000001$, the value 25 will hardly ever be observed, so it won't contribute much to variability in a long sequence of observations. This is why each squared deviation is multiplied by the probability associated with the value to obtain a measure of variability.

DEFINITION

The **variance of a discrete random variable** x , denoted by σ_x^2 , is computed by first subtracting the mean from each possible x value to obtain the deviations, then squaring each deviation and multiplying the result by the probability of the corresponding x value, and finally adding these quantities. Symbolically,

$$\sigma_x^2 = \sum_{\text{all possible } x \text{ values}} (x - \mu)^2 p(x)$$

The **standard deviation of x** , denoted by σ_x , is the square root of the variance.

When the probability distribution describes how x values are distributed among members of a population (so that the probabilities are population relative frequencies) σ_x^2 and σ_x are the population variance and standard deviation (of x), respectively.

Example 7.11 Defective Components Revised

For x = number of defective components in a lot from the first supplier in Example 7.10,

$$\begin{aligned}\sigma_x^2 &= (0 - 1)^2 p(0) + (1 - 1)^2 p(1) + (2 - 1)^2 p(2) + (3 - 1)^2 p(3) \\ &= (1)(.4) + (0)(.3) + (1)(.2) + (4)(.1) \\ &= 1.0\end{aligned}$$

Therefore $\sigma_x = 1.0$. For y = the number of defectives in a lot from the second supplier,

$$\sigma_y^2 = (0 - 1)^2 (.2) + (1 - 1)^2 (.6) + (2 - 1)^2 (.2) = .4$$

Then $\sigma_y = \sqrt{.4} = .632$. The fact that $\sigma_x > \sigma_y$ confirms the impression conveyed by Figure 7.12 concerning the variability of x and y .

Example 7.12 More on Apgar scores

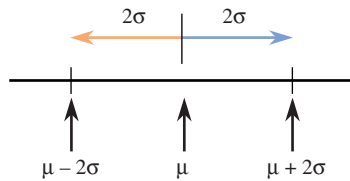


Figure 7.13 Values within 2 standard deviations of the mean.

Reconsider the distribution of Apgar scores for children born at a certain hospital, introduced in Example 7.9. What is the probability that a randomly selected child’s score will be within 2 standard deviations of the mean score? As Figure 7.13 shows, values of x within 2 standard deviations of the mean are those for which $\mu - 2\sigma < x < \mu + 2\sigma$.

From Example 7.9 we already have $\mu_x = 7.16$. The variance is

$$\begin{aligned} \sigma^2 &= \sum (x - \mu)^2 p(x) = \sum (x - 7.16)^2 p(x) \\ &= (0 - 7.16)^2(.002) + (1 - 7.16)^2(.001) + \cdots + (10 - 7.16)^2(.01) \\ &= 1.5684 \end{aligned}$$

and the standard deviation is

$$\sigma = \sqrt{1.5684} = 1.25$$

This gives (using the probabilities given in Example 7.9)

$$\begin{aligned} P(\mu - 2\sigma < x < \mu + 2\sigma) &= P(7.16 - 2.50 < x < 7.16 + 2.50) \\ &= P(4.66 < x < 9.66) \\ &= p(5) + \cdots + p(9) \\ &= .96 \end{aligned}$$



■ Mean and Standard Deviation When x Is Continuous

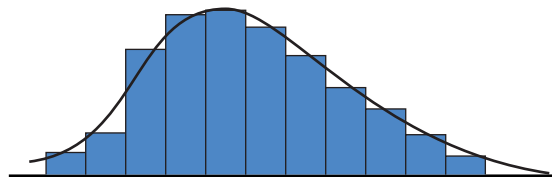


Figure 7.14 Approximating a density curve by a probability histogram.

Figure 7.14 illustrates how the density curve for a continuous random variable can be approximated by a probability histogram of a discrete random variable. Computing the mean value and the standard deviation using this discrete distribution gives approximate values of μ_x and σ_x for the continuous random variable x . If an even more accurate approximating probability histogram is used (narrower rectangles), better approximations of μ_x and σ_x result.

In practice, such an approximation method is often unnecessary. Instead, μ_x and σ_x can be defined and computed using methods from calculus. The details need not concern us; what is important is that μ_x and σ_x play exactly the same role here as they did in the discrete case. The mean value μ_x locates the center of the continuous distribution and gives the approximate long-run average of many observed x values. The standard deviation σ_x measures the extent that the continuous distribution (density curve) spreads out about μ_x and gives information about the amount of variability that can be expected in a long sequence of observed x values.

Example 7.13 A “Concrete” Example

A company receives concrete of a certain type from two different suppliers. Define random variables x and y as follows:

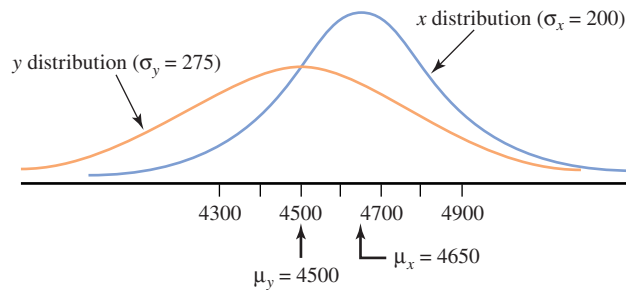
- x = compressive strength of a randomly selected batch from Supplier 1
- y = compressive strength of a randomly selected batch from Supplier 2

Suppose that

$$\begin{array}{ll} \mu_x = 4650 \text{ lb/in.}^2 & \sigma_x = 200 \text{ lb/in.}^2 \\ \mu_y = 4500 \text{ lb/in.}^2 & \sigma_y = 275 \text{ lb/in.}^2 \end{array}$$

The long-run average strength per batch for many, many batches from Supplier 1 will be roughly 4650 lb/in.². This is 150 lb/in.² greater than the long-run average for batches from Supplier 2. In addition, a long sequence of batches from Supplier 1 will exhibit substantially less variability in compressive strength values than will a similar sequence from Supplier 2. The first supplier is preferred to the second both in terms of average value and variability. Figure 7.15 displays density curves that are consistent with this information.

Figure 7.15 Density curves for Example 7.13.



■ Mean and Variance of Linear Functions and Linear Combinations

We have seen how the mean and standard deviation of one or more random variables provide useful information about the variables' long-run behavior, but we might also be interested in the behavior of some function of these variables.

For example, consider the experiment in which a customer of a propane gas company is randomly selected. Suppose that the mean and standard deviation of the random variable

x = number of gallons required to fill a customer's propane tank

are known to be 318 gal and 42 gal, respectively. The company is considering two different pricing models:

Model 1: \$3 per gal

Model 2: service charge of \$50 + \$2.80 per gal

The company is interested in the variable

y = amount billed

For each of the two models, y can be expressed as a function of the random variable x :

Model 1: $y_{\text{model 1}} = 3x$

Model 2: $y_{\text{model 2}} = 50 + 2.8x$

Both of these equations are examples of a linear function of x . The mean and standard deviation of a linear function of x can be computed from the mean and standard deviation of x , as described in the following box.

The Mean, Variance, and Standard Deviation of a Linear Function

If x is a random variable with mean μ_x and variance σ_x and a and b are numerical constants, the random variable y defined by

$$y = a + bx$$

is called a **linear function of the random variable x** .

The mean of $y = a + bx$ is

$$\mu_y = \mu_{a+bx} = a + b\mu_x$$

The variance of y is

$$\sigma_y^2 = \sigma_{a+bx}^2 = b^2\sigma_x^2$$

from which it follows that the standard deviation of y is

$$\sigma_y = \sigma_{a+bx} = |b|\sigma_x$$

We can use the results in the preceding box to compute the mean and standard deviation of the billing amount variable for the propane gas example, as follows:

For Model 1:

$$\begin{aligned}\mu_{model\ 1} &= \mu_{3x} = 3\mu_x = 3(318) = 954 \\ \sigma_{model\ 1}^2 &= \sigma_{3x}^2 = 3^2\sigma_x^2 = 9(42)^2 = 15,876 \\ \sigma_{model\ 1} &= \sqrt{15,876} = 126 = 3(42)\end{aligned}$$

For Model 2:

$$\begin{aligned}\mu_{model\ 2} &= \mu_{50+2.8x} = 50 + 2.8\mu_x = 50 + 2.8(318) = 940.40 \\ \sigma_{model\ 2}^2 &= \sigma_{50+2.8x}^2 = 2.8^2\sigma_x^2 = (2.8)^2(42)^2 = 13,829.76 \\ \sigma_{model\ 2} &= \sqrt{13,829.76} = 117.60 = 2.8(42)\end{aligned}$$

The mean billing amount for Model 1 is a bit higher than for Model 2, as is the variability in billing amounts. Model 2 results in slightly more consistency from bill to bill in the amount charged.

Now let's consider a different type of problem. Suppose that you have three tasks that you plan to complete on the way home from school: stop at the public library to return an overdue book for which you must pay a fine, deposit your most recent paycheck at the bank, and stop by the office supply store to purchase paper for your computer printer. Define the following variables:

$$\begin{aligned}x_1 &= \text{time required to return book and pay fine} \\ x_2 &= \text{time required to deposit paycheck} \\ x_3 &= \text{time required to buy printer paper}\end{aligned}$$

We can then define a new variable, y , to represent the total amount of time to complete these tasks:

$$y = x_1 + x_2 + x_3$$

Defined in this way, y is an example of a linear combination of random variables.

If x_1, x_2, \dots, x_n are random variables and a_1, a_2, \dots, a_n are numerical constants, the random variable y defined as

$$y = a_2x_1 + a_1x_2 + \cdots + a_nx_n$$

is a **linear combination of the x_i 's**.

For example, $y = 10x_1 - 5x_2 + 8x_3$ is a linear combination of x_1, x_2 and x_3 with $a_1 = 10, a_2 = -5$ and $a_3 = 8$. It is easy to compute the mean of a linear combination of x_i if the individual means $\mu_1, \mu_2, \dots, \mu_n$ are known. The variance and standard deviation of a linear combination of the x_i are also easily computed *if the x_i are independent*. Two random variables x_i and x_j are independent if any event defined solely by x_i is independent of any event defined solely by x_j . When the x_i are not independent, computation of the variance and standard deviation of a linear combination of the x_i is more complicated; this case is not considered here.

Mean, Variance, and Standard Deviation for Linear Combinations

If x_1, x_2, \dots, x_n are random variables with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, and

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

then

1. $\mu_y = \mu_{a_1x_1 + a_2x_2 + \cdots + a_nx_n} = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n$

This result is true regardless of whether the x_i 's are independent.

2. When x_1, x_2, \dots, x_n are independent random variables,

$$\begin{aligned}\sigma_y^2 &= \sigma_{a_1x_1 + a_2x_2 + \cdots + a_nx_n}^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \\ \sigma_y &= \sigma_{a_1x_1 + a_2x_2 + \cdots + a_nx_n} = \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2}\end{aligned}$$

This result is true only when the x_i 's are independent.

Examples 7.14–7.16 illustrate the use of these rules.

Example 7.14 Freeway Traffic

Three different roads feed into a particular freeway entrance. Suppose that during a fixed time period, the number of cars coming from each road onto the freeway is a random variable with mean values as follows:

| | | | |
|------|-----|------|-----|
| Road | 1 | 2 | 3 |
| Mean | 800 | 1000 | 600 |

With x_i representing the number of cars entering from road i , we can define $y = x_1 + x_2 + x_3$, the total number of cars entering the freeway. The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+x_2+x_3} \\ &= \mu_{x_1} + \mu_{x_2} + \mu_{x_3} \\ &= 800 + 1000 + 600 \\ &= 2400\end{aligned}$$



Example 7.15 Combining Exam Subscores

A nationwide standardized exam consists of a multiple-choice section and a free response section. For each section, the mean and standard deviation are reported to be

| | Mean | Standard Deviation |
|-----------------|------|--------------------|
| Multiple Choice | 38 | 6 |
| Free Response | 30 | 7 |

Let's define x_1 and x_2 as the multiple-choice score and the free-response score, respectively, of a student selected at random from those taking this exam. We are also interested in the variable $y =$ total score. Suppose that the total score is computed as $y = x_1 + 2x_2$. What are the mean and standard deviation of y ?

Because $y = x_1 + 2x_2$ is a linear combination of x_1 and x_2 , the mean of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+2x_2} \\ &= \mu_{x_1} + 2\mu_{x_2} \\ &= 38 + 2(30) \\ &= 98\end{aligned}$$

What about the variance and standard deviation of y ? To use Rule 2 in the preceding box, x_1 and x_2 must be independent. It is unlikely that the value of x_1 (a student's multiple-choice score) would be unrelated to the value of x_2 (the same student's free-response score), because it seems probable that students who score well on one section of the exam will also tend to score well on the other section. Therefore, it would not be appropriate to calculate the variance and standard deviation from the given information.



Example 7.16 Luggage Weights

A commuter airline flies small planes between San Luis Obispo and San Francisco. For small planes, the baggage weight is a concern, especially on foggy mornings, because the weight of the plane has an effect on how quickly the plane can ascend. Suppose that it is known that the variable $x =$ weight of baggage checked by a

randomly selected passenger has a mean and standard deviation of 42 and 16, respectively. Consider a flight on which 10 passengers, all traveling alone, are flying. If we use x_i to denote the baggage weight for passenger i (for i ranging from 1 to 10), the total weight of checked baggage, y , is then

$$y = x_1 + x_2 + \cdots + x_{10}$$

Note that y is a linear combination of the x_i . The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1} + \mu_{x_2} + \cdots + \mu_{x_{10}} \\ &= 42 + 42 + \cdots + 42 \\ &= 420\end{aligned}$$

Since the ten passengers are all traveling alone, it is reasonable to think that the ten baggage weights are unrelated and that the x_i are independent. (This would not be a reasonable assumption if the 10 passengers were not traveling alone.) Then the variance of y is

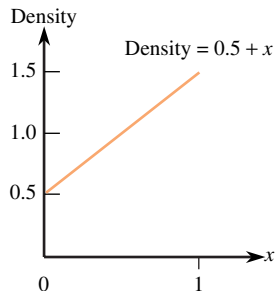
$$\begin{aligned}\sigma_y^2 &= \sigma_{x_1}^2 + \sigma_{x_2}^2 + \cdots + \sigma_{x_{10}}^2 \\ &= 16^2 + 16^2 + \cdots + 16^2 \\ &= 2560\end{aligned}$$

and the standard deviation of y is

$$\sigma_y = \sqrt{2560} = 50.596$$

Exercises 7.27–7.44

7.27 An express mail service charges a special rate for any package that weighs less than 1 lb. Let x denote the weight of a randomly selected parcel that qualifies for this special rate. The probability distribution of x is specified by the following density curve:



Use the fact that area of a trapezoid = (base)(average of two side lengths) to answer each of the following questions:

a. What is the probability that a randomly selected package of this type weighs at most 0.5 lb? between 0.25 and 0.5 lb? at least 0.75 lb?

b. It can be shown that $\mu_x = \frac{7}{12}$ and $\sigma_x^2 = \frac{11}{144}$. What is the probability that the value of x is more than 1 standard deviation from the mean value?

7.28 The probability distribution of x , the number of defective tires on a randomly selected automobile checked at a certain inspection station, is given in the following table:

| | | | | | |
|--------|-----|-----|-----|-----|-----|
| x | 0 | 1 | 2 | 3 | 4 |
| $p(x)$ | .54 | .16 | .06 | .04 | .20 |

- Calculate the mean value of x .
- What is the probability that x exceeds its mean value?

7.29 Exercise 7.9 introduced the following probability distribution for y = the number of broken eggs in a carton:

| | | | | | |
|--------|-----|-----|-----|-----|-----|
| y | 0 | 1 | 2 | 3 | 4 |
| $p(y)$ | .65 | .20 | .10 | .04 | .01 |

- Calculate and interpret μ_y .
- In the long run, for what percentage of cartons is the number of broken eggs less than μ_y ? Does this surprise you?

c. Why doesn't $\mu_y = (0 + 1 + 2 + 3 + 4)/5 = 2.0$. Explain.

7.30 Referring to Exercise 7.29, use the result of Part (a) along with the fact that a carton contains 12 eggs to determine the mean value of $z =$ the number of unbroken eggs. (Hint: z can be written as a linear function of x .)

7.31 The mean value of x , the number of defective tires, whose distribution appears in Exercise 7.28, is $\mu_x = 1.2$. Calculate σ_x^2 and σ_x .

7.32 Exercise 7.8 gave the following probability distribution for $x =$ the number of courses for which a randomly selected student at a certain university is registered:

| | | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|-----|
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $p(x)$ | .02 | .03 | .09 | .25 | .40 | .16 | .05 |

It can be easily verified that $\mu = 4.66$ and $\sigma = 1.20$.

- Because $\mu - \sigma = 3.46$, the x values 1, 2, and 3 are more than 1 standard deviation below the mean. What is the probability that x is more than 1 standard deviation below its mean?
- What x values are more than 2 standard deviations away from the mean value (i.e., either less than $\mu - 2\sigma$ or greater than $\mu + 2\sigma$)? What is the probability that x is more than 2 standard deviations away from its mean value?

7.33 Suppose that for a given computer salesperson, the probability distribution of $x =$ the number of systems sold in one month is given by the following table:

| | | | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $p(x)$ | .05 | .10 | .12 | .30 | .30 | .11 | .01 | .01 |

- Find the mean value of x (the mean number of systems sold).
- Find the variance and standard deviation of x . How would you interpret these values?
- What is the probability that the number of systems sold is within 1 standard deviation of its mean value?
- What is the probability that the number of systems sold is more than 2 standard deviations from the mean?

7.34 A local television station sells 15-sec, 30-sec, and 60-sec advertising spots. Let x denote the length of a randomly selected commercial appearing on this station, and suppose that the probability distribution of x is given by the following table:

| | | | |
|--------|----|----|----|
| x | 15 | 30 | 60 |
| $p(x)$ | .1 | .3 | .6 |

- Find the average length for commercials appearing on this station.
- If a 15-sec spot sells for \$500, a 30-sec spot for \$800, and a 60-sec spot for \$1000, find the average amount paid for commercials appearing on this station. (Hint: Consider a new variable, $y =$ cost, and then find the probability distribution and mean value of y .)

7.35 An author has written a book and submitted it to a publisher. The publisher offers to print the book and gives the author the choice between a flat payment of \$10,000 and a royalty plan. Under the royalty plan the author would receive \$1 for each copy of the book sold. The author thinks that the following table gives the probability distribution of the variable $x =$ the number of books that will be sold:

| | | | | |
|--------|------|------|--------|--------|
| x | 1000 | 5000 | 10,000 | 20,000 |
| $p(x)$ | .05 | .30 | .40 | .25 |

Which payment plan should the author choose? Why?

7.36 A grocery store has an express line for customers purchasing at most five items. Let x be the number of items purchased by a randomly selected customer using this line. Give examples of two different assignments of probabilities such that the resulting distributions have the same mean but quite different standard deviations.

7.37 ▼ A gas station sells gasoline at the following prices (in cents per gallon, depending on the type of gas and service): 315.9, 318.9, 329.9, 339.9, 344.9, and 359.7. Let y denote the price per gallon paid by a randomly selected customer.

- Is y a discrete random variable? Explain.
- Suppose that the probability distribution of y is as follows:

| | | | | | | |
|--------|-------|-------|-------|-------|-------|-------|
| y | 315.9 | 318.9 | 329.9 | 339.9 | 344.9 | 359.7 |
| $p(y)$ | .36 | .24 | .10 | .16 | .08 | .06 |

What is the probability that a randomly selected customer has paid more than \$3.20 per gallon? Less than \$3.40 per gallon?

- Refer to Part (b), and calculate the mean value and standard deviation of y . Interpret these values.

7.38 A chemical supply company currently has in stock 100 lb of a certain chemical, which it sells to customers in 5-lb lots. Let $x =$ the number of lots ordered by a randomly chosen customer. The probability distribution of x is as follows:

| | | | | |
|--------|----|----|----|----|
| x | 1 | 2 | 3 | 4 |
| $p(x)$ | .2 | .4 | .3 | .1 |

- a. Calculate the mean value of x .
 b. Calculate the variance and standard deviation of x .

7.39 Return to Exercise 7.38, and let y denote the amount of material (in pounds) left after the next customer's order is shipped. Find the mean and variance of y . (Hint: y is a linear function of x .)

7.40 An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space. Let x = the amount of storage space purchased by the next customer to buy a freezer. Suppose that x has the following probability distribution:

| | | | |
|--------|------|------|------|
| x | 13.5 | 15.9 | 19.1 |
| $p(x)$ | .2 | .5 | .3 |

- a. Calculate the mean and standard deviation of x .
 b. If the price of the freezer depends on the size of the storage space, x , such that $\text{Price} = 25x - 8.5$, what is the mean value of the variable *Price* paid by the next customer?
 c. What is the standard deviation of the price paid?

7.41 ▼ To assemble a piece of furniture, a wood peg must be inserted into a predrilled hole. Suppose that the diameter of a randomly selected peg is a random variable with mean 0.25 in. and standard deviation 0.006 in. and that the diameter of a randomly selected hole is a random variable with mean 0.253 in. and standard deviation 0.002 in. Let x_1 = peg diameter, and let x_2 = denote hole diameter.

- a. Why would the random variable y , defined as $y = x_2 - x_1$, be of interest to the furniture manufacturer?
 b. What is the mean value of the random variable y ?
 c. Assuming that x_1 and x_2 are independent, what is the standard deviation of y ?
 d. Is it reasonable to think that x_1 and x_2 are independent? Explain.
 e. Based on your answers to Parts (b) and (c), do you think that finding a peg that is too big to fit in the predrilled hole would be a relatively common or a relatively rare occurrence? Explain.

7.42 A multiple-choice exam consists of 50 questions. Each question has five choices, of which only one is correct. Suppose that the total score on the exam is computed as

$$y = x_1 - \frac{1}{4}x_2$$

where x_1 = number of correct responses and x_2 = number of incorrect responses. (Calculating a total score by

subtracting a term based on the number of incorrect responses is known as a correction for guessing and is designed to discourage test takers from choosing answers at random.)

- a. It can be shown that if a totally unprepared student answers all 50 questions by just selecting one of the five answers at random, then $\mu_{x_1} = 10$ and $\mu_{x_2} = 40$. What is the mean value of the total score, y ? Does this surprise you? Explain.
 b. Explain why it is unreasonable to use the formulas given in this section to compute the variance or standard deviation of y .

7.43 Consider a large ferry that can accommodate cars and buses. The toll for cars is \$3, and the toll for buses is \$10. Let x and y denote the number of cars and buses, respectively, carried on a single trip. Cars and buses are accommodated on different levels of the ferry, so the number of buses accommodated on any trip is independent of the number of cars on the trip. Suppose that x and y have the following probability distributions:

| | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|
| x | 0 | 1 | 2 | 3 | 4 | 5 |
| $p(x)$ | .05 | .10 | .25 | .30 | .20 | .10 |
| y | 0 | 1 | 2 | | | |
| $p(y)$ | .50 | .30 | .20 | | | |

- a. Compute the mean and standard deviation of x .
 b. Compute the mean and standard deviation of y .
 c. Compute the mean and variance of the total amount of money collected in tolls from cars.
 d. Compute the mean and variance of the total amount of money collected in tolls from buses.
 e. Compute the mean and variance of z = total number of vehicles (cars and buses) on the ferry.
 f. Compute the mean and variance of w = total amount of money collected in tolls.

7.44 Consider a game in which a red die and a blue die are rolled. Let x_R denote the value showing on the uppermost face of the red die, and define x_B similarly for the blue die.

- a. The probability distribution of x_R is

| | | | | | | |
|----------|-----|-----|-----|-----|-----|-----|
| x_R | 1 | 2 | 3 | 4 | 5 | 6 |
| $p(x_R)$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

Find the mean, variance, and standard deviation of x_R .

- b. What are the values of the mean, variance, and standard deviation of x_B ? (You should be able to answer this question without doing any additional calculations.)

c. Suppose that you are offered a choice of the following two games:

Game 1: Costs \$7 to play, and you win y_1 dollars, where $y_1 = x_R + x_B$.

Game 2: Doesn't cost anything to play initially, but you "win" $3y_2$ dollars, where $y_2 = x_R - x_B$. If y_2 is negative, you must pay that amount; if it is positive, you receive that amount.

For Game 1, the net amount won in a game is $w_1 = y_1 - 7 = x_R + x_B - 7$. What are the mean and standard deviation of w_1 ?

d. For Game 2, the net amount won in a game is $w_2 = 3y_2 = 3(x_R - x_B)$. What are the mean and standard deviation of w_2 ?

e. Based on your answers to Parts (c) and (d), if you had to play, which game would you choose and why?

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.5

Binomial and Geometric Distributions

In this section we introduce two of the more commonly encountered discrete probability distributions: the binomial distribution and the geometric distribution. These distributions arise when the experiment of interest consists of making a sequence of dichotomous observations (two possible values for each observation). The process of making a single such observation is called a *trial*. For example, one characteristic of blood type is Rh factor, which can be either positive or negative. We can think of an experiment that consists of noting the Rh factor for each of 25 blood donors as a sequence of 25 dichotomous trials, where each trial consists of observing the Rh factor (positive or negative) of a single donor.

We could also conduct a different experiment that consists of observing the Rh factor of blood donors until a donor who is Rh-negative is encountered. This second experiment can also be viewed as a sequence of dichotomous trials, but the total number of trials in this experiment is not predetermined, as it was in the previous example, where we knew in advance that there would be 25 trials. Experiments of the two types just described are characteristic of those leading to the binomial and the geometric probability distributions, respectively.

■ Binomial Distributions

Suppose that we decide to record the gender of each of the next 25 newborn children at a particular hospital. What is the chance that at least 15 are female? What is the chance that between 10 and 15 are female? How many among the 25 can we expect to be female? These and other similar questions can be answered by studying the *binomial probability distribution*. This distribution arises when the experiment of interest is a *binomial experiment*, that is, an experiment having the characteristics listed in the following box.

Properties of a Binomial Experiment

A binomial experiment consists of a sequence of trials with the following conditions:

1. There are a fixed number of observations called trials.
2. Each trial can result in one of only two mutually exclusive outcomes labeled success (S) and failure (F).
3. Outcomes of different trials are independent.
4. The probability that a trial results in S is the same for each trial.

The **binomial random variable** x is defined as

x = number of successes observed when a binomial experiment is performed

The probability distribution of x is called the *binomial probability distribution*.

The term *success* here does not necessarily have any of its usual connotations. Which of the two possible outcomes is labeled “success” is determined by the random variable of interest. For example, if the variable counts the number of female births among the next 25 births at a particular hospital, then a female birth would be labeled a success (because this is what the variable counts). If male births were counted instead, a male birth would be labeled a success and a female birth a failure.

One illustration of a binomial probability distribution was given in Example 7.5. There, we considered x = number among four customers who selected an electric (as opposed to gas) hot tub. This is a binomial experiment with four trials and $P(\text{success}) = P(E) = .4$. The 16 possible outcomes, along with their probabilities, were displayed in Table 7.1.

Consider now the case of five customers, a binomial experiment with five trials. Here the binomial distribution tells us the probability associated with each of the possible x values 0, 1, 2, 3, 4, and 5. There are 32 possible outcomes, and 5 of them yield $x = 1$: *SFFFF*, *FSFFF*, *FFSFF*, *FFFSF*, and *FFFFS*.

By independence, the first of these outcomes has probability

$$\begin{aligned} P(SFFFF) &= P(S)P(F)P(F)P(F)P(F) \\ &= (.4)(.6)(.6)(.6)(.6) \\ &= (.4)(.6)^4 \\ &= .05184 \end{aligned}$$

The probability calculation will be the same for any outcome with only one success ($x = 1$). It does not matter where in the sequence the single success occurs. Thus

$$\begin{aligned} p(1) &= P(x = 1) \\ &= P(SFFFF \text{ or } FSFFF \text{ or } FFSFF \text{ or } FFFSF \text{ or } FFFFF) \\ &= .05184 + .05184 + .05184 + .05184 + .05184 \\ &= (5)(.05184) \\ &= .25920 \end{aligned}$$

Similarly, there are ten outcomes for which $x = 2$, because there are 10 ways to select two from among the five trials to be the S 's: *SSFFF*, *SFSFF*, . . . , and *FFFSF*. The probability of each results from multiplying together (.4) two times and (.6) three times. For example,

$$\begin{aligned} P(SSFFF) &= (.4)(.4)(.6)(.6)(.6) \\ &= (.4)^2(.6)^3 \\ &= .03456 \end{aligned}$$

and so

$$\begin{aligned} p(2) &= P(x = 2) \\ &= P(SSFFF) + \cdots + P(FFFSF) \\ &= (10)(.4)^2(.6)^3 \\ &= .34560 \end{aligned}$$

The general form of the distribution here is

$$\begin{aligned} p(x) &= P(x \text{ S's among the five trials}) \\ &= (\text{no. of outcomes with } x \text{ S's}) \cdot (\text{probability of any particular outcome with } x \text{ S's}) \\ &= (\text{no. of outcomes with } x \text{ S's}) \cdot (.4)^x (.6)^{5-x} \end{aligned}$$

This form was seen previously where $p(2) = 10(.4)^2(.6)^3$.

Let n denote the number of trials in the experiment. Then the number of outcomes with x S's is the number of ways of selecting x from among the n trials to be the success trials. A simple expression for this quantity is

$$\text{number of outcomes with } x \text{ successes} = \frac{n!}{x!(n-x)!}$$

where, for any positive whole number m , the symbol $m!$ (read “ m factorial”) is defined by

$$m! = m(m-1)(m-2) \cdots (2)(1)$$

and $0! = 1$.

The Binomial Distribution

Let

n = number of independent trials in a binomial experiment

π = constant probability that any particular trial results in a success*

Then

$$\begin{aligned} p(x) &= P(x \text{ successes among } n \text{ trials}) \\ &= \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x} \quad x = 0, 1, 2, \dots, n \end{aligned}$$

The expressions $\binom{n}{x}$ or ${}_n C_x$ are sometimes used in place of $\frac{n!}{x!(n-x)!}$. Both are read as “ n choose x ” and represent the number of ways of choosing x items from a set of n . The binomial probability function can then be written as

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad x = 0, 1, 2, \dots, n$$

or

$$p(x) = {}_n C_x \pi^x (1-\pi)^{n-x} \quad x = 0, 1, 2, \dots, n$$

*Some sources use p to represent the probability of success rather than π . We prefer the use of Greek letters for characteristics of a population or probability distribution, thus the use of π .

Notice that the probability distribution is specified using a formula that allows calculation of the various probabilities rather than by giving a table or a probability histogram.

Example 7.17 Computer Monitors

Sixty percent of all computer monitors sold by a large computer retailer have a flat panel display and 40% have a CRT display. The type of monitor purchased by each of the next 12 customers will be noted. Define a random variable x by

x = number of monitors among these 12 that have a flat panel display

Because x counts the number of flat panel displays, we use S to denote the sale of a flat panel monitor. Then x is a binomial random variable with $n = 12$ and $\pi = P(S) = .60$. The probability distribution of x is given by

$$p(x) = \frac{12!}{x!(12-x)!} (.6)^x (.4)^{12-x} \quad x = 0, 1, 2, \dots, 12$$

The probability that exactly four monitors are flat panel displays is

$$\begin{aligned} p(4) &= P(x = 4) \\ &= \frac{12!}{4!8!} (.6)^4 (.4)^8 \\ &= (495) (.6)^4 (.4)^8 \\ &= .042 \end{aligned}$$

If group after group of 12 purchases is examined, the long-run percentage of those with exactly four flat panel monitors will be 4.2%. According to this calculation, 495 of the possible outcomes (there are $2^{12} = 4096$) have $x = 4$.

The probability that between four and seven (inclusive) are flat panel displays is

$$P(4 \leq x \leq 7) = P(x = 4 \text{ or } x = 5 \text{ or } x = 6 \text{ or } x = 7)$$

Since these outcomes are disjoint, this is equal to

$$\begin{aligned} P(4 \leq x \leq 7) &= p(4) + p(5) + p(6) + p(7) \\ &= \frac{12!}{4!8!} (.6)^4 (.4)^8 + \dots + \frac{12!}{7!5!} (.6)^7 (.4)^5 \\ &= .042 + .101 + .177 + .227 \\ &= .547 \end{aligned}$$

Notice that

$$\begin{aligned} P(4 < x < 7) &= P(x = 5 \text{ or } x = 6) \\ &= p(5) + p(6) \\ &= .278 \end{aligned}$$

so the probability depends on whether $<$ or \leq appears. (This is typical of *discrete* random variables.)

The binomial distribution formula can be tedious to use unless n is small. Appendix Table 9 gives binomial probabilities for selected n in combination with various values of π . Appendix Table 9 should help you practice using the binomial distribution without getting bogged down in arithmetic.

Using Appendix Table 9

To find $p(x)$ for any particular value of x ,

1. Locate the part of the table corresponding to your value of n (5, 10, 15, 20, or 25).
2. Move down to the row labeled with your value of x .
3. Go across to the column headed by the specified value of π .

The desired probability is at the intersection of the designated x row and π column. For example, when $n = 20$ and $\pi = .8$,

$$p(15) = P(x = 15) = (\text{entry at intersection of } n = 15 \text{ row and } \pi = .8 \text{ column}) = .175$$

Although $p(x)$ is positive for every possible x value, many probabilities are zero to three decimal places, so they appear as .000 in the table. More extensive binomial tables are available. Alternatively, most statistics software packages and graphing calculators are programmed to calculate these probabilities.

■ **Sampling Without Replacement** Usually, sampling is carried out without replacement; that is, once an element has been selected for the sample, it is not a candidate for future selection. If the sampling was accomplished by selecting an element from the population, observing whether it is a success or a failure, and then returning it to the population before the next selection is made, the variable $x =$ number of successes observed in the sample would fit all the requirements of a binomial random variable. When sampling is done without replacement, the trials (individual selections) are not independent. In this case, the number of successes observed in the sample does not have a binomial distribution but rather a different type of distribution called a *hypergeometric distribution*. The probability calculations for this distribution are even more tedious than for the binomial distribution. Fortunately, when the sample size n is much smaller than N , the population size, probabilities calculated using the binomial distribution and the hypergeometric distribution are very close in value. They are so close, in fact, that statisticians often ignore the difference and use the binomial probabilities in place of the hypergeometric probabilities. Most statisticians recommend the following guideline for determining whether the binomial probability distribution is appropriate when sampling without replacement.

Let x denote the number of S 's in a sample of size n selected without replacement from a population consisting of N individuals or objects. If $(n/N) \leq 0.05$, i.e., at most 5% of the population is sampled, then the binomial distribution gives a good approximation to the probability distribution of x .

Example 7.18 Security Systems

A *Los Angeles Times* poll (November 10, 1991) reported that almost 20% of Southern California homeowners questioned had installed a home security system. Suppose that exactly 20% of all such homeowners have a system. Consider a random

sample of $n = 20$ homeowners (much less than 5% of the population). Then x , the number of homeowners in the sample who have a security system, has (approximately) a binomial distribution with $n = 20$ and $\pi = .20$. The probability that five of those sampled have a system is

$$\begin{aligned} p(5) &= P(x = 5) \\ &= (\text{entry in } x \text{ row and } \pi = .20 \text{ column in Appendix Table 9 } (n = 20)) \\ &= .175 \end{aligned}$$

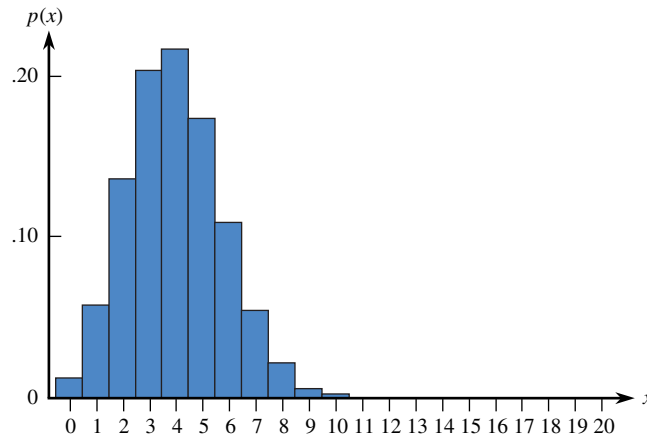
The probability that at least 40% of those in the sample—that is, eight or more—have a system is

$$\begin{aligned} P(x \geq 8) &= P(x = 8, 9, 10, \dots, 19, \text{ or } 20) \\ &= p(8) + p(9) + \dots + p(20) \\ &= .022 + .007 + .002 + .000 + \dots + .000 \\ &= .031 \end{aligned}$$

If, in fact, $\pi = .20$, only about 3% of all samples of size 20 would result in at least 8 homeowners having a security system. Because $P(x \geq 8)$ is so small when $\pi = .20$, if $x \geq 8$ were actually observed, we would have to wonder whether the reported value of $\pi = .20$ is correct. Although it is possible that we could observe $x \geq 8$ when $\pi = .20$ (this would happen about 3% of the time in the long run), it might also be the case that π is actually greater than .20. In Chapter 10, we show how hypothesis-testing methods can be used to decide which of two contradictory claims about a population (e.g., $\pi = .20$ or $\pi > .20$) is more plausible.

The binomial formula or tables can be used to compute each of the 21 probabilities $p(0), p(1), \dots, p(20)$. Figure 7.16 shows the probability histogram for the binomial distribution with $n = 20$ and $\pi = .20$. Notice that the distribution is skewed to the right. (The binomial distribution is symmetric only when $\pi = .5$.)

Figure 7.16 The binomial probability histogram when $n = 20$ and $\pi = .20$.



■ **Mean and Standard Deviation of a Binomial Random Variable** A binomial random variable x based on n trials has possible values $0, 1, 2, \dots, n$, so the mean value is

$$\mu_x = \sum xp(x) = (0)p(0) + (1)p(1) + \dots + (n)p(n)$$

and the variance of x is

$$\begin{aligned}\sigma_x^2 &= \sum (x - \mu_x)^2 \cdot p(x) \\ &= (0 - \mu_x)^2 p(0) + (1 - \mu_x)^2 p(1) + \cdots + (n - \mu_x)^2 p(n)\end{aligned}$$

These expressions appear to be very tedious to evaluate for any particular values of n and π . Fortunately, algebraic manipulation results in considerable simplification, making summation unnecessary.

The mean value and the standard deviation of a binomial random variable are, respectively,

$$\mu_x = n\pi \quad \text{and} \quad \sigma_x = \sqrt{n\pi(1 - \pi)}$$

Example 7.19 Credit Cards Paid in Full

Newsweek (December 2, 1991) reported that one-third of all credit card users pay their bills in full each month. This figure is, of course, an average across different cards and issuers. Suppose that 30% of all individuals holding Visa cards issued by a certain bank pay in full each month. A random sample of $n = 25$ cardholders is to be selected. The bank is interested in the variable $x =$ number in the sample who pay in full each month. Even though sampling is done without replacement, the sample size $n = 25$ is most likely very small compared to the total number of credit card holders, so we can approximate the probability distribution of x using a binomial distribution with $n = 25$ and $\pi = .3$. We have defined “paid in full” as a success because this is the outcome counted by the random variable x . The mean value of x is then

$$\mu_x = n\pi = 25(.30) = 7.5$$

and the standard deviation is

$$\sigma_x = \sqrt{n\pi(1 - \pi)} = \sqrt{25(.30)(.70)} = \sqrt{5.25} = 2.29$$

The probability that x is farther than 1 standard deviation from its mean value is

$$\begin{aligned}P(x < \mu_x - \sigma_x \text{ or } x > \mu_x + \sigma_x) &= P(x < 5.21 \text{ or } x > 9.79) \\ &= P(x \leq 5) + P(x \geq 10) \\ &= p(0) + \cdots + p(5) + p(10) + \cdots + p(25) \\ &= .382 \quad (\text{using Appendix Table 9})\end{aligned}$$

The value of σ_x is 0 when $\pi = 0$ or $\pi = 1$. In these two cases, there is no uncertainty in x : We are sure to observe $x = 0$ when $\pi = 0$ and $x = n$ when $\pi = 1$. It is also easily verified that $\pi(1 - \pi)$ is largest when $\pi = .5$. Thus the binomial distribution spreads out the most when sampling from a 50–50 population. The farther π is from .5, the less spread out and the more skewed the distribution.

■ Geometric Distributions

A binomial random variable is defined as the number of successes in n independent trials, where each trial can result in either a success or a failure and the probability of success is the same for each trial. Suppose, however, that we are not interested in the number of successes in a fixed number of trials but rather in the number of trials that must be carried out before a success occurs. Two examples are counting the number of boxes of cereal that must be purchased before finding one with a rare toy and counting the number of games that a professional bowler must play before achieving a score over 250.

The variable

$$x = \text{number of trials to first success}$$

is called a *geometric random variable*, and the probability distribution that describes its behavior is called a *geometric probability distribution*.

Suppose an experiment consists of a sequence of trials with the following conditions:

1. The trials are independent.
2. Each trial can result in one of two possible outcomes, success and failure.
3. The probability of success is the same for all trials.

A **geometric random variable** is defined as

$$x = \text{number of trials until the first success is observed (including the success trial)}$$

The probability distribution of x is called the *geometric probability distribution*.

For example, suppose that 40% of the students who drive to campus at your university carry jumper cables. Your car has a dead battery and you don't have jumper cables, so you decide to stop students who are headed to the parking lot and ask them whether they have a pair of jumper cables. You might be interested in the number of students you would have to stop before finding one who has jumper cables. If we define success as a student with jumper cables, a trial would consist of asking an individual student for help. The random variable $x = \text{number of students who must be stopped before finding one with jumper cables}$ is an example of a geometric random variable, because it can be viewed as the number of trials to the first success in a sequence of independent trials.

The probability distribution of a geometric random variable is easy to construct. We use π to denote the probability of success on any given trial. Possible outcomes can be denoted as follows:

| Outcome | $x = \text{Number of Trials to First Success}$ |
|-----------|--|
| S | 1 |
| FS | 2 |
| FFS | 3 |
| \vdots | \vdots |
| $FFFFFFS$ | 7 |
| \vdots | \vdots |

Each possible outcome consists of 0 or more failures followed by a single success. So,

$$\begin{aligned} p(x) &= P(x \text{ trials to first success}) \\ &= P(FF \dots FS) \end{aligned}$$

↑
 $x - 1$ failures followed by a success on trial x

Because the probability of success is π for each trial, the probability of failure for each trial is $1 - \pi$. Because the trials are independent,

$$\begin{aligned} p(x) &= P(x \text{ trials to first success}) = P(FF \dots FS) \\ &= P(F)P(F) \cdots P(F)P(S) \\ &= (1 - \pi)(1 - \pi) \cdots (1 - \pi)\pi \\ &= (1 - \pi)^{x-1}\pi \end{aligned}$$

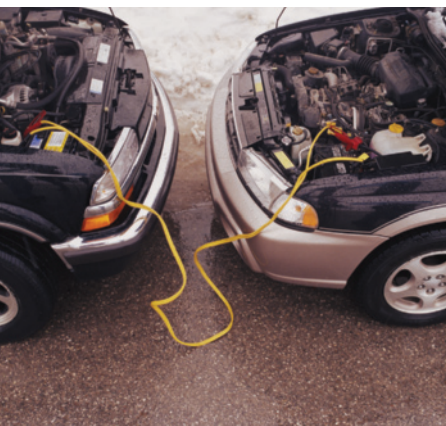
This leads us to the formula for the geometric probability distribution.

Geometric Probability Distribution

If x is a geometric random variable with probability of success = π for each trial, then

$$p(x) = (1 - \pi)^{x-1}\pi \quad x = 1, 2, 3, \dots$$

Example 7.20 Jumper Cables



© Stone +/Greg Ceo/Getty Images

Consider the jumper cable problem described previously. For this problem, $\pi = .4$, because 40% of the students who drive to campus carry jumper cables. The probability distribution of

x = number of students who must be stopped before finding a student with jumper cables

is

$$p(x) = (.6)^{x-1}(.4) \quad x = 1, 2, 3, \dots$$

The probability distribution can now be used to compute various probabilities. For example, the probability that the first student stopped has jumper cables (i.e., $x = 1$) is

$$p(1) = (.6)^{1-1}(.4) = (.6)^0(.4) = .4$$

The probability that three or fewer students must be stopped is

$$\begin{aligned} P(x \leq 3) &= p(1) + p(2) + p(3) \\ &= (.6)^0(.4) + (.6)^1(.4) + (.6)^2(.4) \\ &= .4 + .24 + .144 \\ &= .784 \end{aligned}$$

Exercises 7.45–7.63

7.45 Consider two binomial experiments.

- The first binomial experiment consists of six trials. How many outcomes have exactly one success, and what are these outcomes?
- The second binomial experiment consists of 20 trials. How many outcomes have exactly 10 successes? exactly 15 successes? exactly 5 successes?

7.46 Suppose that in a certain metropolitan area, 9 out of 10 households have a VCR. Let x denote the number among four randomly selected households that have a VCR, so x is a binomial random variable with $n = 4$ and $\pi = .9$.

- Calculate $p(2) = P(x = 2)$, and interpret this probability.
- Calculate $p(4)$, the probability that all four selected households have a VCR.
- Determine $P(x \leq 3)$.

7.47 ▼ The *Los Angeles Times* (December 13, 1992) reported that what airline passengers like to do most on long flights is rest or sleep; in a survey of 3697 passengers, almost 80% did so. Suppose that for a particular route the actual percentage is exactly 80%, and consider randomly selecting six passengers. Then x , the number among the selected six who rested or slept, is a binomial random variable with $n = 6$ and $\pi = .8$.

- Calculate $p(4)$, and interpret this probability.
- Calculate $p(6)$, the probability that all six selected passengers rested or slept.
- Determine $P(x \geq 4)$.

7.48 Refer to Exercise 7.47, and suppose that 10 rather than 6 passengers are selected ($n = 10$, $\pi = .8$), so that Appendix Table 9 can be used.

- What is $p(8)$?
- Calculate $P(x \leq 7)$.
- Calculate the probability that more than half of the selected passengers rested or slept.

7.49 Twenty-five percent of the customers entering a grocery store between 5 P.M. and 7 P.M. use an express checkout. Consider five randomly selected customers, and let x denote the number among the five who use the express checkout.

- What is $p(2)$, that is, $P(x = 2)$?
- What is $P(x \leq 1)$?
- What is $P(2 \leq x)$? (Hint: Make use of your computation in Part (b).)
- What is $P(x \neq 2)$?

7.50 A breeder of show dogs is interested in the number of female puppies in a litter. If a birth is equally likely to result in a male or a female puppy, give the probability distribution of the variable $x =$ number of female puppies in a litter of size 5.

7.51 The article “FBI Says Fewer than 25 Failed Polygraph Test” (*San Luis Obispo Tribune*, July 29, 2001) states that false-positives in polygraph tests (i.e., tests in which an individual fails even though he or she is telling the truth) are relatively common and occur about 15% of the time. Suppose that such a test is given to 10 trustworthy individuals.

- What is the probability that all 10 pass?
- What is the probability that more than 2 fail, even though all are trustworthy?
- The article indicated that 500 FBI agents were required to take a polygraph test. Consider the random variable $x =$ number of the 500 tested who fail. If all 500 agents tested are trustworthy, what are the mean and standard deviation of x ?
- The headline indicates that fewer than 25 of the 500 agents tested failed the test. Is this a surprising result if all 500 are trustworthy? Answer based on the values of the mean and standard deviation from Part (c).

7.52 Industrial quality control programs often include inspection of incoming materials from suppliers. If parts are purchased in large lots, a typical plan might be to select 20 parts at random from a lot and inspect them. A lot might be judged acceptable if one or fewer defective parts are found among those inspected. Otherwise, the lot is rejected and returned to the supplier. Use Appendix Table 9 to find the probability of accepting lots that have each of the following (Hint: Identify success with a defective part):

- 5% defective parts
- 10% defective parts
- 20% defective parts

7.53 An experiment was conducted to investigate whether a graphologist (a handwriting analyst) could distinguish a normal person’s handwriting from that of a psychotic. A well-known expert was given 10 files, each containing handwriting samples from a normal person and from a person diagnosed as psychotic, and asked to identify the psychotic’s handwriting. The graphologist made correct identifications in 6 of the 10 trials (data taken from *Statistics in the Real World*, by R. J. Larsen and D. F. Stroup [New York: Macmillan, 1976]). Does this evidence indi-

cate that the graphologist has an ability to distinguish the handwriting of psychotics? (Hint: What is the probability of correctly guessing 6 or more times out of 10? Your answer should depend on whether this probability is relatively small or relatively large.)

7.54 Suppose that the probability is .1 that any given citrus tree will show measurable damage when the temperature falls to 30°F. If the temperature does drop to 30°F, what is the expected number of citrus trees showing damage in orchards of 2000 trees? What is the standard deviation of the number of trees that show damage?

7.55 Thirty percent of all automobiles undergoing an emissions inspection at a certain inspection station fail the inspection.

- Among 15 randomly selected cars, what is the probability that at most 5 fail the inspection?
- Among 15 randomly selected cars, what is the probability that between 5 and 10 (inclusive) fail to pass inspection?
- Among 25 randomly selected cars, what is the mean value of the number that pass inspection, and what is the standard deviation of the number that pass inspection?
- What is the probability that among 25 randomly selected cars, the number that pass is within 1 standard deviation of the mean value?

7.56 You are to take a multiple-choice exam consisting of 100 questions with 5 possible responses to each question. Suppose that you have not studied and so must guess (select one of the five answers in a completely random fashion) on each question. Let x represent the number of correct responses on the test.

- What kind of probability distribution does x have?
- What is your expected score on the exam? (Hint: Your expected score is the mean value of the x distribution.)
- Compute the variance and standard deviation of x .
- Based on your answers to Parts (b) and (c), is it likely that you would score over 50 on this exam? Explain the reasoning behind your answer.

7.57 Suppose that 20% of the 10,000 signatures on a certain recall petition are invalid. Would the number of invalid signatures in a sample of size 1000 have (approximately) a binomial distribution? Explain.

7.58 A coin is spun 25 times. Let x be the number of spins that result in heads (H). Consider the following rule for deciding whether or not the coin is fair:

Judge the coin fair if $8 \leq x \leq 17$.

Judge the coin biased if either $x \leq 7$ or $x \geq 18$.

a. What is the probability of judging the coin biased when it is actually fair?

b. What is the probability of judging the coin fair when $P(H) = .9$, so that there is a substantial bias? Repeat for $P(H) = .1$.

c. What is the probability of judging the coin fair when $P(H) = .6$? when $P(H) = .4$? Why are the probabilities so large compared to the probabilities in Part (b)?

d. What happens to the “error probabilities” of Parts (a) and (b) if the decision rule is changed so that the coin is judged fair if $7 \leq x \leq 18$ and unfair otherwise? Is this a better rule than the one first proposed? Explain.

7.59 A city ordinance requires that a smoke detector be installed in all residential housing. There is concern that too many residences are still without detectors, so a costly inspection program is being contemplated. Let π be the proportion of all residences that have a detector. A random sample of 25 residences is selected. If the sample strongly suggests that $\pi < .80$ (less than 80% have detectors), as opposed to $\pi \geq .80$, the program will be implemented. Let x be the number of residences among the 25 that have a detector, and consider the following decision rule: Reject the claim that $\pi = .8$ and implement the program if $x \leq 15$.

- What is the probability that the program is implemented when $\pi = .80$?
- What is the probability that the program is not implemented if $\pi = .70$? if $\pi = .60$?
- How do the “error probabilities” of Parts (a) and (b) change if the value 15 in the decision rule is changed to 14?

7.60 Suppose that 90% of all registered California voters favor banning the release of information from exit polls in presidential elections until after the polls in California close. A random sample of 25 California voters is to be selected.

- What is the probability that more than 20 voters favor the ban?
- What is the probability that at least 20 voters favor the ban?
- What are the mean value and standard deviation of the number of voters who favor the ban?
- If fewer than 20 voters in the sample favor the ban, is this at odds with the assertion that (at least) 90% of the populace favors the ban? (Hint: Consider $P(x < 20)$ when $\pi = .9$.)

7.61 Sophie is a dog that loves to play catch. Unfortunately, she isn’t very good, and the probability that she

catches a ball is only .1. Let x be the number of tosses required until Sophie catches a ball.

- Does x have a binomial or a geometric distribution?
- What is the probability that it will take exactly two tosses for Sophie to catch a ball?
- What is the probability that more than three tosses will be required?

7.62 Suppose that 5% of cereal boxes contain a prize and the other 95% contain the message, “Sorry, try again.” Consider the random variable x , where x = number of boxes purchased until a prize is found.

- What is the probability that at most two boxes must be purchased?
- What is the probability that exactly four boxes must be purchased?

- What is the probability that more than four boxes must be purchased?

7.63 ▼ The article on polygraph testing of FBI agents referenced in Exercise 7.51 indicated that the probability of a false-positive (a trustworthy person who nonetheless fails the test) is .15. Let x be the number of trustworthy FBI agents tested until someone fails the test.

- What is the probability distribution of x ?
- What is the probability that the first false-positive will occur when the third person is tested?
- What is the probability that fewer than four are tested before the first false-positive occurs?
- What is the probability that more than three agents are tested before the first false-positive occurs?

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.6 Normal Distributions

Normal distributions formalize the notion of mound-shaped histograms introduced in Chapter 4. Normal distributions are widely used for two reasons. First, they provide a reasonable approximation to the distribution of many different variables. They also play a central role in many of the inferential procedures that will be discussed in later chapters.

Normal distributions are continuous probability distributions that are bell shaped and symmetric, as shown in Figure 7.17. Normal distributions are sometimes referred to as a *normal curves*.

There are many different normal distributions, and they are distinguished from one another by their mean μ and standard deviation σ . The mean μ of a normal distribution describes where the corresponding curve is centered, and the standard deviation σ describes how much the curve spreads out around that center. As with all continuous probability distributions, the total area under any normal curve is equal to 1. Three normal distributions are shown in Figure 7.18. Notice that the smaller the standard deviation, the taller and narrower the corresponding curve. Recall that areas under a continuous probability distribution curve represent probabilities, so when the standard deviation is small, a larger area is concentrated near the center of the curve and the chance of observing a value near the mean is much greater (because μ is at the center).

The value of μ is the number on the measurement axis lying directly below the top of the bell. The value of σ can also be ascertained from a picture of the curve. Consider the normal curve in Figure 7.19. Starting at the top of the bell (above $\mu = 100$) and moving to the right, the curve turns downward until it is above the value 110. After that point, it continues to decrease in height but is turning upward rather than downward. Similarly, to the left of $\mu = 100$, the curve turns downward until it reaches 90 and then begins to turn upward. The curve changes from turning downward to turning upward at a distance of 10 on either side of μ , so $\sigma = 10$. In general, σ is the distance to either side of μ at which a normal curve changes from turning downward to turning upward.



Figure 7.17 A normal distribution.

Figure 7.18 Three normal distributions.

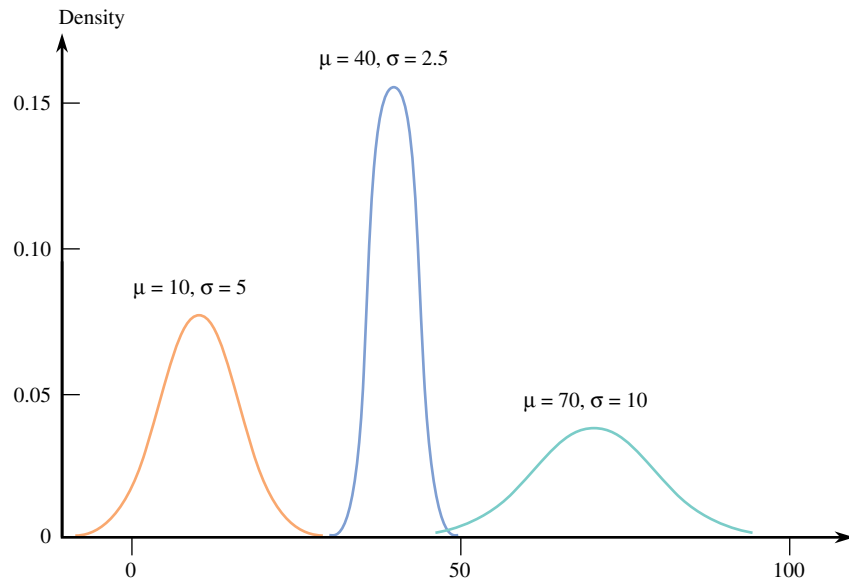
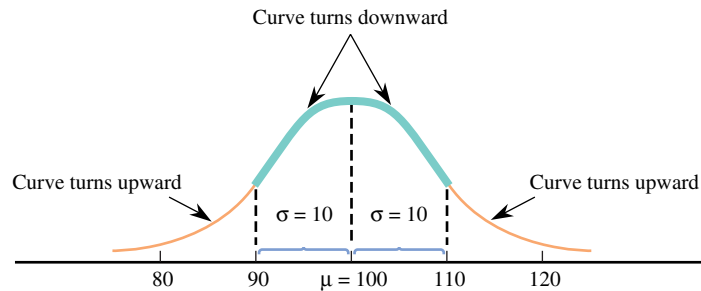
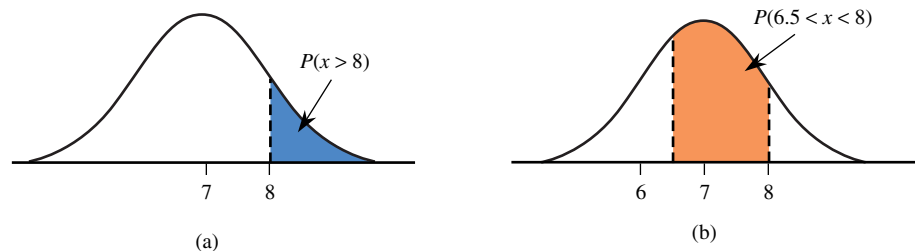


Figure 7.19 Mean μ and standard deviation σ for a normal curve.



If a particular normal distribution is to be used to describe the behavior of a random variable, a mean and a standard deviation must be specified. For example, a normal distribution with mean 7 and standard deviation 1 might be used as a model for the distribution of $x =$ birth weight. If this model is a reasonable description of the probability distribution, we could use areas under the normal curve with $\mu = 7$ and $\sigma = 1$ to approximate various probabilities related to birth weight. The probability that a birth weight is over 8 lb (expressed symbolically as $P(x > 8)$) corresponds to the shaded area in Figure 7.20(a). The shaded area in Figure 7.20(b) is the (approximate) probability $P(6.5 < x < 8)$ of a birth weight falling between 6.5 and 8 lb.

Figure 7.20 Normal distribution for birth weight:
 (a) shaded area = $P(x > 8)$;
 (b) shaded area = $P(6.5 < x < 8)$.



Unfortunately, direct computation of such probabilities (areas under a normal curve) is not simple. To overcome this difficulty, we rely on a table of areas for a reference normal distribution, called the *standard normal distribution*.

DEFINITION

The **standard normal distribution** is the normal distribution with

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

The corresponding density curve is called the *standard normal curve*. It is customary to use the letter z to represent a variable whose distribution is described by the standard normal curve. The term z curve is often used in place of *standard normal curve*.

Few naturally occurring variables have distributions that are well described by the standard normal distribution, but this distribution is important because it is also used in probability calculations for other normal distributions. When we are interested in finding a probability based on some other normal curve, we first translate our problem into an equivalent problem that involves finding an area under the standard normal curve. A table for the standard normal distribution is then used to find the desired area. To be able to do this, we must first learn to work with the standard normal distribution.

■ The Standard Normal Distribution

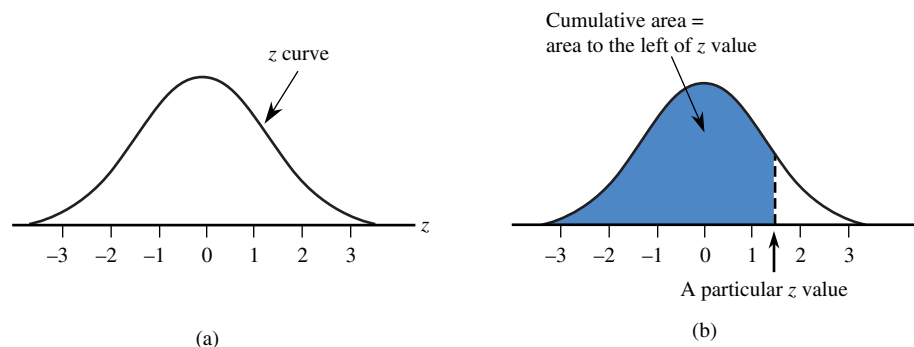
In working with normal distributions, we need two general skills:

1. We must be able to use the normal distribution to compute probabilities, which are areas under a normal curve and above given intervals.
2. We must be able to characterize extreme values in the distribution, such as the largest 5%, the smallest 1%, and the most extreme 5% (which would include the largest 2.5% and the smallest 2.5%).

Let's begin by looking at how to accomplish these tasks when the distribution of interest is the standard normal distribution.

The standard normal or z curve is shown in Figure 7.21(a). It is centered at $\mu = 0$, and the standard deviation, $\sigma = 1$, is a measure of the extent to which it spreads out

Figure 7.21 (a) A standard normal (z) curve and (b) a cumulative area.



about its mean (in this case, 0). Note that this picture is consistent with the Empirical Rule of Chapter 4: About 95% of the area (probability) is associated with values that are within 2 standard deviations of the mean (between -2 and 2), and almost all of the area is associated with values that are within 3 standard deviations of the mean (between -3 and 3).

Appendix Table 2 tabulates cumulative z curve areas of the sort shown in Figure 7.21(b) for many different values of z . The smallest value for which the cumulative area is given is -3.89 , a value far out in the lower tail of the z curve. The next smallest value for which the area appears is -3.88 , then -3.87 , then -3.86 , and so on in increments of 0.01 , terminating with the cumulative area to the left of 3.89 .

Using the Table of Standard Normal Curve Areas

For any number z^* between -3.89 and 3.89 and rounded to two decimal places, Appendix Table 2 gives

$$(\text{area under } z \text{ curve to the left of } z^*) = P(z < z^*) = P(z \leq z^*)$$

where the letter z is used to represent a random variable whose distribution is the standard normal distribution.

To find this probability, locate the following:

1. The row labeled with the sign of z^* and the digit to either side of the decimal point (for example, -1.7 or 0.5)
2. The column identified with the second digit to the right of the decimal point in z^* (for example, $.06$ if $z^* = -1.76$)

The number at the intersection of this row and column is the desired probability, $P(z < z^*)$.

A portion of the table of standard normal curve areas appears in Figure 7.22. To find the area under the z curve to the left of 1.42 , look in the row labeled 1.4 and the column labeled $.02$ (the highlighted row and column in Figure 7.22). From the table, the corresponding cumulative area is $.9222$. So

$$z \text{ curve area to the left of } 1.42 = .9222$$

We can also use the table to find the area to the right of a particular value. Because the total area under the z curve is 1 , it follows that

$$\begin{aligned} (z \text{ curve area to the right of } 1.42) &= 1 - (z \text{ curve area to the left of } 1.42) \\ &= 1 - .9222 \\ &= .0778 \end{aligned}$$

These probabilities can be interpreted to mean that in a long sequence of observations, roughly 92.22% of the observed z values will be smaller than 1.42 , and 7.78% will be larger than 1.42 .

Figure 7.22 Portion of the table of standard normal curve areas.

| z^* | .00 | .01 | .02 | .03 | .04 | .05 |
|-------|-------|-------|-------|-------|-------|-------|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 |

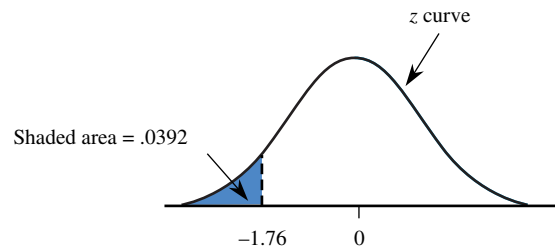
$P(z < 1.42)$

Example 7.21 Finding Standard Normal Curve Areas

The probability $P(z < -1.76)$ is found at the intersection of the -1.7 row and the $.06$ column of the z table. The result is

$$P(z < -1.76) = .0392$$

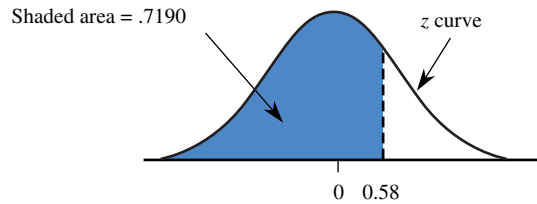
as shown in the following figure:



In other words, in a long sequence of observations, roughly 3.9% of the observed z values will be smaller than -1.76 . Similarly,

$$P(z \leq 0.58) = \text{entry in } 0.5 \text{ row and } .08 \text{ column of Table 2} = .7190$$

as shown in the following figure:



Now consider $P(z < -4.12)$. This probability does not appear in Appendix Table 2; there is no -4.1 row. However, it must be less than $P(z < -3.89)$, the smallest z value in the table, because -4.12 is farther out in the lower tail of the z curve. Since $P(z < -3.89) = .0000$ (that is, zero to four decimal places), it follows that

$$P(z < -4.12) \approx 0$$

Similarly,

$$P(z < 4.18) > P(z < 3.89) = 1.0000$$

from which we conclude that

$$P(z < 4.18) \approx 1$$



As illustrated in Example 7.21, we can use the cumulative areas tabulated in Appendix Table 2 to calculate other probabilities involving z . The probability that z is larger than a value c is

$$P(z > c) = \text{area under the } z \text{ curve to the right of } c = 1 - P(z \leq c)$$

In other words, the area to the right of a value (a right-tail area) is 1 minus the corresponding cumulative area. This is illustrated in Figure 7.23.

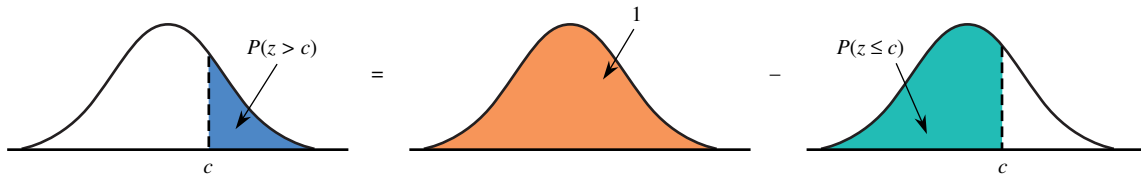


Figure 7.23 The relationship between an upper-tail area and a cumulative area.

Similarly, the probability that z falls in the interval between a lower limit a and an upper limit b is

$$\begin{aligned} P(a < z < b) &= \text{area under the } z \text{ curve and above the interval from } a \text{ to } b \\ &= P(z < b) - P(z < a) \end{aligned}$$

That is, $P(a < z < b)$ is the difference between two cumulative areas, as illustrated in Figure 7.24.

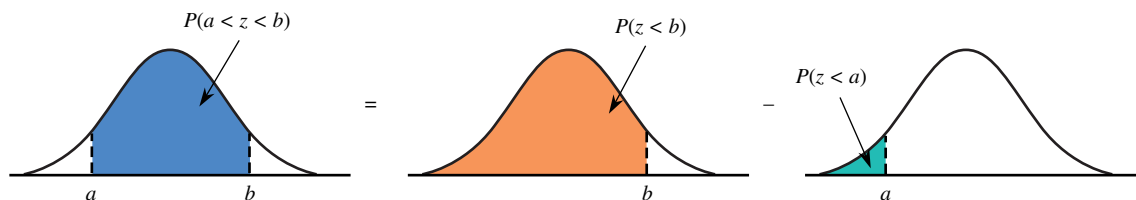


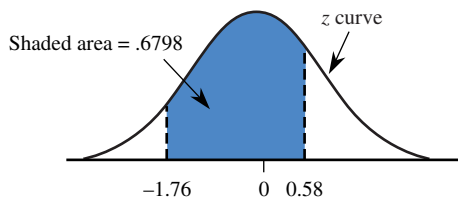
Figure 7.24 $P(a < z < b)$ as the difference between the two cumulative areas.

Example 7.22 More About Standard Normal Curve Areas

The probability that z is between -1.76 and 0.58 is

$$\begin{aligned} P(-1.76 < z < 0.58) &= P(z < 0.58) - P(z < -1.76) \\ &= .7190 - .0392 \\ &= .6798 \end{aligned}$$

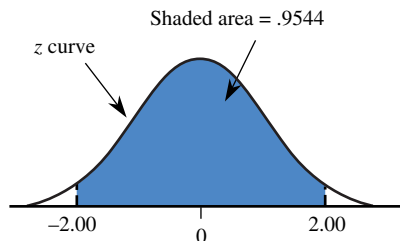
as shown in the following figure:



The probability that z is between -2 and $+2$ (within 2 standard deviations of the mean, since $\mu = 0$ and $\sigma = 1$) is

$$\begin{aligned} P(-2.00 < z < 2.00) &= P(z < 2.00) - P(z < -2.00) \\ &= .9772 - .0228 \\ &= .9544 \\ &\approx .95 \end{aligned}$$

as shown in the following figure:

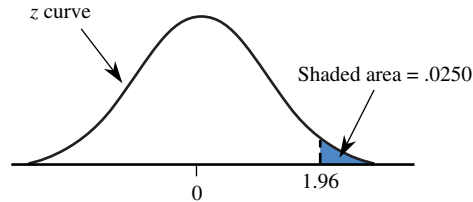


This last probability is the basis for one part of the Empirical Rule, which states that when a histogram is well approximated by a normal curve, roughly 95% of the values are within 2 standard deviations of the mean.

The probability that the value of z exceeds 1.96 is

$$\begin{aligned} P(z > 1.96) &= 1 - P(z < 1.96) \\ &= 1 - .9750 \\ &= .0250 \end{aligned}$$

as shown in the following figure:



That is, 2.5% of the area under the z curve lies to the right of 1.96 in the upper tail.

Similarly,

$$\begin{aligned} P(z > -1.28) &= \text{area to the right of } -1.28 \\ &= 1 - P(z < -1.28) \\ &= 1 - .1003 \\ &= .8997 \\ &\approx .90 \end{aligned}$$

■ Identifying Extreme Values

Suppose that we want to describe the values included in the smallest 2% of a distribution or the values making up the most extreme 5% (which includes the largest 2.5% and the smallest 2.5%). Let's see how we can identify extreme values in the distribution by working through Examples 7.23 and 7.24.

Example 7.23 Identifying Extreme Values

Suppose that we want to describe the values that make up the smallest 2% of the standard normal distribution. Symbolically, we are trying to find a value (call it z^*) such that

$$P(z < z^*) = .02$$

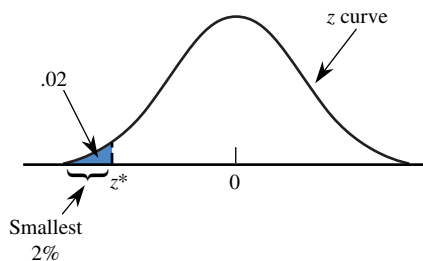


Figure 7.25 The smallest 2% of the standard normal distribution.

This is illustrated in Figure 7.25, which shows that the cumulative area for z^* is .02. Therefore we look for a cumulative area of .0200 in the body of Appendix Table 2. The closest cumulative area in the table is .0202, in the -2.0 row and .05 column; we will use $z^* = -2.05$, the best approximation from the table. Variable values less than -2.05 make up the smallest 2% of the standard normal distribution.

Now suppose that we had been interested in the largest 5% of all z values. We would then be trying to find a value of z^* for which

$$P(z > z^*) = .05$$

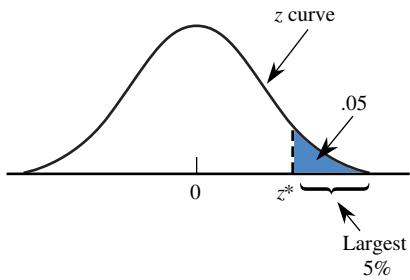


Figure 7.26 The largest 5% of the standard normal distribution.

as illustrated in Figure 7.26. Because Appendix Table 2 always works with cumulative area (area to the left), the first step is to determine

$$\text{area to the left of } z^* = 1 - .05 = .95$$

Looking for the cumulative area closest to .95 in Appendix Table 2, we find that .95 falls exactly halfway between .9495 (corresponding to a z value of 1.64) and .9505 (corresponding to a z value of 1.65). Because .9500 is exactly halfway between the two areas, we use a z value that is halfway between 1.64 and 1.65. (If one value had been closer to .9500 than the other, we would just use the z value corresponding to the closest area). This gives

$$z^* = \frac{1.64 + 1.65}{2} = 1.645$$

Values greater than 1.645 make up the largest 5% of the standard normal distribution. By symmetry, -1.645 separates the smallest 5% of all z values from the others.

Example 7.24 More Extremes

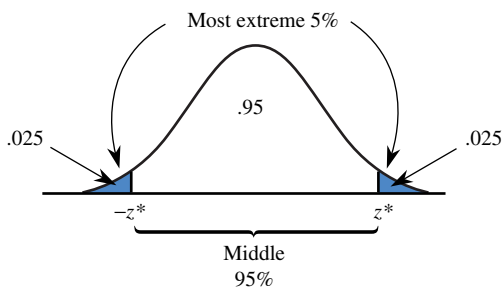


Figure 7.27 The most extreme 5% of the standard normal distribution.

Sometimes we are interested in identifying the most extreme (unusually large *or* small) values in a distribution. Consider describing the values that make up the most extreme 5% of the standard normal distribution. That is, we want to separate the middle 95% from the extreme 5%. This is illustrated in Figure 7.27.

Because the standard normal distribution is symmetric, the most extreme 5% is equally divided between the high side and the low side of the distribution, resulting in an area of .025 for each of the tails of the z curve. Symmetry about 0 implies that if z^* denotes the value that separates the largest 2.5%, the value that separates the smallest 2.5% is simply $-z^*$.

To find z^* , first determine the cumulative area for z^* , which is

$$\text{area to the left of } z^* = .95 + .025 = .975$$

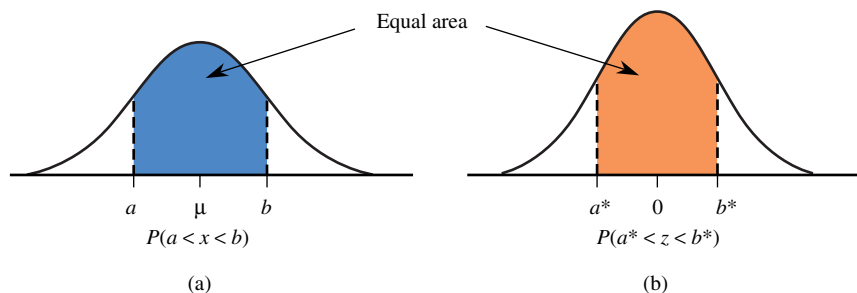
The cumulative area .9750 appears in the 1.9 row and .06 column of Appendix Table 2, so $z^* = 1.96$. For the standard normal distribution, 95% of the variable values fall between -1.96 and 1.96 ; the most extreme 5% are those values that are either greater than 1.96 or less than -1.96 .

Other Normal Distributions

We now show how z curve areas can be used to calculate probabilities and to describe values for any normal distribution. Remember that the letter z is reserved for those variables that have a standard normal distribution; the letter x is used more generally for any variable whose distribution is described by a normal curve with mean μ and standard deviation σ .

Suppose that we want to compute $P(a < x < b)$, the probability that the variable x lies in a particular range. This probability corresponds to an area under a normal curve and above the interval from a to b , as shown in Figure 7.28(a).

Figure 7.28 Equality of nonstandard and standard normal curve areas.



Our strategy for obtaining this probability is to find an equivalent problem involving the standard normal distribution. Finding an equivalent problem means determining an interval (a^*, b^*) that has the same probability for z (same area under the z curve) as does the interval (a, b) in our original normal distribution (Figure 7.28(b)). The asterisk is used to distinguish a and b , the values from the original normal distribution with mean μ and standard deviation σ , from a^* and b^* , the values from the z curve. To find a^* and b^* , we simply calculate z scores for the endpoints of the interval for which a probability is desired. This process is called **standardizing** the endpoints. For example, suppose that the variable x has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 5$. To calculate

$$P(98 < x < 107)$$

we first translate this problem into an equivalent problem for the standard normal distribution. Recall from Chapter 4 that a z score, or standardized score, tells how many standard deviations away from the mean a value lies; the z score is calculated by first subtracting the mean and then dividing by the standard deviation. Converting the lower endpoint $a = 98$ to a z score gives

$$a^* = \frac{98 - 100}{5} = \frac{-2}{5} = -.40$$

and converting the upper endpoint yields

$$b^* = \frac{107 - 100}{5} = \frac{7}{5} = 1.40$$

Then

$$P(98 < x < 107) = P(-.40 < z < 1.40)$$

The probability $P(-.40 < z < 1.40)$ can now be evaluated using Appendix Table 2.

Finding Probabilities

To calculate probabilities for any normal distribution, standardize the relevant values and then use the table of z curve areas. More specifically, if x is a variable whose behavior is described by a normal distribution with mean μ and standard deviation σ , then

$$\begin{aligned} P(x < b) &= P(z < b^*) \\ P(a < x) &= P(a^* < z) \quad [\text{Equivalently, } P(x > a) = P(z > a^*)] \\ P(a < x < b) &= P(a^* < z < b^*) \end{aligned}$$

(continued)

where z is a variable whose distribution is standard normal and

$$a^* = \frac{a - \mu}{\sigma} \quad b^* = \frac{b - \mu}{\sigma}$$

Example 7.25 Children's Heights



© Nicola Sutton/Life File/Getty Images

Data from the article “The Osteological Paradox: Problems in Inferring Prehistoric Health from Skeletal Samples” (*Current Anthropology* [1992]: 343–370) suggest that a reasonable model for the probability distribution of the continuous numerical variable $x =$ height of a randomly selected 5-year-old child is a normal distribution with a mean of $\mu = 100$ cm and standard deviation $\sigma = 6$ cm. What proportion of the heights is between 94 and 112 cm?

To answer this question, we must find

$$P(94 < x < 112)$$

First, we translate the interval endpoints to equivalent endpoints for the standard normal distribution:

$$a^* = \frac{a - \mu}{\sigma} = \frac{94 - 100}{6} = -1.00$$

$$b^* = \frac{b - \mu}{\sigma} = \frac{112 - 100}{6} = 2.00$$

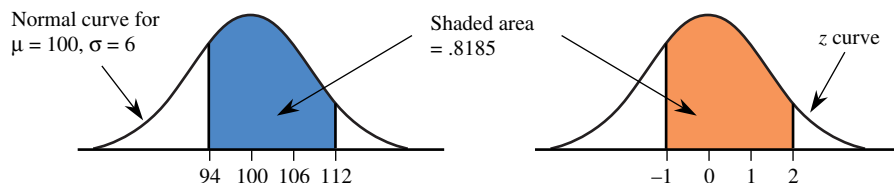
Then

$$\begin{aligned} P(94 < x < 112) &= P(-1.00 < z < 2.00) \\ &= (z \text{ curve area to the left of } 2.00) \\ &\quad - (z \text{ curve area to the left of } -1.00) \\ &= .9772 - .1587 \\ &= .8185 \end{aligned}$$

The probabilities for x and z are shown in Figure 7.29. If height were observed for many children from this population, about 82% of them would fall between 94 and 112 cm.

Figure 7.29

$P(94 < x < 112)$ and corresponding z curve area for the height problem of Example 7.25.



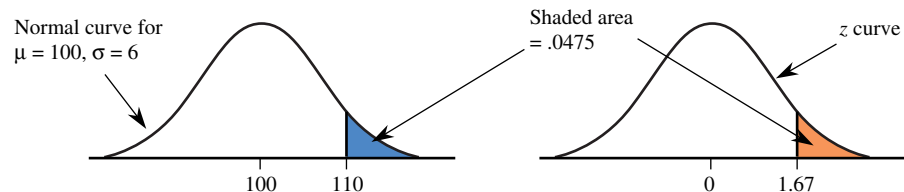
What is the probability that a randomly chosen child will be taller than 110 cm? To evaluate $P(x > 110)$, we first compute

$$a^* = \frac{a - \mu}{\sigma} = \frac{110 - 100}{6} = 1.67$$

Then (see Figure 7.30)

$$\begin{aligned}
 P(x > 110) &= P(z > 1.67) \\
 &= z \text{ curve area to the right of } 1.67 \\
 &= 1 - (z \text{ curve area to the left of } 1.67) \\
 &= 1 - .9525 \\
 &= .0475
 \end{aligned}$$

Figure 7.30 $P(x > 110)$ and corresponding z curve area for the height problem of Example 7.25.



Example 7.26 IQ Scores

Although there is some controversy regarding the appropriateness of IQ scores as a measure of intelligence, IQ scores are commonly used for a variety of purposes. One commonly used IQ scale has a mean of 100 and a standard deviation of 15, and IQ scores are approximately normally distributed. (IQ score is actually a discrete variable [because it is based on the number of correct responses on a test], but its population distribution closely resembles a normal curve.) If we define the random variable

x = IQ score of a randomly selected individual

then x has approximately a normal distribution with $\mu = 100$ and $\sigma = 15$.

One way to become eligible for membership in Mensa, an organization purportedly for those of high intelligence, is to have a Stanford–Binet IQ score above 130. What proportion of the population would qualify for Mensa membership? An answer to this question requires evaluating $P(x > 130)$. This probability is shown in Figure 7.31. With $a = 130$,

$$a^* = \frac{a - \mu}{\sigma} = \frac{130 - 100}{15} = 2.00$$

So (see Figure 7.32)

$$\begin{aligned}
 P(x > 130) &= P(z > 2.00) \\
 &= z \text{ curve area to the right of } 2.00 \\
 &= 1 - (z \text{ curve area to the left of } 2.00) \\
 &= 1 - .9772 \\
 &= .0228
 \end{aligned}$$

Only 2.28% of the population would qualify for Mensa membership.

Suppose that we are interested in the proportion of the population with IQ scores below 80—that is, $P(x < 80)$. With $b = 80$,

$$b^* = \frac{b - \mu}{\sigma} = \frac{80 - 100}{15} = -1.33$$

Figure 7.31 Normal distribution and desired proportion for Example 7.26.

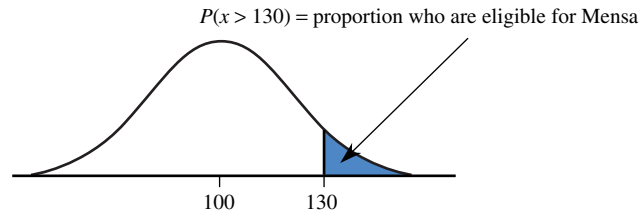
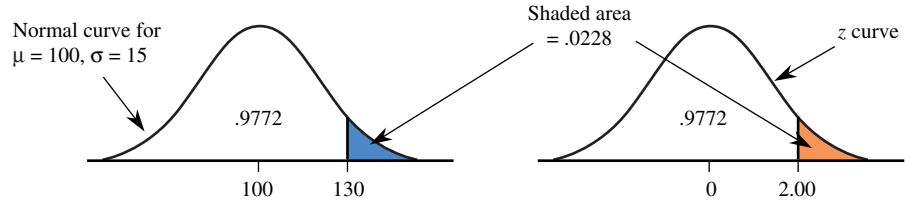


Figure 7.32 $P(x > 130)$ and corresponding z curve area for the IQ problem of Example 7.26.

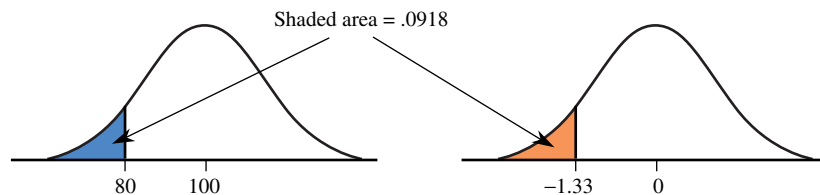


So

$$\begin{aligned} P(x < 80) &= P(z < -1.33) \\ &= z \text{ curve area to the left of } -1.33 \\ &= .0918 \end{aligned}$$

as shown in Figure 7.33. This probability (.0918) tells us that just a little over 9% of the population has an IQ score below 80.

Figure 7.33 $P(x < 80)$ and corresponding z curve area for the IQ problem of Example 7.26.



Now consider the proportion of the population with IQs between 75 and 125. Using $a = 75$ and $b = 125$, we obtain

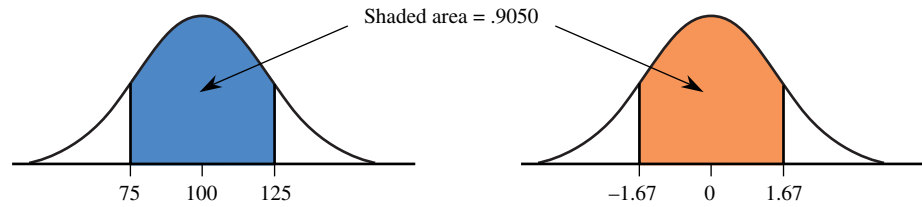
$$a^* = \frac{75 - 100}{15} = -1.67 \quad b^* = \frac{125 - 100}{15} = 1.67$$

so

$$\begin{aligned} P(75 < x < 125) &= P(-1.67 < z < 1.67) \\ &= z \text{ curve area between } -1.67 \text{ and } 1.67 \\ &= (z \text{ curve area to the left of } 1.67) \\ &\quad - (z \text{ curve area to the left of } -1.67) \\ &= .9525 - .0475 \\ &= .9050 \end{aligned}$$

This is illustrated in Figure 7.34. The calculation tells us that 90.5% of the population has an IQ score between 75 and 125. Of the 9.5% whose IQ score is not between 75 and 125, half of them (4.75%) have scores over 125, and the other half have scores below 75.

Figure 7.34
 $P(75 < x < 125)$ and corresponding z curve area for the IQ problem of Example 7.26.



When we translate from a problem involving a normal distribution with mean μ and standard deviation σ to a problem involving the standard normal distribution, we convert to z scores:

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

Because a z score can be interpreted as giving the distance of an x value from the mean in units of the standard deviation, a z score of 1.4 corresponds to an x value that is 1.4 standard deviations above the mean, and a z score of -2.1 corresponds to an x value that is 2.1 standard deviations below the mean.

Suppose that we are trying to evaluate $P(x < 60)$ for a variable whose distribution is normal with $\mu = 50$ and $\sigma = 5$. Converting the endpoint 60 to a z score gives

$$z = \frac{60 - 50}{5} = 2$$

which tells us that the value 60 is 2 standard deviations above the mean. We then have

$$P(x < 60) = P(z < 2)$$

where z is a standard normal variable. Notice that for the standard normal distribution, the value 2 is 2 standard deviations above the mean, because the mean is 0 and the standard deviation is 1. The value $z = 2$ is located the same distance (measured in standard deviations) from the mean of the standard normal distribution as is the value $x = 60$ from the mean in the normal distribution with $\mu = 50$ and $\sigma = 5$. This is why the translation using z scores results in an equivalent problem involving the standard normal distribution.

■ Describing Extreme Values in a Normal Distribution

To describe the extreme values for a normal distribution with mean μ and standard deviation σ , we first solve the corresponding problem for the standard normal distribution and then translate our answer into one for the normal distribution of interest. This process is illustrated in Example 7.27.

Example 7.27 Registration Times

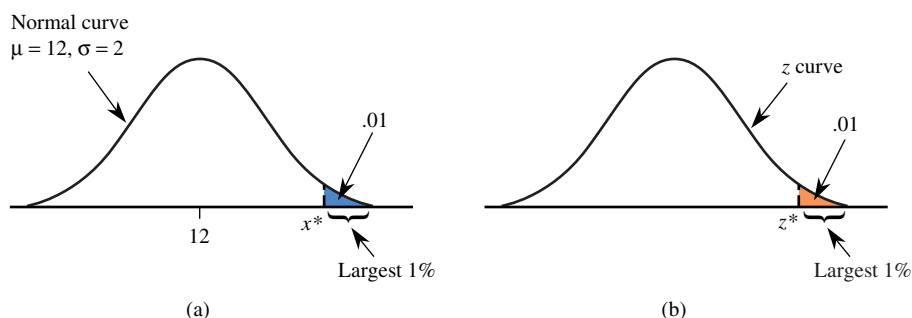
Data on the length of time required to complete registration for classes using a telephone registration system suggest that the distribution of the variable

$$x = \text{time to register}$$

for students at a particular university can be well approximated by a normal distribution with mean $\mu = 12$ min and standard deviation $\sigma = 2$ min. (The normal

distribution might not be an appropriate model for $x =$ time to register at another university. Many factors influence the shape, center, and spread of such a distribution.) Because some students do not sign off properly, the university would like to disconnect students automatically after some amount of time has elapsed. It is decided to choose this time such that only 1% of the students are disconnected while they are still attempting to register. To determine the amount of time that should be allowed before disconnecting a student, we need to describe the largest 1% of the distribution of time to register. These are the individuals who will be mistakenly disconnected. This is illustrated in Figure 7.35(a). To determine the value of x^* , we first solve the analogous problem for the standard normal distribution, as shown in Figure 7.35(b).

Figure 7.35 Capturing the largest 1% in a normal distribution for the problem in Example 7.27.



By looking at Appendix Table 2 for a cumulative area of .99, we find the closest entry (.9901) in the 2.3 row and the .03 column, from which $z^* = 2.33$. For the standard normal distribution, the largest 1% of the distribution is made up of those values greater than 2.33. An equivalent statement is that the largest 1% are those with z scores greater than 2.33. This implies that in the distribution of time to register x (or any other normal distribution), the largest 1% are those values with z scores greater than 2.33 or, equivalently, those x values more than 2.33 standard deviations above the mean. Here, the standard deviation is 2, so 2.33 standard deviations is $2.33(2)$, and it follows that

$$x^* = 12 + 2.33(2) = 12 + 4.66 = 16.66$$

The largest 1% of the distribution for time to register is made up of values that are greater than 16.66 min. If the university system was set to disconnect students after 16.66 min, only 1% of the students registering would be disconnected before completing their registration.

A general formula for converting a z score back to an x value results from solving $z^* = \frac{x^* - \mu}{\sigma}$ for x^* , as shown in the accompanying box.

To convert a z score z^* back to an x value, use

$$x = \mu + z^*\sigma$$

Example 7.28 Motor Vehicle Emissions



© Hisham F. Ibrahim/Getty Images

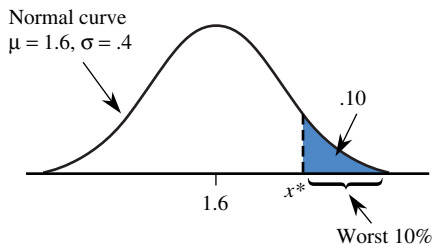
Data from the article “Determining Statistical Characteristics of a Vehicle Emissions Audit Procedure” (*Technometrics* [1980]: 483–493) suggest that the emissions of nitrogen oxides, which are major constituents of smog, can be plausibly modeled using a normal distribution. Let x denote the amount of this pollutant emitted by a randomly selected vehicle. The distribution of x can be described by a normal distribution with $\mu = 1.6$ and $\sigma = 0.4$.

Suppose that the EPA wants to offer some sort of incentive to get the worst polluters off the road. What emission levels constitute the worst 10% of the vehicles? The worst 10% would be the 10% with the highest emissions level, as shown in the illustration in the margin.

For the standard normal distribution, the largest 10% are those with z values greater than $z^* = 1.28$ (from Appendix Table 2, based on a cumulative area of .90).

Then

$$\begin{aligned} x^* &= \mu + z^*\sigma \\ &= 1.6 + 1.28(.4) \\ &= 1.6 + .512 \\ &= 2.112 \end{aligned}$$



In the population of vehicles of the type considered, about 10% would have oxide emission levels greater than 2.112.

Exercises 7.64–7.80

7.64 Determine the following standard normal (z) curve areas:

- The area under the z curve to the left of 1.75
- The area under the z curve to the left of -0.68
- The area under the z curve to the right of 1.20
- The area under the z curve to the right of -2.82
- The area under the z curve between -2.22 and 0.53
- The area under the z curve between -1 and 1
- The area under the z curve between -4 and 4

7.65 Determine each of the following areas under the standard normal (z) curve:

- To the left of -1.28
- To the right of 1.28

- Between -1 and 2
- To the right of 0
- To the right of -5
- Between -1.6 and 2.5
- To the left of 0.23

7.66 Let z denote a random variable that has a standard normal distribution. Determine each of the following probabilities:

- $P(z < 2.36)$
- $P(z \leq 2.36)$
- $P(z < -1.23)$
- $P(1.14 < z < 3.35)$
- $P(-0.77 \leq z \leq -0.55)$

- f. $P(z > 2)$
 g. $P(z \geq -3.38)$
 h. $P(z < 4.98)$

7.67 Let z denote a random variable having a normal distribution with $\mu = 0$ and $\sigma = 1$. Determine each of the following probabilities:

- a. $P(z < 0.10)$
 b. $P(z < -0.10)$
 c. $P(0.40 < z < 0.85)$
 d. $P(-0.85 < z < -0.40)$
 e. $P(-0.40 < z < 0.85)$
 f. $P(z > -1.25)$
 g. $P(z < -1.50 \text{ or } z > 2.50)$

7.68 Let z denote a variable that has a standard normal distribution. Determine the value z^* to satisfy the following conditions:

- a. $P(z < z^*) = .025$
 b. $P(z < z^*) = .01$
 c. $P(z < z^*) = .05$
 d. $P(z > z^*) = .02$
 e. $P(z > z^*) = .01$
 f. $P(z > z^* \text{ or } z < -z^*) = .20$

7.69 Determine the value z^* that

- a. Separates the largest 3% of all z values from the others
 b. Separates the largest 1% of all z values from the others
 c. Separates the smallest 4% of all z values from the others
 d. Separates the smallest 10% of all z values from the others

7.70 Determine the value of z^* such that

- a. $-z^*$ and z^* separate the middle 95% of all z values from the most extreme 5%
 b. $-z^*$ and z^* separate the middle 90% of all z values from the most extreme 10%
 c. $-z^*$ and z^* separate the middle 98% of all z values from the most extreme 2%
 d. $-z^*$ and z^* separate the middle 92% of all z values from the most extreme 8%

7.71 Because $P(z < .44) = .67$, 67% of all z values are less than .44, and .44 is the 67th percentile of the standard normal distribution. Determine the value of each of the following percentiles for the standard normal distribution (Hint: If the cumulative area that you must look for does not appear in the z table, use the closest entry):

- a. The 91st percentile (Hint: Look for area .9100.)

- b. The 77th percentile
 c. The 50th percentile
 d. The 9th percentile
 e. What is the relationship between the 70th z percentile and the 30th z percentile?

7.72 Consider the population of all 1-gal cans of dusty rose paint manufactured by a particular paint company. Suppose that a normal distribution with mean $\mu = 5$ ml and standard deviation $\sigma = 0.2$ ml is a reasonable model for the distribution of the variable $x =$ amount of red dye in the paint mixture. Use the normal distribution model to calculate the following probabilities:

- a. $P(x < 5.0)$
 b. $P(x < 5.4)$
 c. $P(x \leq 5.4)$
 d. $P(4.6 < x < 5.2)$
 e. $P(x > 4.5)$
 f. $P(x > 4.0)$

7.73 Consider babies born in the “normal” range of 37–43 weeks gestational age. Extensive data support the assumption that for such babies born in the United States, birth weight is normally distributed with mean 3432 g and standard deviation 482 g (“Are Babies Normal?” *The American Statistician* [1999]: 298–302).

- a. What is the probability that the birth weight of a randomly selected baby of this type exceeds 4000 g? is between 3000 and 4000 g?
 b. What is the probability that the birth weight of a randomly selected baby of this type is either less than 2000 g or greater than 5000 g?
 c. What is the probability that the birth weight of a randomly selected baby of this type exceeds 7 lb? (Hint: 1 lb = 453.59 g.)
 d. How would you characterize the most extreme 0.1% of all birth weights?
 e. If x is a random variable with a normal distribution and a is a numerical constant ($a \neq 0$), then $y = ax$ also has a normal distribution. Use this formula to determine the distribution of birth weight expressed in pounds (shape, mean, and standard deviation), and then recalculate the probability from Part (c). How does this compare to your previous answer?

7.74 A machine that cuts corks for wine bottles operates in such a way that the distribution of the diameter of the corks produced is well approximated by a normal distribution with mean 3 cm and standard deviation 0.1 cm. The specifications call for corks with diameters between 2.9 and 3.1 cm. A cork not meeting the specifications is considered defective. (A cork that is too small leaks and causes the wine to deteriorate; a cork that is too large

doesn't fit in the bottle.) What proportion of corks produced by this machine are defective?

7.75 Refer to Exercise 7.74. Suppose that there are two machines available for cutting corks. The machine described in the preceding problem produces corks with diameters that are approximately normally distributed with mean 3 cm and standard deviation 0.1 cm. The second machine produces corks with diameters that are approximately normally distributed with mean 3.05 cm and standard deviation 0.01 cm. Which machine would you recommend? (Hint: Which machine would produce fewer defective corks?)

7.76 A gasoline tank for a certain car is designed to hold 15 gal of gas. Suppose that the variable x = actual capacity of a randomly selected tank has a distribution that is well approximated by a normal curve with mean 15.0 gal and standard deviation 0.1 gal.

- What is the probability that a randomly selected tank will hold at most 14.8 gal?
- What is the probability that a randomly selected tank will hold between 14.7 and 15.1 gal?
- If two such tanks are independently selected, what is the probability that both hold at most 15 gal?

7.77 ▼ The time that it takes a randomly selected job applicant to perform a certain task has a distribution that can be approximated by a normal distribution with a mean value of 120 sec and a standard deviation of 20 sec. The fastest 10% are to be given advanced training. What task times qualify individuals for such training?

7.78 A machine that produces ball bearings has initially been set so that the true average diameter of the bearings it produces is 0.500 in. A bearing is acceptable if its diameter is within 0.004 in. of this target value. Suppose, however, that the setting has changed during the course of

production, so that the distribution of the diameters produced is well approximated by a normal distribution with mean 0.499 in. and standard deviation 0.002 in. What percentage of the bearings produced will not be acceptable?

7.79 ▼ Suppose that the distribution of net typing rate in words per minute (wpm) for experienced typists can be approximated by a normal curve with mean 60 wpm and standard deviation 15 wpm (“Effects of Age and Skill in Typing”, *Journal of Experimental Psychology* [1984]: 345–371).

- What is the probability that a randomly selected typist's net rate is at most 60 wpm? less than 60 wpm?
- What is the probability that a randomly selected typist's net rate is between 45 and 90 wpm?
- Would you be surprised to find a typist in this population whose net rate exceeded 105 wpm? (Note: The largest net rate in a sample described in the paper is 104 wpm.)
- Suppose that two typists are independently selected. What is the probability that both their typing rates exceed 75 wpm?
- Suppose that special training is to be made available to the slowest 20% of the typists. What typing speeds would qualify individuals for this training?

7.80 Consider the variable x = time required for a college student to complete a standardized exam. Suppose that for the population of students at a particular university, the distribution of x is well approximated by a normal curve with mean 45 min and standard deviation 5 min.

- If 50 min is allowed for the exam, what proportion of students at this university would be unable to finish in the allotted time?
- How much time should be allowed for the exam if we wanted 90% of the students taking the test to be able to finish in the allotted time?
- How much time is required for the fastest 25% of all students to complete the exam?

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

7.7

Checking for Normality and Normalizing Transformations

Some of the most frequently used statistical methods are valid only when a sample x_1, x_2, \dots, x_n has come from a population distribution that is at least approximately normal. One way to see whether an assumption of population normality is plausible is to construct a **normal probability plot** of the data. One version of this plot uses

quantities called **normal scores**. The values of the normal scores depend on the sample size n . For example, the normal scores when $n = 10$ are as follows:

| | | | | |
|--------|--------|-------|-------|-------|
| −1.539 | −1.001 | −.656 | −.376 | −.123 |
| .123 | .376 | .656 | 1.001 | 1.539 |

To interpret these numbers, think of selecting sample after sample from a standard normal distribution, each one consisting of $n = 10$ observations. Then -1.539 is the long-run average of the smallest observation from each sample, -1.001 is the long-run average of the second smallest observation from each sample, and so on. In other words, -1.539 is the mean value of the smallest observation in a sample of size 10 from the z distribution, -1.001 is the mean value of the second smallest observation, and so on.

Extensive tabulations of normal scores for many different sample sizes are available. Alternatively, many software packages (such as MINITAB and SAS) and some graphing calculators can compute these scores on request and then construct a normal probability plot. Not all calculators and software packages use the same algorithm to compute normal scores. However, this does not change the overall character of a normal probability plot, so either the tabulated values or those given by the computer or calculator can be used.

After the sample observations are ordered from smallest to largest, the smallest normal score is paired with the smallest observation, the second smallest normal score with the second smallest observation, and so on. The first number in a pair is the normal score, and the second number in the pair is the observed data value. A normal probability plot is just a scatterplot of the (normal score, observed value) pairs.

If the sample has been selected from a *standard* normal distribution, the second number in each pair should be reasonably close to the first number (ordered observation \approx corresponding mean value). Then the n plotted points will fall near a line with slope equal to 1 (a 45° line) passing through $(0, 0)$. When the sample has been obtained from *some* normal population distribution (but not necessarily the standard normal distribution), the plotted points should be close to *some* straight line.

DEFINITION

A **normal probability plot** is a scatter plot of the (normal score, observed value) pairs. A substantial linear pattern in a normal probability plot suggests that population normality is plausible. On the other hand, a systematic departure from a straight-line pattern (such as curvature in the plot) casts doubt on the legitimacy of assuming a normal population distribution.

Example 7.29 Window Widths

The following 10 observations are widths of contact windows in integrated circuit chips:

3.21 2.49 2.94 4.38 4.02 3.62 3.30 2.85 3.34 3.81

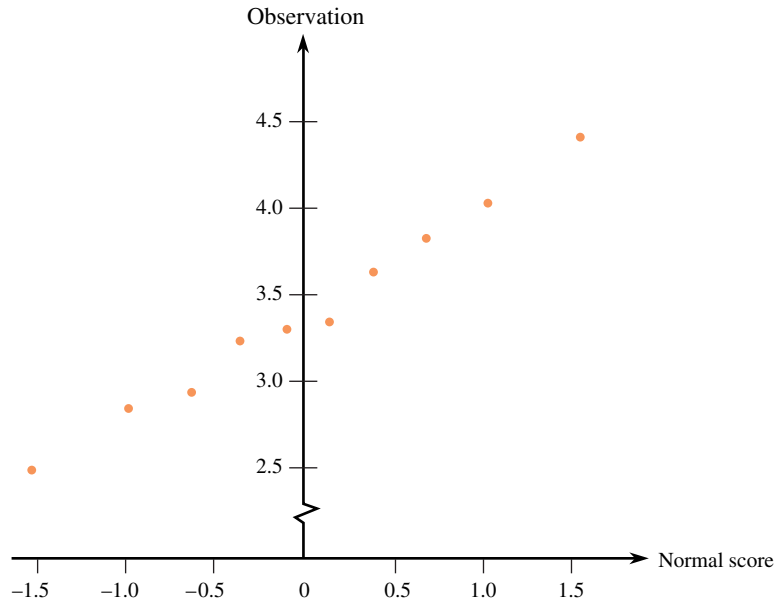
The 10 pairs for the normal probability plot are then

| | |
|------------------|-----------------|
| $(-1.539, 2.49)$ | $(0.123, 3.34)$ |
| $(-1.001, 2.85)$ | $(0.376, 3.62)$ |
| $(-0.656, 2.94)$ | $(0.656, 3.81)$ |

(-0.376, 3.21) (1.001, 4.02)
 (-0.123, 3.30) (1.539, 4.38)

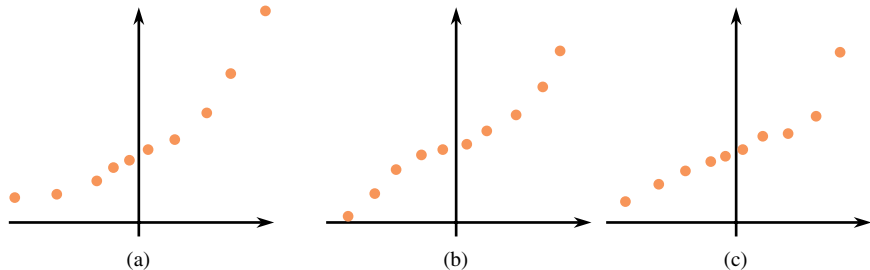
The normal probability plot is shown in Figure 7.36. The linearity of the plot supports the assumption that the window width distribution from which these observations were drawn is normal.

Figure 7.36 A normal probability plot for Example 7.29.



The decision as to whether a plot shows a substantial linear pattern is somewhat subjective. Particularly when n is small, normality should not be ruled out unless the departure from linearity is clear-cut. Figure 7.37 displays several plots that suggest a nonnormal population distribution.

Figure 7.37 Plots suggesting nonnormality: (a) indication that the population distribution is skewed; (b) indication that the population distribution has heavier tails than a normal curve; (c) presence of an outlier.



■ **Using the Correlation Coefficient to Check Normality**

The correlation coefficient r was introduced in Chapter 5 as a quantitative measure of the extent to which the points in a scatterplot fall close to a straight line. Consider the n (normal score, observed value) pairs:

- (smallest normal score, smallest observation)
- ⋮
- (largest normal score, largest observation)

Then the correlation coefficient can be computed as discussed in Chapter 5. The normal probability plot always slopes upward (because it is based on values ordered from

smallest to largest), so r will be a positive number. A value of r quite close to 1 indicates a very strong linear relationship in the normal probability plot. If r is too much smaller than 1, normality of the underlying distribution is questionable.

How far below 1 does r have to be before we begin to seriously doubt the plausibility of normality? The answer depends on the sample size n . If n is small, an r value somewhat below 1 is not surprising, even when the distribution is normal, but if n is large, only an r value very close to 1 supports the assumption of normality. For selected values of n , Table 7.2 gives critical values to which r can be compared to check for normality. If your sample size is in between two tabulated values of n , use the critical value for the larger sample size. (For example, if $n = 46$, use the value .966 for sample size 50.)

Table 7.2 Values to Which r Can Be Compared to Check for Normality*

| n | 5 | 10 | 15 | 20 | 25 | 30 | 40 | 50 | 60 | 75 |
|--------------------------------|------|------|------|------|------|------|------|------|------|------|
| Critical r | .832 | .880 | .911 | .929 | .941 | .949 | .960 | .966 | .971 | .976 |

*Source: *MINITAB User's Manual*.

If

$r < \text{critical } r \text{ for corresponding } n$

considerable doubt is cast on the assumption of population normality.

How were the critical values in Table 7.2 obtained? Consider the critical value .941 for $n = 25$. Suppose that the underlying distribution is actually normal. Consider obtaining a large number of different samples, each one consisting of 25 observations, and computing the value of r for each one. Then it can be shown that only 1% of the samples result in an r value less than the critical value .941. That is, .941 was chosen to guarantee a 1% error rate: In only 1% of all cases will we judge normality implausible when the distribution is really normal. The other critical values are also chosen to yield a 1% error rate for the corresponding sample sizes.

It might have occurred to you that another type of error is possible: obtaining a large value of r and concluding that normality is a reasonable assumption when the distribution is actually nonnormal. This type of error is more difficult to control than the type mentioned previously, but the procedure we have described generally does a good job in both respects.

Example 7.30 Window Widths Continued

The sample size for the contact window width data of Example 7.29 is $n = 10$. The critical r , from Table 7.2 is then .880. The correlation coefficient calculated using the (normal score, observed value) pairs is $r = .995$. Because r is larger than the critical r for a sample of size 10, it is plausible that the population distribution of window widths from which this sample was drawn is approximately normal.

■ Transforming Data to Obtain a Distribution That Is Approximately Normal

Many of the most frequently used statistical methods are valid only when the sample is selected at random from a population whose distribution is at least approximately normal. When a sample histogram shows a distinctly nonnormal shape, it is common to use a transformation or reexpression of the data. By *transforming* data, we mean applying some specified mathematical function (such as the square root, logarithm, or reciprocal) to each data value to produce a set of transformed data. We can then study and summarize the distribution of these transformed values using methods that require normality. We saw in Chapter 5 that, with bivariate data, one or both of the variables can be transformed in an attempt to find two variables that are linearly related. With univariate data, a transformation is usually chosen to yield a distribution of transformed values that is more symmetric and more closely approximated by a normal curve than was the original distribution.

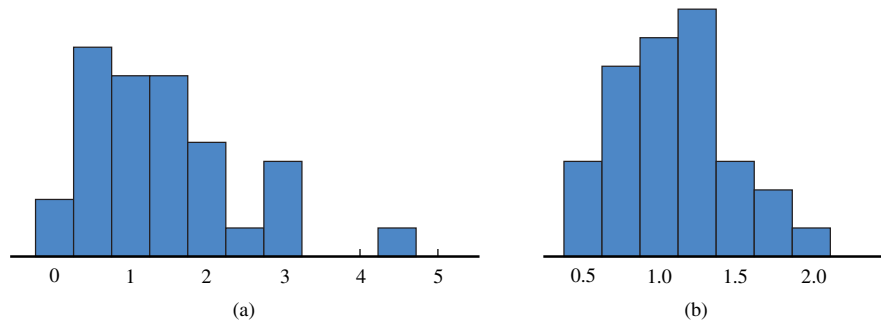
Example 7.31 Rainfall Data

● Data that have been used by several investigators to introduce the concept of transformation (e.g., “Exploratory Methods for Choosing Power Transformations,” *Journal of the American Statistical Association* [1982]: 103–108) consist of values of March precipitation for Minneapolis–St. Paul over a period of 30 years. These values are given in Table 7.3, along with the square root of each value. Histograms of both the original and the transformed data appear in Figure 7.38. The distribution of the original data is clearly skewed, with a long upper tail. The square-root transformation results in a substantially more symmetric distribution, with a typical (i.e., central) value near the 1.25 boundary between the third and fourth class intervals.

Table 7.3 Original and Square-Root-Transformed Values of March Precipitation in Minneapolis–St. Paul over a 30-year Period

| Year | Precipitation | $\sqrt{\text{Precipitation}}$ | Year | Precipitation | $\sqrt{\text{Precipitation}}$ |
|------|---------------|-------------------------------|------|---------------|-------------------------------|
| 1 | .77 | .88 | 16 | 1.62 | 1.27 |
| 2 | 1.74 | 1.32 | 17 | 1.31 | 1.14 |
| 3 | .81 | .90 | 18 | .32 | .57 |
| 4 | 1.20 | 1.10 | 19 | .59 | .77 |
| 5 | 1.95 | 1.40 | 20 | .81 | .90 |
| 6 | 1.20 | 1.10 | 21 | 2.81 | 1.68 |
| 7 | .47 | .69 | 22 | 1.87 | 1.37 |
| 8 | 1.43 | 1.20 | 23 | 1.18 | 1.09 |
| 9 | 3.37 | 1.84 | 24 | 1.35 | 1.16 |
| 10 | 2.20 | 1.48 | 25 | 4.75 | 2.18 |
| 11 | 3.00 | 1.73 | 26 | 2.48 | 1.57 |
| 12 | 3.09 | 1.76 | 27 | .96 | .98 |
| 13 | 1.51 | 1.23 | 28 | 1.89 | 1.37 |
| 14 | 2.10 | 1.45 | 29 | .90 | .95 |
| 15 | .52 | .72 | 30 | 2.05 | 1.43 |

Figure 7.38 Histograms of the precipitation data used in Example 7.31: (a) untransformed data; (b) square-root transformed data.



Logarithmic transformations are also common and, as with bivariate data, either the natural logarithm or the base 10 logarithm can be used. A logarithmic transformation is usually applied to data that are positively skewed (a long upper tail). This affects values in the upper tail substantially more than values in the lower tail, yielding a more symmetric—and often more nearly normal—distribution.

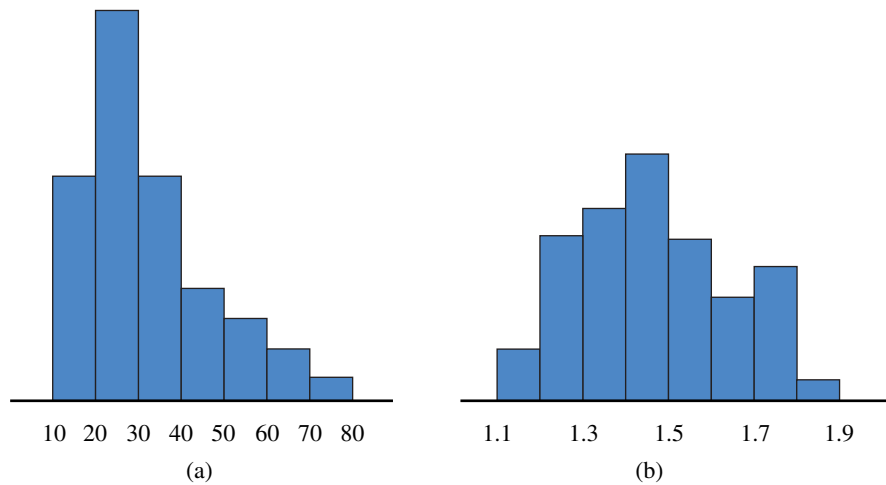
Example 7.32 Beryllium Exposure

- Exposure to beryllium is known to produce adverse effects on lungs as well as on other tissues and organs in both laboratory animals and humans. The article “Time Lapse Cinematographic Analysis of Beryllium: Lung Fibroblast Interactions” (*Environmental Research* [1983]: 34–43) reported the results of experiments designed to study the behavior of certain individual cells that had been exposed to beryllium. An important characteristic of such an individual cell is its interdivision time (IDT). IDTs were determined for a large number of cells under both exposed (treatment) and unexposed (control) conditions. The authors of the article stated, “The IDT distributions are seen to be skewed, but the natural logs do have an approximate normal distribution.” The same property holds for \log_{10} transformed data. We give representative IDT data in Table 7.4 and the resulting histograms in Figure 7.39, which are in agreement with the authors’ statement.

Table 7.4 Original and $\log_{10}(\text{IDT})$ Values

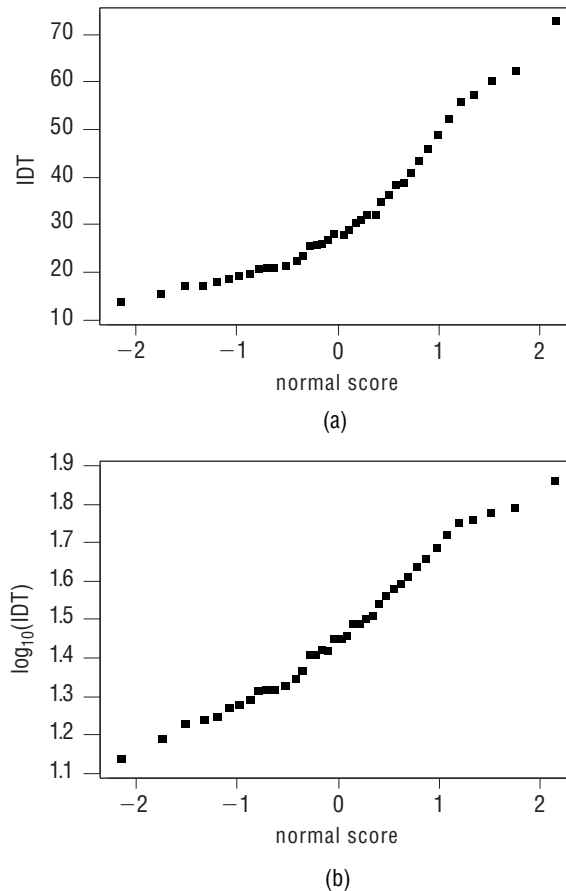
| IDT | $\log_{10}(\text{IDT})$ | IDT | $\log_{10}(\text{IDT})$ | IDT | $\log_{10}(\text{IDT})$ |
|------|-------------------------|------|-------------------------|------|-------------------------|
| 28.1 | 1.45 | 31.2 | 1.49 | 13.7 | 1.14 |
| 46.0 | 1.66 | 25.8 | 1.41 | 16.8 | 1.23 |
| 34.8 | 1.54 | 62.3 | 1.79 | 28.0 | 1.45 |
| 17.9 | 1.25 | 19.5 | 1.29 | 21.1 | 1.32 |
| 31.9 | 1.50 | 28.9 | 1.46 | 60.1 | 1.78 |
| 23.7 | 1.37 | 18.6 | 1.27 | 21.4 | 1.33 |
| 26.6 | 1.42 | 26.2 | 1.42 | 32.0 | 1.51 |
| 43.5 | 1.64 | 17.4 | 1.24 | 38.8 | 1.59 |
| 30.6 | 1.49 | 55.6 | 1.75 | 25.5 | 1.41 |
| 52.1 | 1.72 | 21.0 | 1.32 | 22.3 | 1.35 |
| 15.5 | 1.19 | 36.3 | 1.56 | 19.1 | 1.28 |
| 38.4 | 1.58 | 72.8 | 1.86 | 48.9 | 1.69 |
| 21.4 | 1.33 | 20.7 | 1.32 | 57.3 | 1.76 |
| 40.9 | 1.61 | | | | |

Figure 7.39 Histograms of the IDT data used in Example 7.32: (a) untransformed data; (b) \log_{10} transformed data.



The sample size for the IDT data is $n = 40$. The correlation coefficient for the (normal score, original [untransformed] data) pairs is .950, which is less than the critical r for $n = 40$ (critical $r = .960$). The correlation coefficient using the transformed data is .998, which is much larger than the critical r , supporting the assertion that $\log_{10}(\text{IDT})$ has approximately a normal distribution. Figure 7.40 displays

Figure 7.40 MINITAB-generated normal probability plots for Example 7.32: (a) original IDT data; (b) \log_{10} -transformed IDT.



MINITAB normal probability plots for the original data and for the transformed data. The plot for the transformed data is clearly more linear in appearance than the plot for the original data.

■ Selecting a Transformation

Occasionally, a particular transformation can be dictated by some theoretical argument, but often this is not the case and you may wish to try several different transformations to find one that is satisfactory. Figure 7.41, from the article “Distribution of Sperm Counts in Suspected Infertile Men” (*Journal of Reproduction and Fertility* [1983]: 91–96), shows what can result from such a search. Other investigators in this field had previously used all three of the transformations illustrated.

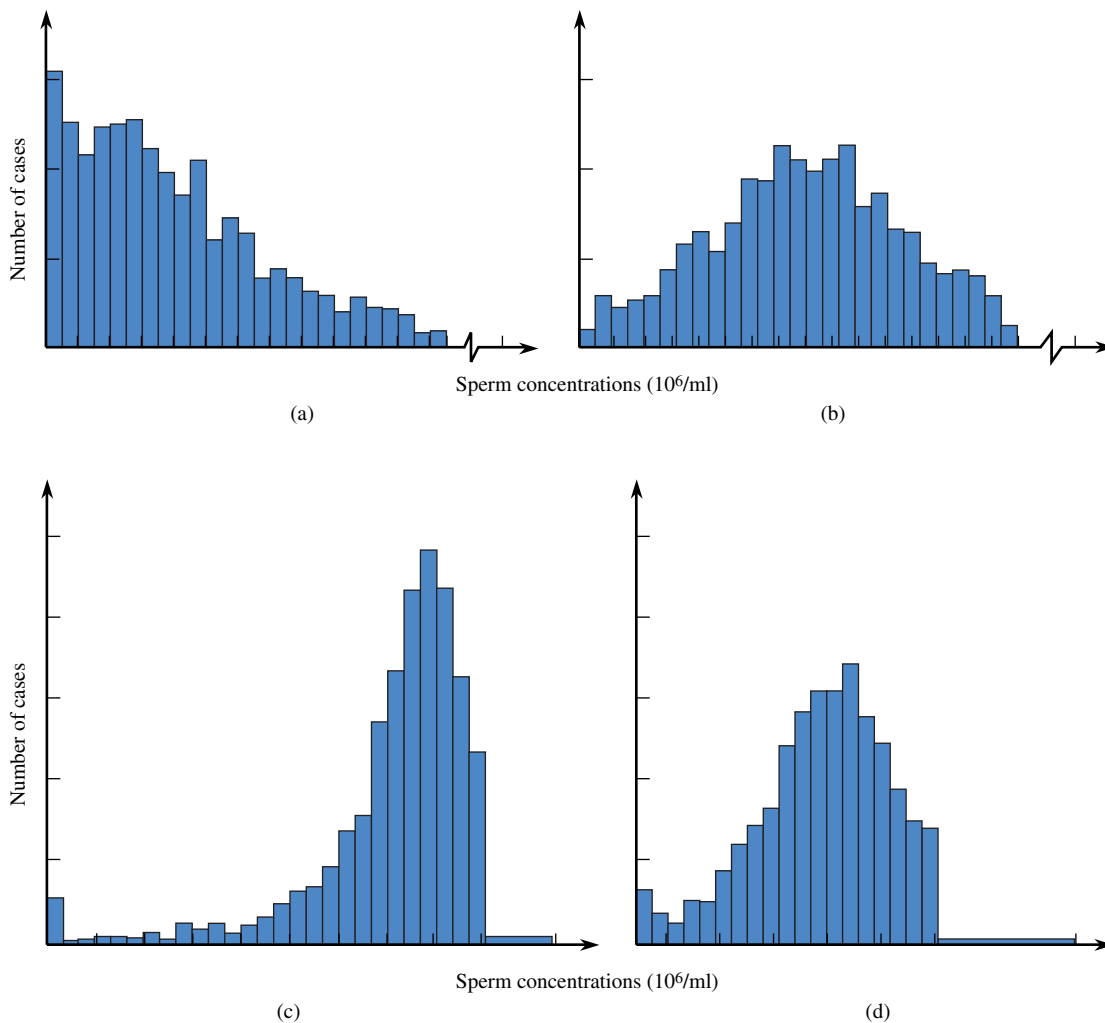
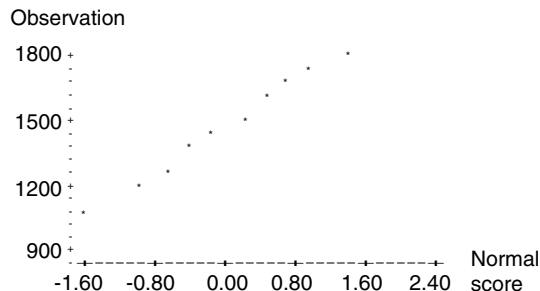


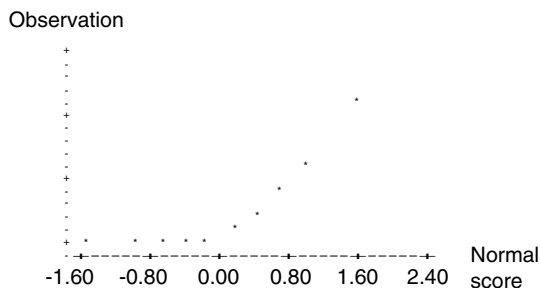
Figure 7.41 Histograms of sperm concentrations for 1711 suspected infertile men: (a) untransformed data (highly skewed); (b) log-transformed data (reasonably symmetric); (c) square-root-transformed data; (d) cube-root-transformed data.

Exercises 7.81–7.92

7.81 Ten measurements of the steam rate (in pounds per hour) of a distillation tower were used to construct the following normal probability plot (“A Self-Descaling Distillation Tower,” *Chemical Engineering Process* [1968]: 79–84). Based on the plot, do you think it is reasonable to assume that the normal distribution provides an adequate description of the steam rate distribution? Explain.



7.82 The following normal probability plot was constructed using part of the data appearing in the paper “Trace Metals in Sea Scallops” (*Environmental Concentration and Toxicology* 19: 1326–1334).



The variable under study was the amount of cadmium in North Atlantic scallops. Do the sample data suggest that the cadmium concentration distribution is not normal? Explain.

7.83 ● Consider the following 10 observations on the lifetime (in hours) for a certain type of component: 152.7, 172.0, 172.5, 173.3, 193.0, 204.7, 216.5, 234.9, 262.6, 422.6. Construct a normal probability plot, and comment on the plausibility of a normal distribution as a model for component lifetime.

7.84 The paper “The Load-Life Relationship for M50 Bearings with Silicon Nitride Ceramic Balls” (*Lubrication*

Engineering [1984]: 153–159) reported the following data on bearing load life (in millions of revolutions); the corresponding normal scores are also given:

| x | Normal Score | x | Normal Score |
|-------|--------------|-------|--------------|
| 47.1 | -1.867 | 240.0 | 0.062 |
| 68.1 | -1.408 | 240.0 | 0.187 |
| 68.1 | -1.131 | 278.0 | 0.315 |
| 90.8 | -0.921 | 278.0 | 0.448 |
| 103.6 | -0.745 | 289.0 | 0.590 |
| 106.0 | -0.590 | 289.0 | 0.745 |
| 115.0 | -0.448 | 367.0 | 0.921 |
| 126.0 | -0.315 | 385.9 | 1.131 |
| 146.6 | -0.187 | 392.0 | 1.408 |
| 229.0 | -0.062 | 395.0 | 1.867 |

Construct a normal probability plot. Is normality plausible?

7.85 ● The following observations are DDT concentrations in the blood of 20 people:

24 26 30 35 35 38 39 40 40 41 42 52
56 58 61 75 79 88 102 42

Use the normal scores from Exercise 7.84 to construct a normal probability plot, and comment on the appropriateness of a normal probability model.

7.86 ● Consider the following sample of 25 observations on the diameter x (in centimeters) of a disk used in a certain system:

16.01 16.08 16.13 15.94 16.05 16.27 15.89
15.84 15.95 16.10 15.92 16.04 15.82 16.15
16.06 15.66 15.78 15.99 16.29 16.15 16.19
16.22 16.07 16.13 16.11

The 13 largest normal scores for a sample of size 25 are 1.965, 1.524, 1.263, 1.067, 0.905, 0.764, 0.637, 0.519, 0.409, 0.303, 0.200, 0.100, and 0. The 12 smallest scores result from placing a negative sign in front of each of the given nonzero scores. Construct a normal probability plot. Does it appear plausible that disk diameter is normally distributed? Explain.

7.87 ● Example 7.31 examined rainfall data for Minneapolis–St. Paul. The square-root transformation was used to obtain a distribution of values that was more symmetric

than the distribution of the original data. Another power transformation that has been suggested by meteorologists is the cube root: transformed value = (original value)^{1/3}. The original values and their cube roots (the transformed values) are given in the following table:

| Original | Transformed | Original | Transformed |
|----------|-------------|----------|-------------|
| 0.32 | 0.68 | 1.51 | 1.15 |
| 0.47 | 0.78 | 1.62 | 1.17 |
| 0.52 | 0.80 | 1.74 | 1.20 |
| 0.59 | 0.84 | 1.87 | 1.23 |
| 0.77 | 0.92 | 1.89 | 1.24 |
| 0.81 | 0.93 | 1.95 | 1.25 |
| 0.81 | 0.93 | 2.05 | 1.27 |
| 0.90 | 0.97 | 2.10 | 1.28 |
| 0.96 | 0.99 | 2.20 | 1.30 |
| 1.18 | 1.06 | 2.48 | 1.35 |
| 1.20 | 1.06 | 2.81 | 1.41 |
| 1.20 | 1.06 | 3.00 | 1.44 |
| 1.31 | 1.09 | 3.09 | 1.46 |
| 1.35 | 1.11 | 3.37 | 1.50 |
| 1.43 | 1.13 | 4.75 | 1.68 |

Construct a histogram of the transformed data. Compare your histogram to those given in Figure 7.38. Which of the cube-root and square-root transformations appear to result in the more symmetric histogram(s)?

7.88 ● The following data are a sample of survival times (days from diagnosis) for patients suffering from chronic leukemia of a certain type (*Statistical Methodology for Survival Time Studies* [Bethesda, MD: National Cancer Institute, 1986]):

| | | | | | | | |
|------|------|------|------|------|------|------|------|
| 7 | 47 | 58 | 74 | 177 | 232 | 273 | 285 |
| 317 | 429 | 440 | 445 | 455 | 468 | 495 | 497 |
| 532 | 571 | 579 | 581 | 650 | 702 | 715 | 779 |
| 881 | 900 | 930 | 968 | 1077 | 1109 | 1314 | 1334 |
| 1367 | 1534 | 1712 | 1784 | 1877 | 1886 | 2045 | 2056 |
| 2260 | 2429 | 2509 | | | | | |

- Construct a relative frequency distribution for this data set, and draw the corresponding histogram.
- Would you describe this histogram as having a positive or a negative skew?
- Would you recommend transforming the data? Explain.

7.89 ● In a study of warp breakage during the weaving of fabric (*Technometrics* [1982]: 63), 100 pieces of yarn were tested. The number of cycles of strain to breakage

was recorded for each yarn sample. The resulting data are given in the following table:

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 86 | 146 | 251 | 653 | 98 | 249 | 400 | 292 | 131 | 176 |
| 76 | 264 | 15 | 364 | 195 | 262 | 88 | 264 | 42 | 321 |
| 180 | 198 | 38 | 20 | 61 | 121 | 282 | 180 | 325 | 250 |
| 196 | 90 | 229 | 166 | 38 | 337 | 341 | 40 | 40 | 135 |
| 597 | 246 | 211 | 180 | 93 | 571 | 124 | 279 | 81 | 186 |
| 497 | 182 | 423 | 185 | 338 | 290 | 398 | 71 | 246 | 185 |
| 188 | 568 | 55 | 244 | 20 | 284 | 93 | 396 | 203 | 829 |
| 239 | 236 | 277 | 143 | 198 | 264 | 105 | 203 | 124 | 137 |
| 135 | 169 | 157 | 224 | 65 | 315 | 229 | 55 | 286 | 350 |
| 193 | 175 | 220 | 149 | 151 | 353 | 400 | 61 | 194 | 188 |

- Construct a frequency distribution using the class intervals 0 to < 100, 100 to < 200, and so on.
- Draw the histogram corresponding to the frequency distribution in Part (a). How would you describe the shape of this histogram?
- Find a transformation for these data that results in a more symmetric histogram than what you obtained in Part (b).

7.90 The article “The Distribution of Buying Frequency Rates” (*Journal of Marketing Research* [1980]: 210–216) reported the results of a 3½-year study of dentifrice purchases. The investigators conducted their research using a national sample of 2071 households and recorded the number of toothpaste purchases for each household participating in the study. The results are given in the following frequency distribution:

| Number of Purchases | Number of Households (Frequency) |
|---------------------|----------------------------------|
| 10 to <20 | 904 |
| 20 to <30 | 500 |
| 30 to <40 | 258 |
| 40 to <50 | 167 |
| 50 to <60 | 94 |
| 60 to <70 | 56 |
| 70 to <80 | 26 |
| 80 to <90 | 20 |
| 90 to <100 | 13 |
| 100 to <110 | 9 |
| 110 to <120 | 7 |
| 120 to <130 | 6 |
| 130 to <140 | 6 |
| 140 to <150 | 3 |
| 150 to <160 | 0 |
| 160 to <170 | 2 |

- a. Draw a histogram for this frequency distribution. Would you describe the histogram as positively or negatively skewed?
- b. Does the square-root transformation result in a histogram that is more symmetric than that of the original data? (Be careful! This one is a bit tricky, because you don't have the raw data; transforming the endpoints of the class intervals will result in class intervals that are not necessarily of equal widths, so the histogram of the transformed values will have to be drawn with this in mind.)

7.91 ● The paper “Temperature and the Northern Distributions of Wintering Birds” (*Ecology* [1991]: 2274–2285) gave the following body masses (in grams) for 50 different bird species:

| | | | | | | | |
|-------|------|------|------|------|------|------|------|
| 7.7 | 10.1 | 21.6 | 8.6 | 12.0 | 11.4 | 16.6 | 9.4 |
| 11.5 | 9.0 | 8.2 | 20.2 | 48.5 | 21.6 | 26.1 | 6.2 |
| 19.1 | 21.0 | 28.1 | 10.6 | 31.6 | 6.7 | 5.0 | 68.8 |
| 23.9 | 19.8 | 20.1 | 6.0 | 99.6 | 19.8 | 16.5 | 9.0 |
| 448.0 | 21.3 | 17.4 | 36.9 | 34.0 | 41.0 | 15.9 | 12.5 |
| 10.2 | 31.0 | 21.5 | 11.9 | 32.5 | 9.8 | 93.9 | 10.9 |
| 19.6 | 14.5 | | | | | | |

- a. Construct a stem-and-leaf display in which 448.0 is listed separately beside the display as an outlier on the high side, the stem of an observation is the tens digit, the

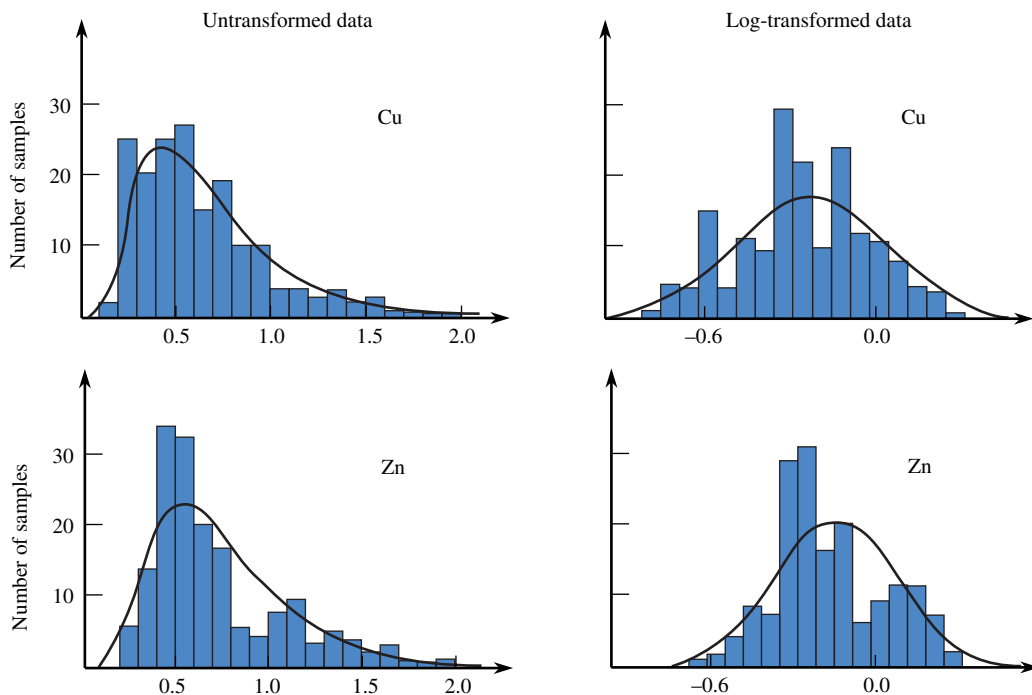
leaf is the ones digit, and the tenths digit is suppressed (e.g., 21.5 has stem 2 and leaf 1). What do you perceive as the most prominent feature of the display?

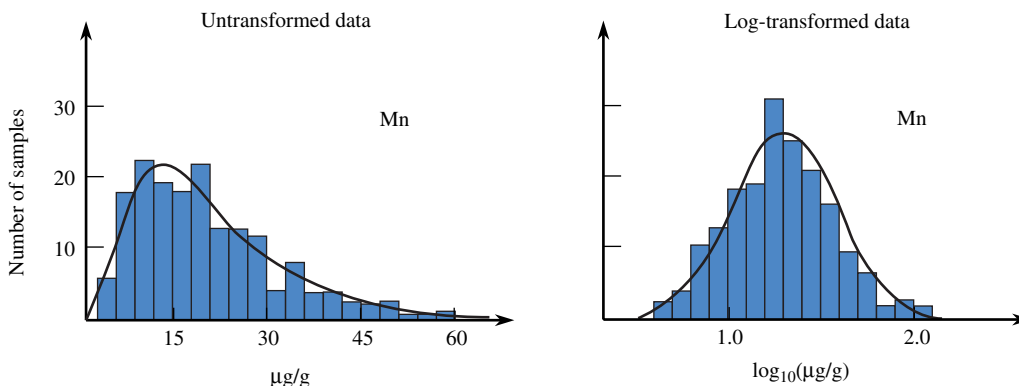
- b. Draw a histogram based on class intervals 5 to <10, 10 to <15, 15 to <20, 20 to <25, 25 to <30, 30 to <40, 40 to <50, 50 to <100, and 100 to <500. Is a transformation of the data desirable? Explain.

- c. Use a calculator or statistical computer package to calculate logarithms of these observations, and construct a histogram. Is the log transformation successful in producing a more symmetric distribution?

- d. Consider transformed value = $\frac{1}{\sqrt{\text{original value}}}$ and construct a histogram of the transformed data. Does it appear to resemble a normal curve?

7.92 The following figure appeared in the paper “EDTA-Extractable Copper, Zinc, and Manganese in Soils of the Canterbury Plains” (*New Zealand Journal of Agricultural Research* [1984]: 207–217). A large number of topsoil samples were analyzed for manganese (Mn), zinc (Zn), and copper (Cu), and the resulting data were summarized using histograms. The investigators transformed each data set using logarithms in an effort to obtain more symmetric distributions of values. Do you think the transformations were successful? Explain.





Bold exercises answered in back ● Data set available online but not required ▼ Video solution available

7.8

Using the Normal Distribution to Approximate a Discrete Distribution

The distribution of many random variables can be approximated by a carefully chosen normal distribution. In this section, we show how probabilities for some discrete random variables can be approximated using a normal curve. The most important case of this is the approximation of binomial probabilities.

■ The Normal Curve and Discrete Variables

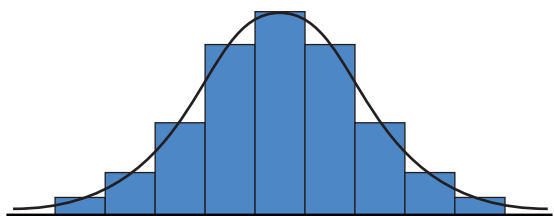


Figure 7.42 A normal curve approximation to a probability histogram.

The probability distribution of a discrete random variable x is represented pictorially by a probability histogram. The probability of a particular value is the area of the rectangle centered at that value. Possible values of x are isolated points on the number line, usually whole numbers. For example, if $x =$ the IQ of a randomly selected 8-year-old child, then x is a discrete random variable, because an IQ score must be a whole number.

Often a probability histogram can be well approximated by a normal curve, as illustrated in Figure 7.42. In such cases, it is customary to say that x has approximately a normal distribution. The normal distribution can then be used to calculate approximate probabilities of events involving x .

Example 7.33 Express Mail Packages

The number of express mail packages mailed at a certain post office on a randomly selected day is approximately normally distributed with mean 18 and standard deviation 6. Let's first calculate the approximate probability that $x = 20$. Figure 7.43(a) shows a portion of the probability histogram for x with the approximating normal curve superimposed. The area of the shaded rectangle is $P(x = 20)$. The left edge of this rectangle is at 19.5 on the horizontal scale, and the right edge is at 20.5.

Therefore, the desired probability is approximately the area under the normal curve between 19.5 and 20.5. Standardizing these limits gives

$$\frac{20.5 - 18}{6} = .42 \quad \frac{19.5 - 18}{6} = .25$$

from which we get

$$P(x = 20) \approx P(.25 < z < .42) = .6628 - .5987 = .0641$$

In a similar fashion, Figure 7.43(b) shows that $P(x \leq 10)$ is approximately the area under the normal curve to the left of 10.5. Then

$$P(x \leq 10) \approx P\left(z \leq \frac{10.5 - 18}{6}\right) = P(z \leq -1.25) = .1056$$

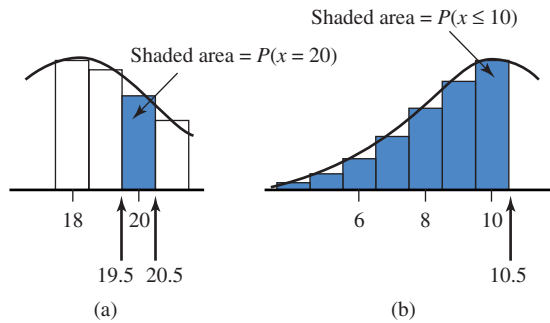


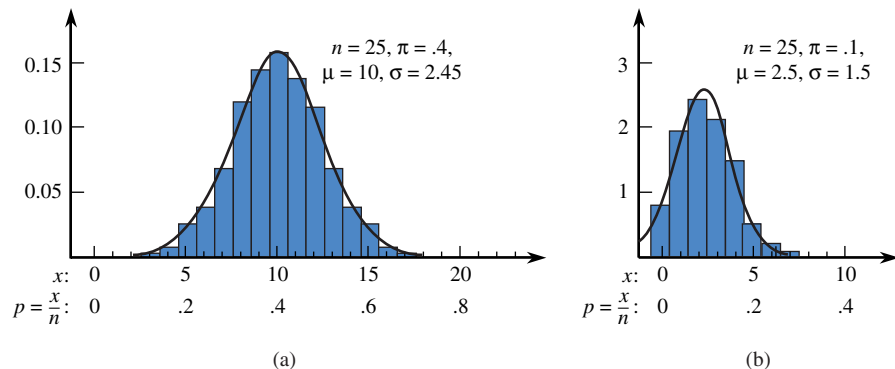
Figure 7.43 The normal approximation for Example 7.33.

The calculation of probabilities in Example 7.33 illustrates the use of what is known as a **continuity correction**. Because the rectangle for $x = 10$ extends to 10.5 on the right, we use the normal curve area to the left of 10.5 rather than 10. In general, if possible x values are consecutive whole numbers, then $P(a \leq x \leq b)$ will be approximately the normal curve area between limits $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

■ Normal Approximation to a Binomial Distribution

Figure 7.44 shows the probability histograms for two binomial distributions, one with $n = 25$, $\pi = .4$, and the other with $n = 25$, $\pi = .1$. For each distribution, we computed $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1 - \pi)}$ and then we superimposed a normal curve with this μ and σ on the corresponding probability histogram. A normal curve fits the probability histogram well in the first case (Figure 7.44(a)). When this happens, binomial probabilities can be accurately approximated by areas under the normal curve. Because of this, statisticians say that both x (the number of successes) and x/n (the proportion of successes) are approximately normally distributed. In the second case (Figure 7.44(b)), the normal curve does not give a good approximation because the probability histogram is skewed, whereas the normal curve is symmetric.

Figure 7.44 Normal approximations to binomial distributions.



Let x be a binomial random variable based on n trials and success probability π , so that

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

If n and π are such that

$$n\pi \geq 10 \quad \text{and} \quad n(1 - \pi) \geq 10$$

then x has approximately a normal distribution. Combining this result with the continuity correction implies that

$$P(a \leq x \leq b) = P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right)$$

That is, the probability that x is between a and b inclusive is approximately the area under the approximating normal curve between $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

Similarly,

$$P(x \leq b) \approx P\left(z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right) \quad P(a \leq x) \approx P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z\right)$$

When either $n\pi < 10$ or $n(1 - \pi) < 10$, the binomial distribution is too skewed for the normal approximation to give accurate results.

Example 7.34 Premature Babies

Premature babies are those born more than 3 weeks early. *Newsweek* (May 16, 1988) reported that 10% of the live births in the United States are premature. Suppose that 250 live births are randomly selected and that the number x of “preemies” is determined. Because

$$\begin{aligned} n\pi &= 250(.1) = 25 \geq 10 \\ n(1 - \pi) &= 250(.9) = 225 \geq 10 \end{aligned}$$

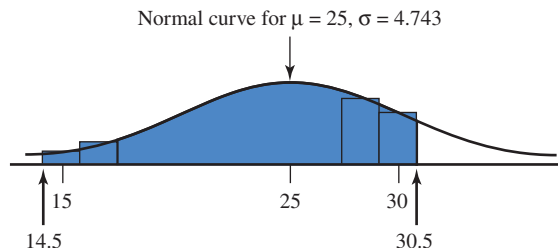
x has approximately a normal distribution, with

$$\begin{aligned} \mu &= 250(.1) = 25 \\ \sigma &= \sqrt{250(.1)(.9)} = 4.743 \end{aligned}$$

The probability that x is between 15 and 30 (inclusive) is

$$\begin{aligned} P(15 \leq x \leq 30) &= P\left(\frac{14.5 - 25}{4.743} \leq z \leq \frac{30.5 - 25}{4.743}\right) \\ &= P(-2.21 \leq z \leq 1.16) \\ &= .8770 - .0136 \\ &= .8634 \end{aligned}$$

as shown in the following figure:



Exercises 7.93–7.101

7.93 Let x denote the IQ for an individual selected at random from a certain population. The value of x must be a whole number. Suppose that the distribution of x can be approximated by a normal distribution with mean value 100 and standard deviation 15. Approximate the following probabilities:

- $P(x = 100)$
- $P(x \leq 110)$
- $P(x < 110)$ (Hint: $x < 110$ is the same as $x \leq 109$.)
- $P(75 \leq x \leq 125)$

7.94 Suppose that the distribution of the number of items x produced by an assembly line during an 8-hr shift can be approximated by a normal distribution with mean value 150 and standard deviation 10.

- What is the probability that the number of items produced is at most 120?
- What is the probability that at least 125 items are produced?
- What is the probability that between 135 and 160 (inclusive) items are produced?

7.95 The number of vehicles leaving a turnpike at a certain exit during a particular time period has approximately a normal distribution with mean value 500 and standard deviation 75. What is the probability that the number of cars exiting during this period is

- At least 650?
- Strictly between 400 and 550? (*Strictly* means that the values 400 and 550 are not included.)
- Between 400 and 550 (inclusive)?

7.96 Let x have a binomial distribution with $n = 50$ and $\pi = .6$, so that $\mu = n\pi = 30$ and $\sigma = \sqrt{n\pi(1 - \pi)} = 3.4641$. Calculate the following probabilities using the normal approximation with the continuity correction:

- $P(x = 30)$
- $P(x = 25)$
- $P(x \leq 25)$
- $P(25 \leq x \leq 40)$
- $P(25 < x < 40)$ (Hint: $25 < x < 40$ is the same as $26 \leq x \leq 39$.)

7.97 Seventy percent of the bicycles sold by a certain store are mountain bikes. Among 100 randomly selected bike purchases, what is the approximate probability that

- At most 75 are mountain bikes?
- Between 60 and 75 (inclusive) are mountain bikes?
- More than 80 are mountain bikes?
- At most 30 are not mountain bikes?

7.98 Suppose that 25% of the fire alarms in a large city are false alarms. Let x denote the number of false alarms in a random sample of 100 alarms. Give approximations to the following probabilities:

- $P(20 \leq x \leq 30)$
- $P(20 < x < 30)$
- $P(35 \leq x)$
- The probability that x is farther than 2 standard deviations from its mean value

7.99 Suppose that 65% of all registered voters in a certain area favor a 7-day waiting period before purchase of a

handgun. Among 225 randomly selected voters, what is the probability that

- At least 150 favor such a waiting period?
- More than 150 favor such a waiting period?
- Fewer than 125 favor such a waiting period?

7.100 Flash bulbs manufactured by a certain company are sometimes defective.

- If 5% of all such bulbs are defective, could the techniques of this section be used to approximate the probability that at least 5 of the bulbs in a random sample of size 50 are defective? If so, calculate this probability; if not, explain why not.
- Reconsider the question posed in Part (a) for the probability that at least 20 bulbs in a random sample of size 500 are defective.

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

Activity 7.1 Rotten Eggs?

Background: *The Salt Lake Tribune* (October 11, 2002) printed the following account of an exchange between a restaurant manager and a health inspector:

The recipe calls for four fresh eggs for each quiche. A Salt Lake County Health Department inspector paid a visit recently and pointed out that research by the Food and Drug Administration indicates that one in four eggs carries salmonella bacterium, so restaurants should never use more than three eggs when preparing quiche. The manager on duty wondered aloud if simply throwing out three eggs from each dozen and using the remaining nine in four-egg quiches would serve the same purpose.

- Working in a group or as a class, discuss the folly of the above statement!
- Suppose the following argument is made for three-egg quiches rather than four-egg quiches: Let $x \leq$ number of eggs that carry salmonella. Then

$$p(0) = p(x = 0) = (0.75)^3 = .422$$

for three-egg quiches and

$$p(0) = p(x = 0) = (0.75)^4 = .316$$

for four-egg quiches. What assumption must be made to justify these probability calculations? Do you think this is reasonable or not? Explain.

7.101 A company that manufactures mufflers for cars offers a lifetime warranty on its products, provided that ownership of the car does not change. Suppose that only 20% of its mufflers are replaced under this warranty.

- In a random sample of 400 purchases, what is the approximate probability that between 75 and 100 (inclusive) mufflers are replaced under warranty?
- Among 400 randomly selected purchases, what is the probability that at most 70 mufflers are ultimately replaced under warranty?
- If you were told that fewer than 50 among 400 randomly selected purchases were ever replaced under warranty, would you question the 20% figure? Explain.

3. Suppose that a carton of one dozen eggs does happen to have exactly three eggs that carry salmonella and that the manager does as he proposes: selects three eggs at random and throws them out, then uses the remaining nine eggs in four-egg quiches. Let $x =$ number of eggs that carry salmonella among four eggs selected at random from the remaining nine.

Working with a partner, conduct a simulation to approximate the distribution of x by carrying out the following sequence of steps:

- Take 12 identical slips of paper and write “Good” on 9 of them and “Bad” on the remaining 3. Place the slips of paper in a paper bag or some other container.
- Mix the slips and then select three at random and remove them from the bag.
- Mix the remaining slips and select four “eggs” from the bags.
- Note the number of bad eggs among the four selected. (This is an observed x value.)
- Replace all slips, so that the bag now contains all 12 “eggs.”
- Repeat Steps (b)–(d) at least 10 times, each time recording the observed x value.

4. Combine the observations from your group with those from the other groups. Use the resulting data to approximate the distribution of x . Comment on the resulting distribution in the context of the risk of salmonella exposure if the manager’s proposed procedure is used.

Summary of Key Terms and Concepts

Term or Formula

Random variable: discrete or continuous

Probability distribution $p(x)$ of a discrete random variable x

Probability distribution of a continuous random variable x

μ_x and σ_x

$$\mu_x = \sum xp(x)$$

$$\sigma_x^2 = \sum (x - \mu_x)^2 p(x)$$

$$\sigma_x = \sqrt{\sigma_x^2}$$

Binomial probability distribution

$$p(x) = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}$$

$$\mu_x = n\pi$$

$$\sigma_x = \sqrt{n\pi(1-\pi)}$$

Normal distribution

Standard normal distribution

z critical value

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

Comment

A numerical variable with a value determined by the outcome of a chance experiment. It is discrete if its possible values are isolated points along the number line and continuous if its possible values form an entire interval on the number line.

A formula, table, or graph that gives the probability associated with each x value. Conditions on $p(x)$ are
 (1) $p(x) \geq 0$, and
 (2) $\sum p(x) = 1$, where the sum is over all possible x values.

Specified by a smooth (density) curve for which the total area under the curve is 1. The probability $P(a < x < b)$ is the area under the curve and above the interval from a to b ; this is also $P(a \leq x \leq b)$.

The mean and standard deviation, respectively, of a random variable x . These quantities describe the center and extent of spread about the center of the variable's probability distribution.

The mean value of a discrete random variable x ; it locates the center of the variable's probability distribution.

The variance and standard deviation, respectively, of a discrete random variable; these are measures of the extent to which the variable's distribution spreads out about the mean μ_x .

This formula gives the probability of observing x successes ($x = 0, 1, \dots, n$) among n trials of a binomial experiment.

The mean and standard deviation of a binomial random variable.

A continuous probability distribution that has a bell-shaped density curve. A particular normal distribution is determined by specifying values of μ and σ .

This is the normal distribution with $\mu = 0$ and $\sigma = 1$. The density curve is called the z curve, and z is the letter commonly used to denote a variable having this distribution. Areas under the z curve to the left of various values are given in Appendix Table 2.

A number on the z measurement scale that captures a specified tail area or central area.

z is obtained by "standardizing": subtracting the mean and then dividing by the standard deviation. When x has a normal distribution, z has a standard normal distribution.

Term or Formula

Normal probability plot

Normal approximation to the binomial distribution


Comment

This fact implies that probabilities involving any normal random variable (any μ or σ) can be obtained from z curve areas.

A picture used to judge the plausibility of the assumption that a sample has been selected from a normal population distribution. If the plot is reasonably straight, this assumption is reasonable.

When both $n\pi \geq 10$ and $n(1 - \pi) \geq 10$, binomial probabilities are well approximated by corresponding areas under a normal curve with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1 - \pi)}$.

Chapter Review Exercises 7.102–7.124

 Know exactly what to study! Take a pre-test and receive your Personalized Learning Plan.

7.102 An article in the *Los Angeles Times* (December 8, 1991) reported that there are 40,000 travel agencies nationwide, of which 11,000 are members of the American Society of Travel Agents (booking a tour through an ASTA member increases the likelihood of a refund in the event of cancellation).

- If x is the number of ASTA members among 5000 randomly selected agencies, could you use the methods of Section 7.8 to approximate $P(1200 < x < 1400)$? Why or why not?
- In a random sample of 100 agencies, what are the mean value and standard deviation of the number of ASTA members?
- If the sample size in Part (b) is doubled, does the standard deviation double? Explain.

7.103 A soft-drink machine dispenses only regular Coke and Diet Coke. Sixty percent of all purchases from this machine are diet drinks. The machine currently has 10 cans of each type. If 15 customers want to purchase drinks before the machine is restocked, what is the probability that each of the 15 is able to purchase the type of drink desired? (Hint: Let x denote the number among the 15 who want a diet drink. For which possible values of x is everyone satisfied?)

7.104 A mail-order computer software business has six telephone lines. Let x denote the number of lines in use at a specified time. The probability distribution of x is as follows:

| | | | | | | | |
|--------|-----|-----|-----|-----|-----|-----|-----|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $p(x)$ | .10 | .15 | .20 | .25 | .20 | .06 | .04 |

Write each of the following events in terms of x , and then calculate the probability of each one:

- At most three lines are in use
- Fewer than three lines are in use
- At least three lines are in use
- Between two and five lines (inclusive) are in use
- Between two and four lines (inclusive) are not in use
- At least four lines are not in use

7.105 Refer to the probability distribution of Exercise 7.104.

- Calculate the mean value and standard deviation of x .
- What is the probability that the number of lines in use is farther than 3 standard deviations from the mean value?

7.106 A new battery's voltage may be acceptable (A) or unacceptable (U). A certain flashlight requires two batteries, so batteries will be independently selected and tested until two acceptable ones have been found. Suppose that 80% of all batteries have acceptable voltages, and let y denote the number of batteries that must be tested.

- What is $p(2)$, that is, $P(y = 2)$?
- What is $p(3)$? (Hint: There are two different outcomes that result in $y = 3$.)
- In order to have $y = 5$, what must be true of the fifth battery selected? List the four outcomes for which $y = 5$, and then determine $p(5)$.

d. Use the pattern in your answers for Parts (a)–(c) to obtain a general formula for $p(y)$.

7.107 A pizza company advertises that it puts 0.5 lb of real mozzarella cheese on its medium pizzas. In fact, the amount of cheese on a randomly selected medium pizza is normally distributed with a mean value of 0.5 lb and a standard deviation of 0.025 lb.

- What is the probability that the amount of cheese on a medium pizza is between 0.525 and 0.550 lb?
- What is the probability that the amount of cheese on a medium pizza exceeds the mean value by more than 2 standard deviations?
- What is the probability that three randomly selected medium pizzas all have at least 0.475 lb of cheese?

7.108 Suppose that fuel efficiency for a particular model car under specified conditions is normally distributed with a mean value of 30.0 mpg and a standard deviation of 1.2 mpg.

- What is the probability that the fuel efficiency for a randomly selected car of this type is between 29 and 31 mpg?
- Would it surprise you to find that the efficiency of a randomly selected car of this model is less than 25 mpg?
- If three cars of this model are randomly selected, what is the probability that all three have efficiencies exceeding 32 mpg?
- Find a number c such that 95% of all cars of this model have efficiencies exceeding c (i.e., $P(x > c) = .95$).

7.109 The amount of time spent by a statistical consultant with a client at their first meeting is a random variable having a normal distribution with a mean value of 60 min and a standard deviation of 10 min.

- What is the probability that more than 45 min is spent at the first meeting?
- What amount of time is exceeded by only 10% of all clients at a first meeting?
- If the consultant assesses a fixed charge of \$10 (for overhead) and then charges \$50 per hour, what is the mean revenue from a client's first meeting?

7.110 The lifetime of a certain brand of battery is normally distributed with a mean value of 6 hr and a standard deviation of 0.8 hr when it is used in a particular cassette player. Suppose that two new batteries are independently selected and put into the player. The player ceases to function as soon as one of the batteries fails.

- What is the probability that the player functions for at least 4 hr?

b. What is the probability that the cassette player works for at most 7 hr?

- Find a number z^* such that only 5% of all cassette players will function without battery replacement for more than z^* hr.

7.111 A machine producing vitamin E capsules operates so that the actual amount of vitamin E in each capsule is normally distributed with a mean of 5 mg and a standard deviation of 0.05 mg. What is the probability that a randomly selected capsule contains less than 4.9 mg of vitamin E? at least 5.2 mg?

7.112 Accurate labeling of packaged meat is difficult because of weight decrease resulting from moisture loss (defined as a percentage of the package's original net weight). Suppose that moisture loss for a package of chicken breasts is normally distributed with mean value 4.0% and standard deviation 1.0%. (This model is suggested in the paper "Drained Weight Labeling for Meat and Poultry: An Economic Analysis of a Regulatory Proposal," *Journal of Consumer Affairs* [1980]: 307–325.) Let x denote the moisture loss for a randomly selected package.

- What is the probability that x is between 3.0% and 5.0%?
- What is the probability that x is at most 4.0%?
- What is the probability that x is at least 7.0%?
- Find a number z^* such that 90% of all packages have moisture losses below z^* %.
- What is the probability that moisture loss differs from the mean value by at least 1%?

7.113 The *Wall Street Journal* (February 15, 1972) reported that General Electric was sued in Texas for sex discrimination over a minimum height requirement of 5 ft 7 in. The suit claimed that this restriction eliminated more than 94% of adult females from consideration. Let x represent the height of a randomly selected adult woman. Suppose that x is approximately normally distributed with mean 66 in. (5 ft 6 in.) and standard deviation 2 in.

- Is the claim that 94% of all women are shorter than 5 ft 7 in. correct?
- What proportion of adult women would be excluded from employment as a result of the height restriction?

7.114 The longest "run" of S 's in the sequence $SSFSSSSFFS$ has length 4, corresponding to the S 's on the fourth, fifth, sixth, and seventh trials. Consider a binomial experiment with $n = 4$, and let y be the length (number of trials) in the longest run of S 's.

- a. When $\pi = .5$, the 16 possible outcomes are equally likely. Determine the probability distribution of y in this case (first list all outcomes and the y value for each one). Then calculate μ_y .
- b. Repeat Part (a) for the case $\pi = .6$.
- c. Let z denote the longest run of either S 's or F 's. Determine the probability distribution of z when $\pi = .5$.

7.115 Two sisters, Allison and Teri, have agreed to meet between 1 and 6 P.M. on a particular day. In fact, Allison is equally likely to arrive at exactly 1 P.M., 2 P.M., 3 P.M., 4 P.M., 5 P.M., or 6 P.M. Teri is also equally likely to arrive at each of these six times, and Allison's and Teri's arrival times are independent of one another. Thus there are 36 equally likely (Allison, Teri) arrival-time pairs, for example, (2, 3) or (6, 1). Suppose that the first person to arrive waits until the second person arrives; let w be the amount of time the first person has to wait.

- a. What is the probability distribution of w ?
- b. How much time do you expect to elapse between the two arrivals?

7.116 Four people—a, b, c, and d—are waiting to give blood. Of these four, a and b have type AB blood, whereas c and d do not. An emergency call has just come in for some type AB blood. If blood samples are taken one by one from the four people in random order for blood typing and x is the number of samples taken to obtain an AB individual (so possible x values are 1, 2, and 3), what is the probability distribution of x ?

7.117 Bob and Lygia are going to play a series of Trivial Pursuit games. The first person to win four games will be declared the winner. Suppose that outcomes of successive games are independent and that the probability of Lygia winning any particular game is $.6$. Define a random variable x as the number of games played in the series.

- a. What is $p(4)$? (Hint: Either Bob or Lygia could win four straight games.)
- b. What is $p(5)$? (Hint: For Lygia to win in exactly five games, what has to happen in the first four games and in Game 5?)
- c. Determine the probability distribution of x .
- d. How many games can you expect the series to last?

7.118 Refer to Exercise 7.117, and let y be the number of games won by the series loser. Determine the probability distribution of y .

7.119 A sporting goods store has a special sale on three brands of tennis balls—call them D, P, and W. Because

the sale price is so low, only one can of balls will be sold to each customer. If 40% of all customers buy Brand W, 35% buy Brand P, and 25% buy Brand D and if x is the number among three randomly selected customers who buy Brand W, what is the probability distribution of x ?

7.120 Suppose that your statistics professor tells you that the scores on a midterm exam were approximately normally distributed with a mean of 78 and a standard deviation of 7. The top 15% of all scores have been designated A's. Your score is 89. Did you receive an A? Explain.

7.121 Suppose that the pH of soil samples taken from a certain geographic region is normally distributed with a mean pH of 6.00 and a standard deviation of 0.10. If the pH of a randomly selected soil sample from this region is determined, answer the following questions about it:

- a. What is the probability that the resulting pH is between 5.90 and 6.15?
- b. What is the probability that the resulting pH exceeds 6.10?
- c. What is the probability that the resulting pH is at most 5.95?
- d. What value will be exceeded by only 5% of all such pH values?

7.122 The lightbulbs used to provide exterior lighting for a large office building have an average lifetime of 700 hr. If length of life is approximately normally distributed with a standard deviation of 50 hr, how often should all the bulbs be replaced so that no more than 20% of the bulbs will have already burned out?

7.123 Suppose that 16% of all drivers in a certain city are uninsured. Consider a random sample of 200 drivers.

- a. What is the mean value of the number who are uninsured, and what is the standard deviation of the number who are uninsured?
- b. What is the (approximate) probability that between 25 and 40 (inclusive) drivers in the sample were uninsured?
- c. If you learned that more than 50 among the 200 drivers were uninsured, would you doubt the 16% figure? Explain.

7.124 Let x denote the duration of a randomly selected pregnancy (the time elapsed between conception and birth). Accepted values for the mean value and standard deviation of x are 266 days and 16 days, respectively. Suppose that the probability distribution of x is (approximately) normal.

- a. What is the probability that the duration of pregnancy is between 250 and 300 days?

- b. What is the probability that the duration of pregnancy is at most 240 days?
- c. What is the probability that the duration of pregnancy is within 16 days of the mean duration?
- d. A “Dear Abby” column dated January 20, 1973, contained a letter from a woman who stated that the duration of her pregnancy was exactly 310 days. (She wrote that the last visit with her husband, who was in the navy, occurred 310 days before birth.) What is the probability that the duration of pregnancy is at least 310 days? Does this probability make you a bit skeptical of the claim?
- e. Some insurance companies will pay the medical expenses associated with childbirth only if the insurance has

been in effect for more than 9 months (275 days). This restriction is designed to ensure that the insurance company pays benefits for only those pregnancies for which conception occurred during coverage. Suppose that conception occurred 2 weeks after coverage began. What is the probability that the insurance company will refuse to pay benefits because of the 275-day insurance requirement?



Personal Tutor

Do you need a live tutor for homework problems?



Are you ready? Take your exam-prep post-test now.

Bold exercises answered in back

● Data set available online but not required

▼ Video solution available

Graphing Calculator Explorations



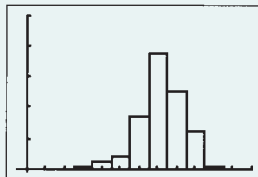
Exploration 7.1 Discrete Probability Distributions

The calculator is at its finest when used with random variables, transforming minutes of mindless calculation into seconds of easy button-pushing. In our calculator presentation of random variables we will capitalize extensively on the list capabilities of your calculator. We will show you not only how to graph a discrete probability distribution, but also how to find the mean and standard deviation of a discrete random variable.

First recall that we have encountered a similar problem before when we considered the problem of graphing a relative frequency histogram of a frequency distribution. At that time frequencies were converted into relative frequencies for plotting; now, these relative frequencies have morphed into probabilities. You begin by entering the possible values of the random variable in your calculator’s equivalent of List1 and the corresponding probabilities of these values in List2. For our example we will use a numerical rating of newborn children called an Apgar score. The Apgar score has eleven possible values, 0, 1, . . . , 10 based on factors such as muscle tone, skin color, etc. Suppose that the scores have the following probability distribution.

| L1 | L2 |
|----|------|
| 0 | .002 |
| 1 | .001 |
| 2 | .002 |
| 3 | .005 |
| 4 | .02 |
| 5 | .04 |
| 6 | .17 |

(a)



(b)

| | | | | | | | | | | | |
|--------|------|------|------|------|-----|-----|-----|-----|-----|-----|-----|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $p(x)$ | .002 | .001 | .002 | .005 | .02 | .04 | .17 | .38 | .25 | .12 | .01 |

In Figure 7.45(a) a portion of the calculator screen after data entry is shown. After you enter the data you can graph the probability distribution by supplying the proper lists in the histogram command as was done for the relative frequency histogram. The graph for the Apgar probability distribution is shown in Figure 7.45(b). The window is set so that the horizontal and vertical axis would show in the screen to give a more informative display. The horizontal axis runs from $-.5$ to 12 , and the vertical axis runs from $-.01$ to 0.5 .

Figure 7.45 (a) Apgar score probability distribution; (b) probability histogram for Apgar score.

Now we turn our attention to calculating the mean and standard deviation of the random variable. The data are already entered and we begin by recalling the definition of the mean of a discrete random variable,

$$\mu_x = \sum_{\text{all possible } x \text{ values}} xp(x)$$

Because we have stored exactly what is needed in Lists 1 and 2, we can virtually duplicate this definition using the language of lists and list operations for our calculator:

$$\mu_x = \sum_{\text{all possible } x \text{ values}} \text{List1} * \text{List2}$$

The strategy for finding the mean of the Apgar random variable, translated from math symbols to English is the following: calculate the products of numbers in our Lists 1 and 2, store the results in List3, and then find the sum of all the numbers in List3. Multiplying to obtain the product is fairly easy as we appeal again to the language of lists:

$$\text{List1} * \text{List2} \rightarrow \text{List3}$$

Now we need to find the sum of the numbers in List3. Exactly how this is done will vary from calculator to calculator. The most likely scenario is for you to calculate the “1-variable statistics” for List3. Calculators will usually report the sum—look for this symbol: $(\sum x)$. Be careful as you scan for the right choice in your calculator window! Don’t be misled by the symbol for the *mean*; we want the *sum*. You should get the value 7.16 for the Apgar mean.

We use a similar strategy to compute the standard deviation. We will find the variance first, then take the square root. The formula for the variance,

$$\sigma_x^2 = \sum_{\text{all possible } x \text{ values}} (x - \mu_x)^2 p(x)$$

also easily translates into the language of lists:

$$\sigma_x^2 = \sum_{\text{all possible } x \text{ values}} (\text{List1} - 7.16)^2 * \text{List2}$$

The list language is only slightly more complicated than for the mean:

$$(\text{List1} - 7.16)^2 * \text{List2} \rightarrow \text{List3}$$

After this production the sum of the numbers in List3 is the variance of the random variable; the square root is the standard deviation. Performing these calculations, you should get a variance of 1.5684 (from $(\sum x)$ for List3; again, don’t be misled and choose the σ_x or s_x . The standard deviation is then found by taking the square root of 1.5684, resulting in 1.2524.



Exploration 7.2 Binomial Probability Calculations

Most calculations having to do with random variables are of one of three types. They are (1) the probability the variable will assume a value between two given numbers, (2) the probability the variable will assume a value less than a given number, or (3) the probability the random variable will assume a value greater than a given number. Because these calculations are so common in statistics your calculator may have a built-in capability for finding these probabilities. In the case of a discrete random variable such as the binomial distribution there is a special case of (1) above, the probability that the random variable will actually assume a particular value.

For the binomial distribution, we will illustrate these calculations using an example. Suppose that 60% of all computer monitors have a flat panel display and 40% have a CRT display. Suppose further that the next 12 purchases monitored, and the random variable is defined as the number of flat panel monitors in the next 12 purchases. As an example, the probability exactly four monitors would be flat panel displays is

$$p(x) = {}_n C_x \pi^x (1 - \pi)^{n-x} = {}_{12} C_4 (.6)^4 (1 - .6)^{12-4} = .042$$

Even if your calculator does not have special binomial functions it is likely to have a key for the combinations, $({}_n C_r)$, possibly cleverly hidden in the “math” or “probability” menu. The calculator keystrokes might look like this (don’t forget that to perform the ${}_n C_r$ calculation above you will have to press n , then the ${}_n C_r$ key, and then r):

$$12 \text{ } {}_n C_r \text{ } 4 * .6^4 * .4^8$$

If your calculator has built-in binomial capabilities you will have fewer keystrokes. Let’s consider these problems one at a time, starting with the function names. If your calculator has a built-in function for binomial calculations, it probably has two: a function for finding the probability that “ x is equal to a given value,” and a function for finding the probability that “ x is less than or equal to a given value.” The first function is known to statisticians as a “density” function, and is commonly abbreviated “pdf” for “probability density function.” The second is known as a “cumulative distribution function” and is commonly labeled “cdf.” These two functions on your calculator will in all likelihood mirror these abbreviations.

The second problem you will face is that to find the binomial probabilities the calculator will need more than just one number, and the order you enter the numbers *does* make a difference! Look in your calculator manual for something that looks like “binomial,” especially with a “pdf” somewhere. The function could be obvious, like “binompdf,” or it may be a little more cryptic, like “binpdf.” Your manual will be very careful to specify both what the needed function parameters are, and the order you should enter them. As an example, one type of calculator has the following:

binompdf(*numtrials*, *p*, [*x*])

The manual informs that “*numtrials*” is the number of trials, “*p*” is the probability of success, and that “*x*” can be either an integer or a list of integers. This information collectively explains what is known as the “syntax” of the function. It is your responsibility to get the numbers right, and get them in the right order! The square brackets, “[],” are a standard notation in the calculator world. They indicate the bracketed quantity is either optional or defaults to a preselected option if you do not enter a number in that space. For our example, the number of trials is 12, and the probability of success is 0.6. Since the probability of exactly 4 flat panel monitors is desired, we enter

binompdf(12, .6, 4)

Our calculator gives us 0.042042, which is the correct answer. This is a good sign! Now try this on your calculator. Remember, you must navigate to the function in the manner presented in your manual, and you have to pay attention to the syntax. While you are learning how to use this function (or any calculator function), it is a very good idea to use examples with known answers and check the results.

Now suppose you wish to find the probability of getting 4 *or fewer* flat panel monitors out of 12. The appropriate function here is the cumulative density function, or “cdf”:

binomcdf(*numtrials*, *p*, [*x*])

Does this look disturbingly familiar? Except for the “c” instead of the “p,” they look exactly alike! The good news is that we already understand the syntax; the bad news is that if we aren’t careful we might get the wrong function in haste. Be careful! The function **binomcdf**(12, .6, 4) gives the answer, .0573099213. If we are not convinced of our prowess with **binomcdf** we can use **binompdf** to check the result:

$$\begin{aligned} & \mathbf{binompdf}(12, .6, 0) + \mathbf{binompdf}(12, .6, 1) + \cdots + \mathbf{binompdf}(12, .6, 4) \\ &= .000016777 + .0003019898 + \cdots + .042042 \\ &= .05731 \end{aligned}$$

Now let’s move on to another of the common calculations with random variables: What is the probability that the random variable will assume a value between 4 and 7? One common source of confusion here is that the word *between* is disturbingly ambiguous. Do we mean to include 4 and 7, or do we mean only the probability of getting a 5 or 6? In this case we wish to be inclusive. To evaluate the probability desired, we will use the *cumulative* distribution function for the binomial, **binomcdf**. (Your calculator function may have a different name!) The logic is elementary, as Sherlock Holmes would say. The probability that the binomial random variable will assume a value between 4 and 7 (inclusive) is equal to the probability of assuming a value less than or equal to 7, minus the probability of assuming a value less than or equal to 3 (*not* 4!). In symbols,

$$p(4 \leq x \leq 7) = p(x \leq 7) - p(x \leq 3)$$

which is found as follows:

$$\mathbf{binomcdf}(12, .6, 7) - \mathbf{binomcdf}(12, .6, 3)$$

gives us 0.5465545.

Our last binomial random variable calculation problem is finding the probability of a value greater than a given value. What, for instance, is the probability of more than 7 monitors having flat panel displays? Using a fundamental property of probability, we know that:

$$p(7 < x) + p(x \leq 7) = 1$$

and thus

$$p(7 < x) = 1 - p(x \leq 7)$$

which we translate into:

$$1 - \mathbf{binomcdf}(12, .6, 7)$$

giving us 0.438178222.

We have gone into some detail to explain how these binomial probability problems can be solved using the calculator. This detail is justified not only because of the importance of the binomial distribution, but also because these same calculator procedures will be used for finding probabilities involving the geometric and normal random variables, yet to come. Because the discussions in this exploration have been detailed, the discussions in those cases will be less so.

In Exploration 7.1 we discussed how to graph a discrete distribution. When we graphed the probability density function for the Apgar scores we manually entered the outcomes and their associated probabilities. Anticipating that you may wish to consider binomial chance experiments with many potential successes, we will streamline the data entry process using some commands and functions we have already discussed in previous calculator explorations.

Graphing a binomial distribution will involve three steps:

1. Construct the list of possible values in List1 using the seq command (or your calculator's equivalent).
2. Construct the probabilities in List2 using the binompdf function (or your calculator's equivalent).
3. Draw the graph (in the form of a histogram) of the probability distribution.

Consider the binomial probability distribution for $n = 20$ and $\pi = .20$. Carrying out the steps below puts the integers 0 to 20 in List1, and $p(x)$ for x values from 0 to 20 in List2.

1. seq(x, x, 0, 20) → List1 puts a sequence of 21 integers into List1. (Remember to verify your calculator syntax and the order of the information to be entered for your calculator!)
2. binompdf(20, .2) → List2. (Remember to verify . . .)
3. Now graph the probability distribution, where List1 contains the possible data values and List2 contains the probabilities.

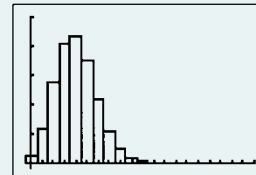
To check your work, partial calculator screen output for this problem is given in Figure 7.46(a), and the graph of the distribution is shown in Figure 7.46(b).

Figure 7.46

(a) Binomial probabilities;
(b) histogram of binomial distribution.

| L1 | L2 |
|----|--------|
| 0 | .01153 |
| 1 | .05765 |
| 2 | .13691 |
| 3 | .20536 |
| 4 | .2182 |
| 5 | .17456 |
| 6 | .1091 |

(a)



(b)



Exploration 7.3 Geometric Probability Calculations

Our calculator exploration of geometric random variables will be an echo of the binomial random variables we have already discussed in Exploration 7.2. We again consider (1) the probability the variable will assume a value between two given numbers, (2) the probability the variable will assume a value less than a given number, and (3) the probability the random variable will assume a value greater than a given number.

Our example here will be about jumper cables. Suppose that 40% of students who drive to campus carry jumper cables. If your car has a dead battery, and you aren't one of the forward thinking 40%, how many students will you have to ask before you find one with jumper cables?

Consider the first problem, the probability of a particular number. The probability the first student stopped has jumper cables is:

$$p(1) = (1 - \pi)^{1-1}\pi = (1 - .4)^{1-1}(.4) = .4$$

The corresponding keystrokes for finding this probability will be something like

$$(1 - 0.4)^0 * 0.4.$$

Now let's consider problems (2) and (3). If your calculator has density and cumulative density functions for the geometric distribution the functions are probably named something like geompdf and geomcdf, similar to the names for the binomial

functions. The calculator syntax for the probability density function will probably look something like

$$\mathbf{geompdf}(p, x)$$

where p is the probability of success and x is in this example the number of students you would ask until success. We want the probability of jumper cables on the very first stop. We enter $\mathbf{geompdf}(.4, 1)$, and the function returns .4.

Now suppose you wish to find the probability of jumper cables after 4 or fewer stops. Using the cumulative density function, “ $\mathbf{geomcdf}$ ” (which has the same parameters as the $\mathbf{geompdf}$ function), we enter $\mathbf{geomcdf}(.4, 4)$, which returns 0.8704. As with the binomial, we can check this by summing:

$$\begin{aligned} \mathbf{geompdf}(.4, 1) + \mathbf{geompdf}(.4, 2) + \mathbf{geompdf}(.4, 4) + \mathbf{geompdf}(.4, 4) \\ = 0.4 + 0.24 + 0.144 + 0.0864 \\ = 0.87041 \end{aligned}$$

The probability that a geometric random variable will assume a value between 4 and 7 (inclusive) is equal to the probability of observing a value less than or equal to 7, minus the probability of observing a value less than or equal to 3 (not 4). In symbols,

$$P(4 \leq x \leq 7) = P(x \leq 7) - P(x \leq 3)$$

which is found using $\mathbf{geomcdf}(.4, 7) - \mathbf{geomcdf}(.4, 3)$, giving 0.1880064.

What is the probability of more than 7 stops before we get jumper cables?

$$1 - \mathbf{geomcdf}(.4, 7) \text{ gives us } 0.0279936.$$

Graphing an entire geometric probability distribution is not possible, since there is an infinite number of possible values—1, 2, 3, . . . Nevertheless, we can graph parts of the distribution. The method for graphing is similar to that for the binomial random variable. Use the \mathbf{seq} function to create a list of integers in List1; then use the $\mathbf{geompdf}$ function to find the corresponding probabilities and store them in List2; and finally plot the distribution as you would a histogram, and as we have previously done with the binomial. These steps for the geometric distribution of Example 7.20 are summarized below:

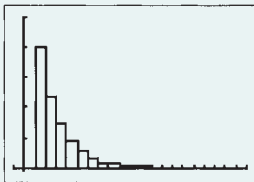
1. $\mathbf{seq}(x, x, 1, 20) \rightarrow \text{List1}$
2. $\mathbf{geompdf}(.4, \text{List1}) \rightarrow \text{List2}$
3. Graph a histogram with the domain List1, and probabilities in List2.

The data editing window is shown in Figure 7.47(a) and a graph of this geometric distribution appears in Figure 7.47(b).

We are really calculating probabilities for only part of the distribution, since the number of possible values is infinite. The graph should tail to the right in a gradual manner, not suddenly drop out of sight. You may notice a sudden plummeting in your graph but it could be that there are more significant probabilities to the right. As an example, suppose we consider the chance experiment of flipping a coin until a head appears. The distribution of $x =$ number of tosses is geometric with success probability .5. If the distribution is plotted using the previous steps but only using a sequence of integers from 1 to 4, the results are shown in Figure 7.48 (a). Clearly, there are values with probabilities different from zero that are not represented in the graph. The solution is to construct the sequence of integers over a larger range of values, say 1 to 16.

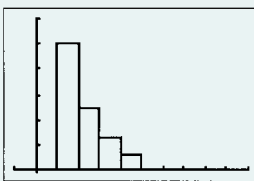
| L1 | L2 |
|----|--------|
| 1 | .4 |
| 2 | .24 |
| 3 | .144 |
| 4 | .0864 |
| 5 | .05184 |
| 6 | .0311 |
| 7 | .01866 |

(a)

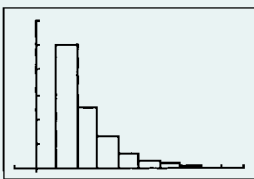


(b)

Figure 7.47 (a) Geometric probability distribution; (b) histogram of geometric probability distribution.



(a)



(b)

Figure 7.48 (a) Incorrect display; (b) correct display.

At some point, of course, the geometric probabilities become very close to zero, as Figure 7.48 (b). If your graph looks similar to the one on the right, tailing off gradually, you can be fairly certain you have captured the essential behavior of the particular geometric distribution.



Exploration 7.4 Normal Curves and the Normal Probability Distribution

The normal distribution is arguably the most famous distribution in all of statistics. As we have learned “the” normal distribution is really a family of distributions with the same shape, but different means and standard deviations. The “standard” normal distribution is the normal probability distribution with $\mu = 0$, and $\sigma = 1.0$. From the calculator perspective working with the normal distribution is slightly different from the binomial and geometric distributions because the normal distribution is continuous. Consequently it will not be graphed as a histogram; normal curves are graphed just as any other function is graphed. The normal curve, however, is not particularly simple. Fortunately your calculator, if it has statistical functions, will have “normal” already in it somewhere. It might be something like this:

normalpdf(x , [μ , σ]).

If you are a glutton for punishment or your calculator does not have a built-in normalpdf function, here is the formula for a normal curve with mean μ and standard deviation σ :

$$y = \frac{1}{\sqrt{2\pi}} e^{[(x-\mu)^2/(2\sigma^2)]}$$

Here are the keystrokes for the formula:

$$y1 = 1/(\text{sqr}(2*\pi)*\sigma)*\exp(-(x-\mu)^2/(2*\sigma^2))$$

Assuming you are smiling because of your foresight in purchasing a calculator with a built-in normalpdf function, let’s put it to good use. The syntax above for the **normalpdf** function might seem complicated but actual use is simple once you get used to it. You should check your calculator manual for two *very important* pieces of information. First, make sure you know the required order for the information you must provide. Second, look closely at the sigma, wherever it is in your calculator’s syntax. Make sure you check whether you must enter (the standard deviation) or (the variance). Now let’s tackle the notation. First of all, if your calculator’s syntax has those square brackets— $[\mu, \sigma]$ —remember that they indicate numbers that are optional. If you leave them out, the normalpdf function will simply default to the standard normal curve, with mean 0 and standard deviation (or variance) 1.

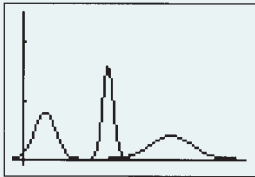
Let’s graph the three normal curves. The first has a mean and standard deviation of 10 and 5, respectively. The second has a mean and standard deviation of 40 and 2.5, and the third a mean and standard deviation of 70 and 10. Navigate your calculator’s menu system to find the normal curve function, and paste this function into the function definition window where you usually define simpler functions. Using the syntax above, you should see your calculator’s equivalent of the following:

$$\begin{aligned} y1 &= \text{normalpdf}(x, 10, 5) \\ y2 &= \text{normalpdf}(x, 40, 2.5) \\ y3 &= \text{normalpdf}(x, 70, 10) \end{aligned}$$

```

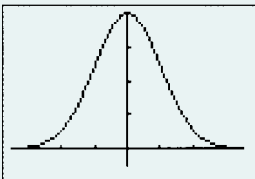
WINDOW
Xmin=-5
Xmax=105
Xscl=50
Ymin=-.01
Ymax=.25
Yscl=.1
Xres=■
    
```

(a)



(b)

Figure 7.49 (a) Window settings; (b) normal curves.



7.50 The standard normal distribution.

```

Normal C.D.

Lower :0
Upper :0
σ :0
μ :0
Execute
    
```

7.51 Setup for normal calculations.

Graphing these functions using the window setting in Figure 7.49(a), we see the graphs in Figure 7.49 (b).

Now let’s graph the standard normal distribution. If your calculator syntax indicates that it defaults to a standard normal, you will only have to enter your calculator equivalent of

$$y1 = \mathbf{normalpdf}(x).$$

It is also possible that your calculator does *not* default to standard normal, in which case you would have to specify the mean and standard deviation as 0 and 1, something like

$$y1 = \mathbf{normalpdf}(x, 0, 1).$$

Set your graphing window with x values running from about -3.5 to 3.5 and the y values from to 0.40 . These values should be fine for the standard normal distribution. If you don’t see a distribution filling the screen as in Figure 7.50, something is amiss and you need to verify your keystrokes and check your calculator’s manual.

Since the normal probability distribution is a continuous distribution the probability that x would be equal to a specific value is, of course, 0. For continuous distributions we are usually interested in finding (1) the area under the curve between two specific values; and (2) the area in the extremes, or “tails,” of the distribution. The function that we will use to find these values will be symbolized with the notation “**normalcdf**,” which stands for the “normal cumulative distribution function.” This actually is a misnomer, because the functions calling themselves “cdf” functions on many calculators actually calculate the probability that the standard normal variable is *between* two values. Calculators seem to get it right for the discrete probability density functions, but for some reason have elected to use similar names for very different kinds of calculation when they get to the continuous probability density functions—don’t let this minor inconvenience confuse you!

Two strategies are used by calculator manufacturers for evaluating the probability that z is between two values, a and b . It is possible your calculator has a table for you to fill in the values as in Figure 7.51. For a calculator utilizing this strategy, you would have to fill in the lower bound, upper bound, and standard deviation (σ) and mean (μ).

Other calculators ask for the mean and standard deviation as parameters of the function. If your calculator uses this strategy, your built-in cumulative distribution function will have syntax something like this:

$$\mathbf{normalcdf}(\text{lower bound, upper bound} [, \mu, \sigma])$$

For a calculator using this syntax you would fill in the lower bound and upper bound with the appropriate values for z , and ignore the optional parameters, since z will have a standard normal distribution. You will specify other values for μ and σ when performing calculations that are not already in terms of z scores. After navigating your calculator’s menus, you will enter something like this:

$$\mathbf{normalcdf}(z\text{-lower, } z\text{-upper}).$$

Let’s find the probability that z is between -1.76 and $.58$. We enter the function as **normalcdf**($-1.76, 0.58$) and the calculator will return a value of 0.6798388789 . We would not suggest writing all those digits; rounding off to $.6798$ is perfectly fine (as you may have surmised from considering Appendix Table 2).

As you might guess from its name, the **normalcdf** function can also be used for calculation of the cumulative distribution function—that is, finding the probability that

z will be below a specific value. Suppose we want to find the probability that z is less than -1.76 . Remembering that the set of possible values for a standard normal random variable is the entire real line, you might think to enter the following:

`normalcdf(-∞, -1.76)`.

If so, your thinking is right on target, except for one thing: there is no “ $-\infty$ ” on your calculator. Some calculators will have a special symbol for “ $-\infty$ ” which the calculator translates internally to its equivalent of a “very small number.” You should check your manual for this number and how to find it. The representation will probably be something like “ $-1E99$ ” or “ $-1e999$ ” which is calculator-speak for -1 times 10 raised to the highest power the calculator can handle. In the case of the standard normal curve, it may be just as easy to enter a different but still very small number in place of the “ $-\infty$,” perhaps `normalcdf(-10, -1.76)`. On our calculator .0392038577 is returned, which agrees with the tabled answer. (If you are squeamish about -10 , use -50 ; using -50 we get .0392038577 also!)

Finding the probability that a z is greater than a particular value is also easy. For example, we find the area to the right of $z = 1.42$ as follows:

`1 - normalcdf(“-∞”, 1.42)`

Using -10 for the lower bound, we get 0.0778038883.

The last type of problem examined will be the identification of extreme values. The easiest way to do this is with a built-in function, typically called “InvNormal,” which stands for “inverse normal.” The “InvNormal” function—or whatever it is named on your calculator—will be the reverse of finding the probability that z is less than a specified value. Earlier in our discussion, we found the probability that z is less than -1.76 to be 0.0392038577. The InvNormal function returns a z value when given the probability. Thus, `InvNormal(.0392038577)` equals -1.76 . Except for the difference in function name, the syntax for this function should be the same as for **normalcdf**:

InvNormal(cumulative probability, [, μ , σ])

On our calculator, `InvNormal(0.0392038577)` returns -1.760000538 .



Exploration 7.5 The Normal Approximation to the Binomial Distribution

In an earlier Exploration we showed you how to use your calculator to find probabilities associated with the binomial and normal distributions, using built-in calculator functions. We generically used the terms *binompdf*, *binomcdf*, *normalpdf*, and *normalcdf* to refer to these functions. In this Exploration we would like to focus on the normal approximation to the binomial distribution. Whenever a continuous distribution is used to approximate a discrete distribution the question naturally occurs, “How good is the approximation?” The answer usually given by statistics instructors is, “it depends.” In the case of the normal approximation to the binomial, the goodness of fit depends on the two quantities which define the binomial distribution: n and π . Most statisticians have a simple “rule of thumb” they apply for approximating the binomial with a normal distribution, such as:

When either $n\pi < 10$ or $n(1 - \pi) < 10$, the binomial distribution is too skewed for the normal approximation to give accurate results.

Different statisticians have different rules of thumb, some feeling comfortable with the accuracy provided by using 5 instead of 10 in the rule of thumb above. In days

of yore—that is, the precalculator days—students would have to accept the rule of thumb as one of the mysteries of statistics. In more modern times a statistics student, armed with her calculator, can not only understand what the rules of thumb are all about, but evaluate the various rules of thumb for a particular n and π pair.

It might be argued that using the normal distribution to approximate a distribution that we can evaluate exactly seems a little foolish. There is something to this argument, but remember: we will not always be able to find exact probabilities in other situations in statistics, and must rely on approximations. Using an approximation involves a fundamental tradeoff between ease of calculation and exactness of answer. An understanding of this with the normal approximation to the binomial will give us a better understanding of the issues involved when we encounter similar tradeoffs in statistics courses yet to come. (At least you'll be more tolerant of those "rules of thumb!")

We shall reacquaint ourselves with some syntax and warm up with a distribution of the number of express mail packages mailed at a certain post office in a day. The number is approximately normally distributed with $\mu = 18$ and $\sigma = 6$. Suppose we wish to find the probability that 20 express mail packages are mailed in a given day. We calculate the probability that in a normal distribution with $\mu = 18$ and $\sigma = 6$, the event $x = 20$ would happen. Remembering the syntax from our earlier discussion,

normalcdf(lower bound, upper bound [, μ , σ])

we enter: normalcdf(19.5, 20.5, 18, 6), and our calculator returns 0.062832569.

We will now compare binomial calculations with the normal approximations. It is reported that 10% of live births in the United States are premature. Suppose we randomly select 250 live births and define the random variable x to be the number of these that are premature. We wish to calculate the probability that x is between 15 and 30 (inclusive). To find the binomial probability we recall that we must use the built-in function we called binomcdf. This function includes the rightmost interval indicated; therefore we subtract the probability of getting x less than or equal to 14 from the probability of getting x less than or equal to 30.

$$\begin{aligned} P(15 \leq x \leq 30) &= \text{binomcdf}(250, .1, 30) - \text{binomcdf}(250, .1, 14) \\ &= 0.8753286537 - 0.00931244487 \\ &= 0.8660162088. \end{aligned}$$

To evaluate this probability using the normal curve approximation we will use the machine accuracy of the calculator with the mean $\mu = 25$ and $\sigma = 4.74341649$:

$$\begin{aligned} &\text{normalcdf}(\text{lower bound, upper bound [, } \mu, \sigma]) \\ &= \text{normalcdf}(14.5, 30.5, 25, 4.74341649) \\ &= 0.8634457937 \end{aligned}$$

The difference between the two probabilities to machine accuracy is 0.0025704151. This does not seem to be a large difference, but *it is* a difference. According to the rule of thumb this approximation meets the test, but the investigator in the context of his or her situation must evaluate the practical importance of the difference.

Now lets redo the calculations, not with a sample size of 250, but a sample size of only 50. Keeping the results proportionally the same by dividing by 5, we will consider approximating the probability of getting between 3 and 6 preemies (inclusive) from a random sample of 50 babies. In this case,

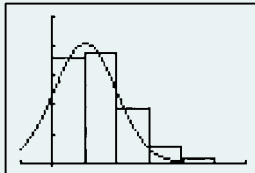
$$\begin{aligned} n\pi &= 50(.10) = 5 < 10 \\ n(1 - \pi) &= 50(1 - .10) = 45 \geq 10 \end{aligned}$$

Since $n\pi < 10$ our rule of thumb would regard the binomial distribution too skewed for the normal curve approximation to give accurate results. Let's see what happens:

$$\begin{aligned} P(3 \leq x \leq 6) &= \text{binomcdf}(50, .1, 6) - \text{binomcdf}(50, .1, 2) \\ &= 0.7702268435 - 0.1117287563 \\ &= 0.6584980872 \end{aligned}$$

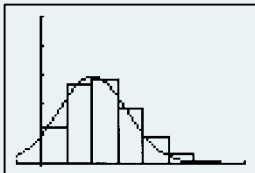
To evaluate this probability using the normal curve approximation we will use the machine accuracy of the calculator with $\mu = 50(.1) = 5$ and $\sigma = \sqrt{50(.1)(.9)} = 2.121320344$

$$\begin{aligned} &\text{normalcdf}(\text{lower bound}, \text{upper bound} [, \mu, \sigma]) \\ &= \text{normalcdf}(2.5, 6.5, 5, 2.121320344) \\ &= 0.6409535402 \end{aligned}$$



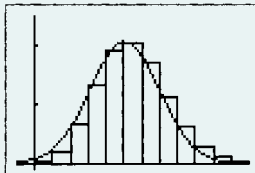
$n = 20; \pi = .05$

(a)



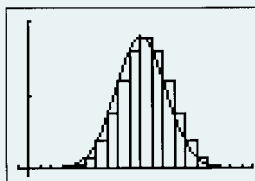
$n = 20; \pi = .10$

(b)



$n = 20; \pi = .25$

(c)



$n = 20; \pi = .50$

(d)

Figure 7.52 Binomial distributions: (a) $n = 20$, $\pi = .05$; (b) $n = 20$, $\pi = .10$; (c) $n = 20$, $\pi = .25$; (d) $n = 20$, $\pi = .50$.

The difference between the binomial and the normal approximation in this case is 0.017544547. It is interesting to note that using a rule of thumb with 5 instead of 10 would call this difference “acceptable.” We do not argue with this rule of thumb in principle, but once again point out that the individual judgment by the investigator on site must be used in evaluating the goodness of the approximation.

Finally, we will superimpose the appropriate normal distribution over the binomial distribution to get a visual sense of the approximation. It is entirely possible that a given approximation will do a better job for different choices of values of the end points of the interval, and the graphs may give us an overall sense of when a normal approximation might be acceptable.

Graphing a binomial distribution and a normal distribution at the same time involves skills we have seen in previous Explorations. (You may want to refer back to the calculator explorations about the binomial and normal distributions to refresh your memory.)

We will graph the binomial and normal distributions for four distributions, each with sample size 20, but with probabilities of success of .05, .1, .25, and .5. We will change the windows to make the graphs fill the windows, but this should not affect any interpretations of the goodness of fit to the binomial by the normal distribution. As a reminder, our binomial preparations for the first graph are

1. $\text{seq}(x, x, 0, 20) \rightarrow \text{List1}$
2. $\text{binompdf}(20, .05) \rightarrow \text{List2}$
3. Specify that we want a histogram with the values in List1, and the corresponding binomial probabilities in List2.

For the normal curve plot, define the graphing function by supplying the mean and standard deviation of the binomial as parameters for the normalpdf function:

$$Y1 = \text{normalpdf}(x, 1, 0.97468)$$

The four plots appear in Figure 7.52.

As can be seen from a comparison of the plots, the normal approximation gets “closer and closer” to the binomial as π gets closer and closer to 0.5. For $\pi = .25$ the rule of thumb is satisfied for $n = 20$, and for $\pi = .5$, the rule of thumb is satisfied using $n = 10$. It is a bit difficult to judge whether or not the normal approximation to the binomial is “adequate” for a particular situation by just looking at the plots.

Modern technology makes it possible to do binomial calculations quickly, so the normal approximation to the binomial is not as widely used as it once was. However, there are other distributions in statistics that are “approximately” normal as long as certain conditions are satisfied. We hope that working with the approximation to the binomial has given you an appreciation for the uses of the normal distribution to approximate these other distributions.