
Volodymyr Goinyk/Shutterstock.com

CHAPTER 14

Probability and Statistics

In the preceding chapters we modeled real-world situations using precise rules, such as equations and functions. But many of our everyday activities are not governed by precise rules, but rather involve randomness. It is remarkable that there are also rules that govern randomness. For example, if we toss a balanced coin many times, we can be pretty sure that "heads" will show up about half of the time. Such patterns in apparently haphazard events allow us to use mathematics to model randomness.

Probability is the mathematical study of chance. Knowing the chance, or probability, of an event happening can be very useful. For example, insurance companies estimate the probability of an automobile accident happening. This allows the company to calculate a reasonable price to charge their customers.

Statistics is the art of collecting and organizing data and then getting information from the data. We usually collect data from a sample of a population. We can then use probability to obtain information about the entire population from properties of the sample. For example, to test the effectiveness of an allergy medication, we try the medication on a sample of the population. The data on the effectiveness of the medication for the individuals in the sample allow us to estimate the probability that the medication is effective for the population in general.

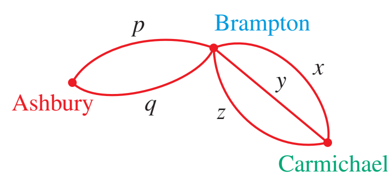
14.1 COUNTING

The Fundamental Counting Principle ▶ Counting Permutations ▶ Counting Combinations ▶ Problem Solving with Permutations and Combinations

Counting the number of apples in a bag or the number of students in an algebra class is easy. But counting all the different ways in which these students can stand in a row is more difficult. It is this latter kind of counting that we'll study in this section.

▼ The Fundamental Counting Principle

Suppose that three towns—Ashbury, Brampton, and Carmichael—are located in such a way that two roads connect Ashbury to Brampton and three roads connect Brampton to Carmichael.



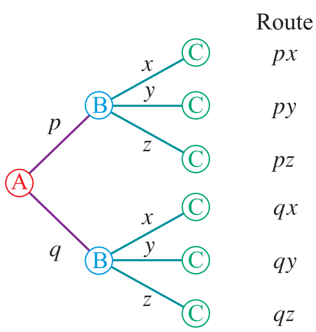


FIGURE 1 Tree diagram

How many different routes can one take to travel from Ashbury to Carmichael via Brampton? The key to answering this question is to consider the problem in stages. At the first stage—from Ashbury to Brampton—there are two choices. For each of these choices there are three choices at the second stage—from Brampton to Carmichael. Thus the number of different routes is $2 \times 3 = 6$. These routes are conveniently enumerated by a *tree diagram* as in Figure 1. The method that we used to solve this problem leads to the following principle.

THE FUNDAMENTAL COUNTING PRINCIPLE

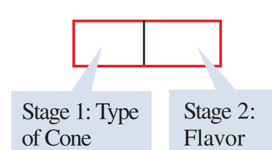
Suppose that two events occur in order. If the first event can occur in m ways and the second can occur in n ways (after the first has occurred), then the two events can occur *in order* in $m \times n$ ways.

There is an immediate consequence of this principle for any number of events: If $E_1, E_2, \ldots, E_k$ are events that occur in order and if E_1 can occur in n_1 ways, E_2 in n_2 ways, and so on, then the events can occur in order in $n_1 \times n_2 \times \cdots \times n_k$ ways.

EXAMPLE 1 | Using the Fundamental Counting Principle

An ice-cream store offers three types of cones and 31 flavors. How many different single-scoop ice-cream cones is it possible to buy at this store?

SOLUTION There are two stages for selecting an ice-cream cone. At the first stage we choose a type of cone, and at the second stage we choose a flavor. We can think of the different stages as boxes:



The first box can be filled in three ways, and the second can be filled in 31 ways:

Stage 1	Stage 2
3	31

By the Fundamental Counting Principle there are $3 \times 31 = 93$ ways of choosing a single-scoop ice-cream cone at this store.

NOW TRY EXERCISE **17**

EXAMPLE 2 | Using the Fundamental Counting Principle

In a certain state, automobile license plates display three letters followed by three digits. How many such plates are possible if repetition of the letters

(a) is allowed? **(b)** is not allowed?

SOLUTION

(a) There are six selection stages, one for each letter or digit on the license plate. As in the preceding example, we sketch a box for each stage:

Letters			Digits		
26	26	26	10	10	10

At the first stage we choose a letter (from 26 possible choices); at the second stage we choose another letter (again from 26 choices); at the third stage we choose another letter (26 choices); at the fourth stage we choose a digit (from 10 possible choices); at the fifth stage we choose a digit (again from 10 choices); and at the sixth stage, we choose another digit (10 choices). By the Fundamental Counting Principle the number of possible license plates is

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17{,}576{,}000$$

(b) If repetition of letters is not allowed, then we arrange the choices as follows:

Letters			Digits		
26	25	24	10	10	10

At the first stage we have 26 letters to choose from, but once the first letter has been chosen, there are only 25 letters to choose from at the second stage. Once the first two letters have been chosen, 24 letters are left to choose from for the third stage. The digits are chosen as before. Thus the number of possible license plates in this case is

$$26 \times 25 \times 24 \times 10 \times 10 \times 10 = 15{,}600{,}000$$

NOW TRY EXERCISE **29**

Let S be a set with n elements. A subset of S can be chosen by making one of two choices for each element: We can choose the element to be *in* or *out* of A. Since S has n elements and we have two choices for each element, by the Fundamental Counting Principle the total number of different subsets is $2 \times 2 \times \cdots \times 2$, where there are n factors. This gives the following formula.

THE NUMBER OF SUBSETS OF A SET

A set with n elements has 2^n different subsets.

EXAMPLE 3 | Finding the Number of Subsets

A pizza parlor offers a basic cheese pizza and a choice of 16 toppings. How many different kinds of pizza can be ordered at this pizza parlor?

SOLUTION We need the number of possible subsets of the 16 toppings (including the empty set, which corresponds to a plain cheese pizza). Thus

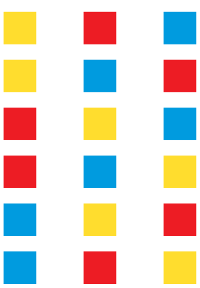
$$2^{16} = 65{,}536$$

different pizzas can be ordered.

NOW TRY EXERCISE **37**

Counting Permutations

Permutations of three colored squares



A **permutation** of a set of distinct objects is an ordering of these objects. For example, some permutations of the letters *ABCD* are

ABDC *BACD* *DCBA* *DABC*

How many such permutations are possible? There are four choices for the first position, three for the second (after the first has been chosen), two for the third (after the first two have been chosen), and only one choice for the fourth letter (the letter that has not yet been chosen). By the Fundamental Counting Principle the number of possible permutations is

$$4 \times 3 \times 2 \times 1 = 4! = 24$$

The same reasoning with 4 replaced by n leads to the following.

The number of permutations of n objects is $n!$

How many permutations consisting of two letters can be made from these same four letters? Some of these permutations are *AB, AC, BD, DB,* . . . . There are 4 choices of the first letter and 3 for the second letter. By the Fundamental Counting Principle there are $4 \times 3 = 12$ such permutations. In general, if a set has n elements, then the number of ways of ordering r elements from the set is denoted by $\boldsymbol{P(n, r)}$ and is called **the number of permutations of n objects taken r at a time**.

PERMUTATIONS OF *n* OBJECTS TAKEN *r* AT A TIME

The number of permutations of n objects taken r at a time is

$$P(n, r) = \frac{n!}{(n - r)!}$$

PROOF There are n objects and r positions to place them in. Thus there are n choices for the first position, $n - 1$ choices for the second, $n - 2$ choices for the third, and so on. The last position can be filled in $n - r + 1$ ways. By the Fundamental Counting Principle we conclude that

$$P(n, r) = n(n-1)(n-2)\cdots(n-r+1)$$

We can express this formula using factorial notation by multiplying numerator and denominator by $(n - r) \cdots 3 \cdot 2 \cdot 1$:

$$P(n, r) = \frac{n(n-1)(n-2)\cdots(n-r+1)(n-r)\cdots 3\cdot 2\cdot 1}{(n-r)\cdots 3\cdot 2\cdot 1} = \frac{n!}{(n-r)!}$$

EXAMPLE 4 | Finding the Number of Permutations

There are six runners in a race that is completed with no tie.

(a) In how many different ways can the race be completed?

(b) In how many different ways can first, second, and third place be decided?

SOLUTION

(a) The number of ways to complete the race is the number of permutations of the six runners: $6! = 720$.

(b) The number of ways in which the first three positions can be decided is:

$$P(6, 3) = \frac{6!}{(6-3)!} = \frac{6 \times 5 \times 4 \times \cancel{3 \times 2 \times 1}}{\cancel{3 \times 2 \times 1}} = 120$$

NOW TRY EXERCISE **41**

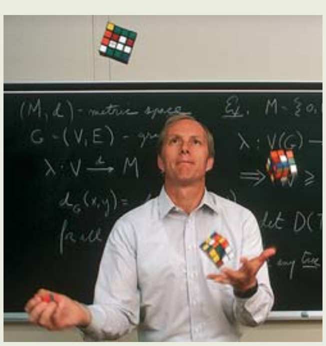
EXAMPLE 5 | Finding the Number of Permutations

A club has nine members. In how many ways can a president, vice president, and secretary be chosen from the members of this club?

SOLUTION We need the number of ways of selecting three members, in order, for the positions of president, vice president, and secretary from the nine club members. This number is

$$P(9, 3) = \frac{9!}{(9-3)!} = \frac{9 \times 8 \times 7 \times \cancel{6 \times 5 \times 4 \times 3 \times 2 \times 1}}{\cancel{6 \times 5 \times 4 \times 3 \times 2 \times 1}} = 504$$

NOW TRY EXERCISE **43**



Ken Regan/Camera 5

RONALD GRAHAM, born in Taft, California, in 1935, is considered the world's leading mathematician in the field of combinatorics, the branch of mathematics that deals with counting. For many years Graham headed the Mathematical Studies Center at Bell Laboratories in Murray Hill, New Jersey, where he solved key problems for the telephone industry. During the *Apollo* program, NASA needed to evaluate mission schedules so that the three astronauts aboard the spacecraft could find the time to perform all the necessary tasks. The number of ways to allot these tasks was astronomical—too vast for even a computer to sort out. Graham, using his knowledge of combinatorics, was able to reassure NASA that there were easy ways of solving their problem that were not too far from the theoretically best possible solution. Besides being a prolific mathematician, Graham is an accomplished juggler (he has been on stage with the Cirque du Soleil and is a past president of the International Jugglers Association). Several of his research papers address the mathematical aspects of juggling. He is also fluent in Mandarin Chinese and Japanese and once spoke with former President Jiang of China in his native language.

EXAMPLE 6 | Finding the Number of Permutations

From 20 raffle tickets in a hat, 4 tickets are to be selected in order. The holder of the first ticket wins a car, the second a motorcycle, the third a bicycle, and the fourth a skateboard. In how many different ways can these prizes be awarded?

SOLUTION The order in which the tickets are chosen determines who wins each prize. So we need to find the number of ways of selecting 4 objects, in order, from 20 objects (the tickets). This number is

$$P(20, 4) = \frac{20!}{(20-4)!} = \frac{20 \times 19 \times 18 \times 17 \times \cancel{16 \times 15 \times 14 \times \cdots \times 3 \times 2 \times 1}}{\cancel{16 \times 15 \times 14 \times \cdots \times 3 \times 2 \times 1}} = 116{,}280$$

NOW TRY EXERCISE **45** ■

Counting Combinations

When counting permutations, we are interested in the number of ways of ordering the elements of a set. In many counting problems, however, order is not important. For example, a poker hand is the same hand regardless of how it is ordered. A poker player who is interested in the number of possible hands wants to know the number of ways of drawing five cards from 52 cards, without regard to the order in which the cards are dealt. We now develop a formula for counting in situations in which order doesn't matter.

A **combination** of r elements of a set is any subset of r elements from the set (without regard to order). If the set has n elements, then the number of combinations of r elements is denoted by $\boldsymbol{C(n, r)}$ and is called the **number of combinations of $\boldsymbol{n}$ elements taken $\boldsymbol{r}$ at a time.** For example, consider a set with the four elements A, B, C, and D. The combinations of these four elements taken three at a time are listed below. Compare this with the permutations of these elements listed in the margin.

ABC	*ABD*	*ACD*	*BCD*
ACB	*ADB*	*ADC*	*BDC*
BAC	*BAD*	*CAD*	*CBD*
BCA	*BDA*	*CDA*	*CDB*
CAB	*DAB*	*DAC*	*DBC*
CBA	*DBA*	*DCA*	*DCB*

ABC   *ABD*   *ACD*   *BCD*

We notice that the number of combinations is a lot fewer than the number of permutations. In fact, each combination of three elements generates $3!$ permutations. So $C(4, 3) = P(4, 3)/3! = 4$. In general, each combination of r objects gives rise to $r!$ permutations of these objects, so we get the following formula.

COMBINATIONS OF $\boldsymbol{n}$ OBJECTS TAKEN $\boldsymbol{r}$ AT A TIME

The number of combinations of n objects taken r at a time is

$$C(n, r) = \frac{n!}{r!\,(n-r)!}$$

The key difference between permutations and combinations is order. If we are interested in ordered arrangements, then we are counting permutations, but if we are concerned with subsets without regard to order, then we are counting combinations. Compare Examples 7 and 8 below (where order doesn't matter) to Examples 5 and 6 (where order does matter).

EXAMPLE 7 | Finding the Number of Combinations

A club has nine members. In how many ways can a committee of three be chosen from the members of this club?

SOLUTION We need the number of ways of choosing three of the nine members. Order is not important here, because the committee is the same no matter how its members are ordered. So we want the number of combinations of nine objects (the club members) taken three at a time. This number is

$$C(9, 3) = \frac{9!}{3!(9-3)!} = \frac{9 \times 8 \times 7 \times \cancel{6 \times 5 \times 4 \times 3 \times 2 \times 1}}{(3 \times 2 \times 1) \times (\cancel{6 \times 5 \times 4 \times 3 \times 2 \times 1})} = 84$$

NOW TRY EXERCISE **53** ■

EXAMPLE 8 | Finding the Number of Combinations

From 20 raffle tickets in a hat, four tickets are to be chosen at random. The holders of the winning tickets get free trips to the Bahamas. In how many ways can the four winners be chosen?

SOLUTION We need to find the number of ways of choosing four winners from 20 entries. The order in which the tickets are chosen doesn't matter, because the same prize is awarded to each of the four winners. So we want the number of combinations of 20 objects (the tickets) taken four at a time. This number is

$$C(20, 4) = \frac{20!}{4!(20-4)!} = \frac{20 \times 19 \times 18 \times 17 \times \cancel{16 \times 15 \times 14 \times \cdots \times 3 \times 2 \times 1}}{(4 \times 3 \times 2 \times 1) \times (\cancel{16 \times 15 \times 14 \times \cdots \times 3 \times 2 \times 1})} = 4845$$

NOW TRY EXERCISE **55** ■

▼ Problem Solving with Permutations and Combinations

The crucial step in solving counting problems is deciding whether to use permutations, combinations, or the Fundamental Counting Principle. In some cases the solution of a problem may require using more than one of these principles. Here are some general guidelines to help us decide how to apply these principles.

GUIDELINES FOR SOLVING COUNTING PROBLEMS

1. **Fundamental Counting Principle.** When consecutive choices are being made, use the Fundamental Counting Principle.
2. **Does Order Matter?** When we want to find the number of ways of picking r objects from n objects, we need to ask ourselves, "Does the order in which we pick the objects matter?"

   If the order matters, we use permutations.

   If the order doesn't matter, we use combinations.

EXAMPLE 9 | Using Combinations

A group of 25 campers consists of 15 women and 10 men. In how many ways can a scouting party of 6 be chosen if it must consist of 3 women and 2 men?

SOLUTION Three women can be chosen from the 15 women in $C(15, 3)$ ways, and two men can be chosen from the 10 men in $C(10, 2)$ ways. It follows by the Fundamental Counting Principle that the number of ways of choosing the scouting party is

$$C(15, 3) \times C(10, 2) = 455 \times 45 = 20{,}475$$

NOW TRY EXERCISE **67** ■

EXAMPLE 10 | Using Permutations and Combinations

A committee of seven—consisting of a chairman, a vice chairman, a secretary, and four other members—is to be chosen from a class of 20 students. In how many ways can the committee be chosen?

SOLUTION In choosing the three officers, order is important. So the number of ways of choosing them is

$$P(20, 3) = 6840$$

Next, we need to choose four other students from the 17 remaining. Since order doesn't matter in choosing these four members, the number of ways of doing this is

$$C(17, 4) = 2380$$

By the Fundamental Counting Principle the number of ways of choosing this committee is

$$P(20, 3) \times C(17, 4) = 6840 \times 2380 = 16{,}279{,}200$$

NOW TRY EXERCISE **69**

EXAMPLE 11 | Using Permutations and Combinations


Jane John

Twelve employees at a company picnic are to stand in a row for a group photograph. In how many ways can this be done if

(a) Jane and John insist on standing next to each other?

(b) Jane and John refuse to stand next to each other?

SOLUTION Since the order in which the people stand is important, we use permutations. But we can't use permutations directly.

(a) Since Jane and John insist on standing together, let's think of them as one object. So we have 11 objects to arrange in a row, and there are $P(11, 11)$ ways of doing this. For each of these arrangements there are two ways of having Jane and John stand together: Jane-John or John-Jane. By the Fundamental Counting Principle the total number of arrangements is

$$2 \times P(11, 11) = 2 \times 11! = 79{,}833{,}600$$

(b) There are $P(12, 12)$ ways of arranging the 12 people. Of these, $2 \times P(11, 11)$ have Jane and John standing together (by part (a)). All the rest have Jane and John standing apart. So the number of arrangements with Jane and John standing apart is

$$P(12, 12) - 2 \times P(11, 11) = 12! - 2 \times 11! = 399{,}168{,}000$$

NOW TRY EXERCISE **77**

14.1 EXERCISES

CONCEPTS

1. The Fundamental Counting Principle says that if one event can occur in m ways and a second event can occur in n ways, then the two events can occur in order in _____ × _____ ways. So if you have two choices for shoes and three choices for hats, then the number of different shoe-hat combinations you can wear is _____ × _____ = _____.

2. The number of ways of arranging r objects from n objects in order is called the number of ____________ of n objects taken r at a time and is given by the formula $P(n, r) =$ ________.

3. The number of ways of choosing r objects from n objects is called the number of ___________ of n objects taken r at a time and is given by the formula $C(n, r) =$ ___________.

4. *True or false*?
(a) In counting combinations, order matters.
(b) In counting permutations, order matters.
(c) For a set of n distinct objects, the number of different combinations of these objects is more than the number of different permutations.
(d) If we have a set with five distinct objects, then the number of different ways of choosing two members of this set is the same as the number of ways of choosing three members.

SKILLS

5–16 ■ Evaluate the expression.

5. $P(8, 3)$	**6.** $P(9, 2)$	**7.** $P(11, 4)$
8. $P(10, 5)$	**9.** $P(100, 1)$	**10.** $P(99, 3)$
11. $C(8, 3)$	**12.** $C(9, 2)$	**13.** $C(11, 4)$
14. $C(10, 5)$	**15.** $C(100, 1)$	**16.** $C(99, 3)$

APPLICATIONS

Exercises 17–36 involve the Fundamental Counting Principle.

17. Ice-Cream Cones A vendor sells ice cream from a cart on the boardwalk. He offers vanilla, chocolate, strawberry, and pistachio ice cream, served in either a waffle, sugar, or plain cone. How many different single-scoop ice-cream cones can you buy from this vendor?

18. Three-Letter Words How many three-letter "words" (strings of letters) can be formed by using the 26 letters of the alphabet if repetition of letters
(a) is allowed? (b) is not allowed?

19. Horse Race Eight horses compete in a race. (Assume the race does not end in a tie.)
(a) How many different orders are possible for completing the race?
(b) In how many different ways can first, second, and third places be decided?

20. Multiple-Choice Test A multiple-choice test has five questions with four choices for each question. In how many different ways can the test be completed?

21. Phone Numbers Telephone numbers consist of seven digits; the first digit cannot be 0 or 1. How many telephone numbers are possible?

22. Running a Race In how many different ways can a race with five runners be completed? (Assume that there is no tie.)

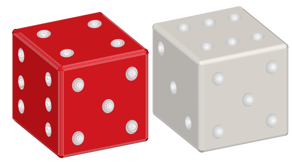
23. Restaurant Meals A restaurant offers the items listed in the table. How many different meals consisting of a main course, a drink, and a dessert can be selected at this restaurant?

Main courses	Drinks	Desserts
Chicken Beef Lasagna Quiche	Iced tea Apple juice Cola Ginger ale Coffee	Ice cream Layer cake Blueberry pie

24. Multiple Routes Towns A, B, C, and D are located in such a way that there are four roads from A to B, five roads from B to C, and six roads from C to D. How many routes are there from town A to town D via towns B and C?

25. Flipping a Coin A coin is flipped five times, and the resulting sequence of heads and tails is recorded. How many such sequences are possible?

26. Rolling a Pair of Dice A red die and a white die are rolled, and the numbers that show are recorded. How many different outcomes are possible? (The singular form of the word *dice* is *die*.)



27. Rolling Three Dice A red die, a blue die, and a white die are rolled, and the numbers that show are recorded. How many different outcomes are possible?

28. Choosing Outfits A girl has five skirts, eight blouses, and 12 pairs of shoes. How many different skirt-blouse-shoe outfits can she wear? (Assume that each item matches all the others, so she is willing to wear any combination.)

29. License Plates Standard automobile license plates in California display a nonzero digit, followed by three letters, followed by three digits. How many different standard plates are possible in this system?

30. ID Numbers A company's employee ID number system consists of one letter followed by three digits. How many different ID numbers are possible with this system?

31. Combination Lock A combination lock has 60 different positions. To open the lock, the dial is turned to a certain number in the clockwise direction, then to a number in the counterclockwise direction, and finally to a third number in the clockwise direction. If successive numbers in the combination cannot be the same, how many different combinations are possible?

32. License Plates A state has registered 8 million automobiles. To simplify the license plate system, a state employee suggests that each plate display only two letters followed by three digits. Will this system create enough different license plates for all the vehicles that are registered?

33. Class Executive In how many ways can a president, vice president, and secretary be chosen from a class of 30 students?

34. Committee Officers A senate subcommittee consists of ten Democrats and seven Republicans. In how many ways can a chairman, vice chairman, and secretary be chosen if the chairman must be a Democrat and the vice chairman must be a Republican?

35. Social Security Numbers Social Security numbers consist of nine digits, with the first digit between 0 and 6, inclusive. How many Social Security numbers are possible?

36. Holiday Photos A couple have seven children: three girls and four boys. In how many ways can the children be arranged for a holiday photo if the girls sit in a row in the front and the boys stand in a row behind the girls?

Exercises 37–40 involve counting subsets.

37. Subsets A set has eight elements.
(a) How many subsets containing five elements does this set have?
(b) How many subsets does this set have?

38. Travel Brochures A travel agency has limited numbers of eight different free brochures about Australia. The agent tells you to take any that you like but no more than one of any kind. In how many different ways can you choose brochures (including not choosing any)?

39. Hamburgers A hamburger chain gives their customers a choice of ten different hamburger toppings. In how many different ways can a customer order a hamburger?

40. To Shop or Not to Shop Each of 20 shoppers in a shopping mall chooses to enter or not to enter the Dressfastic clothing store. How many different outcomes of their decisions are possible?

Exercises 41–52 involve counting permutations.

41. Seating Arrangements Ten people are at a party.
(a) In how many different ways can they be seated in a row of ten chairs?
(b) In how many different ways can six of these people be selected and then seated in a row of six chairs?

42. Three-Letter Words How many three-letter "words" can be made from the letters FGHIJK? (Letters may not be repeated.)

43. Class Officers In how many different ways can a president, vice president, and secretary be chosen from a class of 15 students?

44. Three-Digit Numbers How many different three-digit whole numbers can be formed by using the digits 1, 3, 5, and 7 if no repetition of digits is allowed?

45. Contest Prizes In how many different ways can first, second, and third prizes be awarded in a game with eight contestants?

46. Piano Recital A pianist plans to play eight pieces at a recital. In how many ways can she arrange these pieces in the program?

47. Running a Race In how many different ways can a race with nine runners be completed, assuming that there is no tie?

48. Signal Flags A ship carries five signal flags of different colors. How many different signals can be sent by hoisting exactly three of the five flags on the ship's flagpole in different orders?

49. Contest Prizes In how many ways can first, second, and third prizes be awarded in a contest with 1000 contestants?

50. Class Officers In how many ways can a president, vice president, secretary, and treasurer be chosen from a class of 30 students?

51. Seating Arrangements In how many ways can five students be seated in a row of five chairs if Jack insists on sitting in the first chair?

52. Seating Arrangements In how many ways can the students in Exercise 51 be seated if Jack insists on sitting in the middle chair?

Exercises 53–66 involve counting combinations.

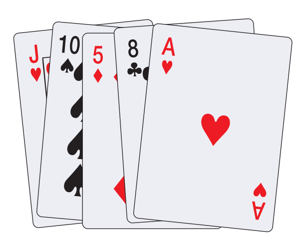
53. Committee In how many ways can a committee of three members be chosen from a club of 25 members?

54. Choosing Books In how many ways can three books be chosen from a group of six different books?

55. Raffle In a raffle with 12 entries, in how many ways can three winners be selected?

56. Choosing a Group In how many ways can six people be chosen from a group of ten?

57. Draw Poker Hands How many different five-card hands can be dealt from a deck of 52 cards?



58. Stud Poker Hands How many different seven-card hands can be picked from a deck of 52 cards?

59. Choosing Exam Questions A student must answer seven of the ten questions on an exam. In how many ways can she choose the seven questions?

60. Three-Topping Pizzas A pizza parlor offers a choice of 16 different toppings. How many three-topping pizzas are possible?

61. Violin Recital A violinist has practiced 12 pieces. In how many ways can he choose eight of these pieces for a recital?

62. Choosing Clothing If a woman has eight skirts, in how many ways can she choose five of these to take on a weekend trip?

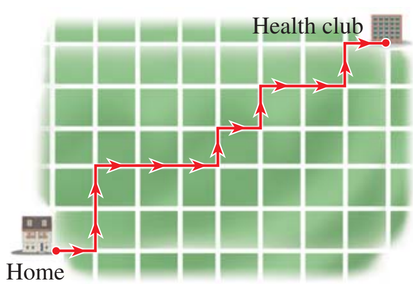
63. Choosing Clothing If a man has ten pairs of pants, in how many ways can he choose three of these to take on a business trip?

64. Field Trip From a class with 30 students, seven are to be chosen to go on a field trip. Find the number of different ways that the seven students can be chosen under the given condition.

(a) Jack must go on the field trip.
(b) Jack is not allowed to go on the field trip.
(c) There are no restrictions on who can go on the field trip.

65. Lottery In the 6/49 lottery game, a player picks six numbers from 1 to 49. How many different choices does the player have?

66. Jogging Routes A jogger jogs every morning to his health club, which is eight blocks east and five blocks north of his home. He always takes a route that is as short as possible, but he likes to vary it (see the figure). How many different routes can he take? [*Hint:* The route shown can be thought of as *ENNEEENENEENE,* where *E* is East and *N* is North.]



Solve Exercises 67–82 by using the appropriate counting principle(s).

67. Choosing a Committee A class has 20 students, of whom 12 are females and 8 are males. In how many ways can a committee of five students be picked from this class under each condition?

(a) No restriction is placed on the number of males or females on the committee.
(b) No males are to be included on the committee.
(c) The committee must have three females and two males.

68. Doubles Tennis From a group of ten male and ten female tennis players, two men and two women are to face each other in a men-versus-women doubles match. In how many different ways can this match be arranged?

69. Choosing a Committee A committee of six is to be chosen from a class of 20 students. The committee is to consist of a chair, a secretary and four other members. In how many different ways can the committee be picked?

70. Choosing a Group Sixteen boys and nine girls go on a camping trip. In how many ways can a group of six be selected to gather firewood, given the following conditions?

(a) The group consists of two girls and four boys.
(b) The group contains at least two girls.

71. Dance Committee A school dance committee is to consist of two freshmen, three sophomores, four juniors, and five seniors. If six freshmen, eight sophomores, twelve juniors, and ten seniors are eligible to be on the committee, in how many ways can the committee be chosen?

72. Casting a Play A group of 22 aspiring thespians contains 10 men and 12 women. For the next play, the director wants to choose a leading man, a leading lady, a supporting male role, a supporting female role, and eight extras—three women and five men. In how many ways can the cast be chosen?

73. Hockey Lineup A hockey team has 20 players, of whom 12 play forward, six play defense, and two are goalies. In how many ways can the coach pick a starting lineup consisting of three forwards, two defense players, and one goalie?

74. Choosing a Pizza A pizza parlor offers four sizes of pizza (small, medium, large, and colossus), two types of crust (thick and thin), and 14 different toppings. How many different pizzas can be made with these choices?

75. Choosing a Committee In how many ways can a committee of four be chosen from a group of ten if Barry and Harry refuse to serve together on the same committee?

76. Parking Committee A five-person committee consisting of students and teachers is being formed to study the issue of student parking privileges. Of those who have expressed an interest in serving on the committee, 12 are teachers and 14 are students. In how many ways can the committee be formed if at least one student and one teacher must be included?

77. Arranging Books In how many ways can five different mathematics books be placed on a shelf if the two algebra books are to be placed next to each other?

78. Arranging a Class Picture In how many ways can ten students be arranged in a row for a class picture if John and Jane want to stand next to each other and Mike and Molly also insist on standing next to each other?

79. Seating Arrangements In how many ways can four men and four women be seated in a row of eight seats for each of the following arrangements?

(a) The first seat is to be occupied by a man.
(b) The first and last seats are to be occupied by women.

80. Seating Arrangements In how many ways can four men and four women be seated in a row of eight seats for each of the following arrangements?

(a) The women are to be seated together.

(b) The men and women are to be seated alternately by gender.

81. Selecting Prizewinners From a group of 30 contestants, six are to be chosen as semifinalists, then two of those are chosen as finalists, and then the top prize is awarded to one of the finalists. In how many ways can these choices be made in sequence?

82. Choosing a Delegation Three delegates are to be chosen from a group of four lawyers, a priest, and three professors. In how many ways can the delegation be chosen if it must include at least one professor?

DISCOVERY ▪ DISCUSSION ▪ WRITING

83. Pairs of Initials Explain why in any group of 677 people, at least two people must have the same pair of initials.

84. Complementary Combinations Without performing any calculations, explain in words why the number of ways of choosing two objects from ten objects is the same as the number of ways of choosing eight objects from ten objects. In general, explain why

$$C(n, r) = C(n, n - r)$$

85. An Identity Involving Combinations Kevin has ten different marbles, and he wants to give three of them to Luke and two to Mark. In how many ways can he choose to do this? There are two ways of analyzing this problem: He could first pick three for Luke and then two for Mark, or he could first pick two for Mark and then three for Luke. Explain how these two viewpoints show that

$$C(10, 3) \cdot C(7, 2) = C(10, 2) \cdot C(8, 3)$$

In general, explain why

$$C(n, r) \cdot C(n - r, k) = C(n, k) \cdot C(n - k, r)$$

86. Why Is $\binom{n}{r}$ the Same as $C(n, r)$? This exercise explains why the binomial coefficients $\binom{n}{r}$ that appear in the expansion of $(x + y)^n$ are the same as $C(n, r)$, the number of ways of choosing r objects from n objects. First, note that expanding a binomial using only the Distributive Property gives

$$\begin{aligned}(x + y)^2 &= (x + y)(x + y)\\ &= (x + y)x + (x + y)y\\ &= xx + xy + yx + yy\\ (x + y)^3 &= (x + y)(xx + xy + yx + yy)\\ &= xxx + xxy + xyx + xyy + yxx\\ &\quad + yxy + yyx + yyy\end{aligned}$$

(a) Expand $(x + y)^5$ using only the Distributive Property.

(b) Write all the terms that represent x^2y^3. These are all the terms that contain two x's and three y's.

(c) Note that the two x's appear in all possible positions. Conclude that the number of terms that represent x^2y^3 is $C(5, 2)$.

(d) In general, explain why $\binom{n}{r}$ in the Binomial Theorem is the same as $C(n, r)$.

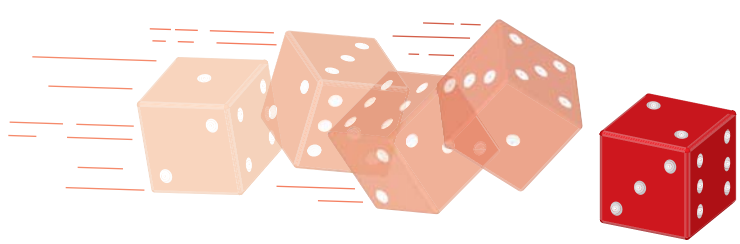
14.2 PROBABILITY

What Is Probability? ▶ Calculating Probability by Counting ▶ The Complement of an Event ▶ The Union of Events ▶ Conditional Probability and the Intersection of Events

In this section we study probability, which is the mathematical study of "chance."

▼ What Is Probability?

Suppose we roll a die, and we're hoping to get a "two." Of course, it's impossible to predict what number will show up. But here's the key idea: If we roll the die many many times, a "two" will show up about one-sixth of the time. If you try this experiment you'll see that it actually works! We say that the *probability* (or chance) of getting a "two" is $\frac{1}{6}$.



The mathematical theory of probability was first discussed in 1654 in a series of letters between Pascal (see page 818) and Fermat (see page 99). Their correspondence was prompted by a question raised by the experienced gambler the Chevalier de Méré. The Chevalier was interested in the equitable distribution of the stakes of an interrupted gambling game (see Problem 3, page 968).

To discuss probability, let's begin by defining some terms. An **experiment** is a process, such as tossing a coin, that gives definite results, called the **outcomes** of the experiment. The **sample space** of an experiment is the set of all possible outcomes. If we let H stand for heads and T for tails, then the sample space of the coin-tossing experiment is $S = \{H, T\}$. The table gives some experiments and their sample spaces.

Experiment	Sample space
Tossing a coin	$\{H, T\}$
Rolling a die	$\{1, 2, 3, 4, 5, 6\}$
Tossing a coin twice and observing the sequence of heads and tails	$\{HH, HT, TH, TT\}$
Picking a card from a deck and observing the suit	{♠, ♥, ♦, ♣}
Administering a drug to three patients and observing whether they recover (R) or not (N)	$\{RRR, RRN, RNR, RNN, NRR, NRN, NNR, NNN\}$

We will be concerned only with experiments for which all the outcomes are **equally likely**. For example, when we toss a perfectly balanced coin, heads and tails are equally likely outcomes in the sense that if this experiment is repeated many times, we expect that about as many heads as tails will show up.

In any given experiment we are often concerned with a particular set of outcomes. We might be interested in a die showing an even number or in picking an ace from a deck of cards. Any particular set of outcomes is a subset of the sample space. This leads to the following definition.

DEFINITION OF AN EVENT

If S is the sample space of an experiment, then an **event** E is any subset of the sample space.

EXAMPLE 1 | Events in a Sample Space

An experiment consists of tossing a coin three times and recording the results in order. List the outcomes in the sample space, then list the outcome in each event.

(a) The event E of getting "exactly two heads."

(b) The event F of getting "at least two heads."

(c) The event G of getting "no heads."

SOLUTION We write H for heads and T for tails. So the outcome HTH means that the three tosses resulted in Heads, Tails, Heads, in that order. The sample space is

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$$

(a) The event E is the subset of the sample space S that consists of all outcomes with exactly two heads. Thus

$$E = \{HHT, HTH, THH\}$$

(b) The event F is the subset of the sample space S that consists of all outcomes with at least two heads. Thus

$$F = \{HHH, HHT, HTH, THH\}$$

PERSI DIACONIS (b. 1945) is currently professor of statistics and mathematics at Stanford University in California. He was born in New York City into a musical family and studied violin until the age of 14. At that time he left home to become a magician. He was a magician (apprentice and master) for ten years. Magic is still his passion, and if there were a professorship for magic, he would certainly qualify for such a post! His interest in card tricks led him to a study of probability and statistics. He is now one of the leading statisticians in the world. With his unusual background he approaches mathematics with an undeniable flair. He says, "Statistics is the physics of numbers. Numbers seem to arise in the world in an orderly fashion. When we examine the world, the same regularities seem to appear again and again." Among his many original contributions to mathematics is a probabilistic study of the perfect card shuffle.

(c) The event G is the subset of the sample space S that consists of all outcomes with no heads. Thus

$$G = \{TTT\}$$

NOW TRY EXERCISE **5**

We are now ready to define the notion of probability. Intuitively, we know that rolling a die may result in any of six equally likely outcomes, so the chance of any particular outcome occurring is $\frac{1}{6}$. What is the chance of showing an even number? Of the six equally likely outcomes possible, three are even numbers. So it is reasonable to say that the chance of showing an even number is $\frac{3}{6} = \frac{1}{2}$. This reasoning is the intuitive basis for the following definition of probability.

DEFINITION OF PROBABILITY

Let S be the sample space of an experiment in which all outcomes are equally likely, and let E be an event. Then the **probability** of E, written $P(E)$, is

$$P(E) = \frac{n(E)}{n(S)} = \frac{\text{number of elements in } E}{\text{number of elements in } S}$$

Notice that $0 \le n(E) \le n(S)$, so the probability $P(E)$ of an event is a number between 0 and 1, that is,

$$0 \le P(E) \le 1$$

The closer the probability of an event is to 1, the more likely the event is to happen; the closer to 0, the less likely. If $P(E) = 1$, then E is called a **certain event**; if $P(E) = 0$, then E is called an **impossible event**.

EXAMPLE 2 | Finding the Probability of an Event

A coin is tossed three times, and the results are recorded in order. Find the probability of the following.

(a) The event E of getting "exactly two heads."

(b) The event F of getting "at least two heads."

(c) The event G of getting "no heads."

SOLUTION By the results of Example 1 the sample space S of this experiment contains 8 outcomes.

(a) The event E of getting "exactly two heads" contains 3 outcomes, so by the definition of probability,

$$P(E) = \frac{n(E)}{n(S)} = \frac{3}{8}$$

(b) The event F of getting "at least two heads" has 4 outcomes, so

$$P(F) = \frac{n(F)}{n(S)} = \frac{4}{8} = \frac{1}{2}$$

(c) The event G of getting "no heads" has one outcome, so

$$P(G) = \frac{n(G)}{n(S)} = \frac{1}{8}$$

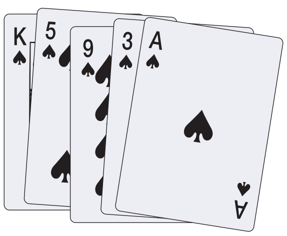
NOW TRY EXERCISE **7**

▼ Calculating Probability by Counting

To find the probability of an event, we do not need to list all the elements in the sample space and the event. We need only the *number* of elements in these sets. The counting techniques that we learned in the preceding sections will be very useful here.

EXAMPLE 3 | Finding the Probability of an Event

A five-card poker hand is drawn from a standard deck of 52 cards. What is the probability that all five cards are spades?



SOLUTION The experiment here consists of choosing five cards from the deck, and the sample space S consists of all possible five-card hands. Thus the number of elements in the sample space is

$$n(S) = C(52, 5) = \frac{52!}{5!(52-5)!} = 2{,}598{,}960$$

The event E that we are interested in consists of choosing five spades. Since the deck contains only 13 spades, the number of ways of choosing five spades is

$$n(E) = C(13, 5) = \frac{13!}{5!(13-5)!} = 1287$$

Thus the probability of drawing five spades is

$$P(E) = \frac{n(E)}{n(S)} = \frac{1287}{2{,}598{,}960} \approx 0.0005$$

NOW TRY EXERCISE **15** ■

What does the answer to Example 3 tell us? Because $0.0005 = \frac{1}{2000}$, we conclude that if you play poker many, many times, on average you will be dealt a hand consisting of only spades about once in every 2000 hands.

EXAMPLE 4 | Finding the Probability of an Event

A bag contains 20 tennis balls, of which four are defective. If two balls are selected at random from the bag, what is the probability that both are defective?

SOLUTION The experiment consists of choosing two balls from 20, so the number of elements in the sample space S is $C(20, 2)$. Since there are four defective balls, the number of ways of picking two defective balls is $C(4, 2)$. Thus the probability of the event E of picking two defective balls is

$$P(E) = \frac{n(E)}{n(S)} = \frac{C(4, 2)}{C(20, 2)} = \frac{6}{190} \approx 0.032$$

NOW TRY EXERCISE **17** ■

▼ The Complement of an Event

The **complement** of an event E is the set of outcomes in the sample space that is not in E. We denote the complement of E by E'.

By solving this equation for $P(E)$, we also have

$$P(E) = 1 - P(E')$$

PROBABILITY OF THE COMPLEMENT OF AN EVENT

Let S be the sample space of an experiment, and let E be an event. Then the probability of E', the complement of E, is

$$P(E') = 1 - P(E)$$

PROOF We calculate the probability of E' using the definition of probability and the fact that $n(E') = n(S) - n(E)$.

$$P(E') = \frac{n(E')}{n(S)} = \frac{n(S) - n(E)}{n(S)} = \frac{n(S)}{n(S)} - \frac{n(E)}{n(S)} = 1 - P(E)$$

This is a very useful result, since it is often difficult to calculate the probability of an event E but easy to find the probability of E'.

EXAMPLE 5 | Finding a Probability Using the Complement of an Event

An urn contains 10 red balls and 15 blue balls. Six balls are drawn at random from the urn. What is the probability that at least one ball is red?

SOLUTION Let E be the event that at least one red ball is drawn. It is tedious to count all the possible ways in which one or more of the balls drawn are red. So let's consider E', the complement of this event—namely, that none of the balls that are chosen is red. The number of ways of choosing 6 blue balls from the 15 blue balls is $C(15, 6)$; the number of ways of choosing 6 balls from the 25 balls is $C(25, 6)$. Thus

$$P(E') = \frac{n(E')}{n(S)} = \frac{C(15, 6)}{C(25, 6)} = \frac{5005}{177{,}100} = \frac{13}{460}$$

By the formula for the complement of an event we have

$$P(E) = 1 - P(E') = 1 - \frac{13}{460} \approx 0.97$$

NOW TRY EXERCISE **19**

The Union of Events

If E and F are events, what is the probability that E or F occurs? The word *or* indicates that we want the probability of the union of these events, that is, $E \cup F$.

PROBABILITY OF THE UNION OF EVENTS

If E and F are events in a sample space S, then the probability of E *or* F is

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

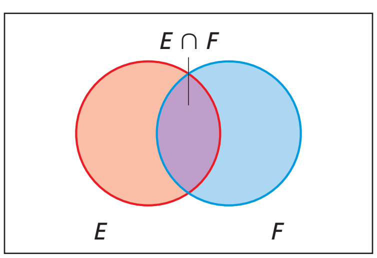


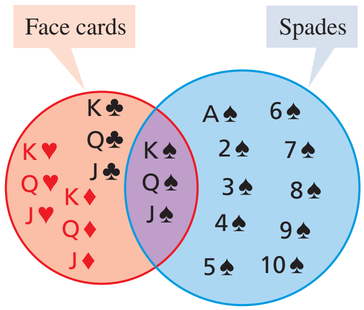
FIGURE 1

PROOF We need to find the number of elements in $E \cup F$. If we simply added the number of elements in E to the number of elements in F, we would be counting the elements in the overlap twice—once in E and once in F (see Figure 1). To get the correct total, we must subtract the number of elements in $E \cap F$. So $n(E \cup F) = n(E) + n(F) - n(E \cap F)$. Using the definition of probability we get

$$P(E \cup F) = \frac{n(E \cup F)}{n(S)} = \frac{n(E) + n(F) - n(E \cap F)}{n(S)} = P(E) + P(F) - P(E \cap F)$$

EXAMPLE 6 | Finding the Probability of the Union of Events

What is the probability that a card drawn at random from a standard 52-card deck is either a face card or a spade?



SOLUTION Let E denote the event "the card is a face card," and let F denote the event "the card is a spade." We want to find the probability of E or F, that is, $P(E \cup F)$.

There are 12 face cards and 13 spades in a 52-card deck, so

$$P(E) = \frac{12}{52} \qquad \text{and} \qquad P(F) = \frac{13}{52}$$

Since 3 cards are simultaneously face cards and spades, we have

$$P(E \cap F) = \frac{3}{52}$$

Now, by the formula for the probability of the union of two events we have

$$\begin{aligned} P(E \cup F) &= P(E) + P(F) - P(E \cap F) \\ &= \frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{11}{26} \end{aligned}$$

NOW TRY EXERCISE **21**

Two events that have no outcome in common are said to be **mutually exclusive** (see Figure 2). In other words, the events E and F are mutually exclusive if $E \cap F = \varnothing$. So if the events E and F are mutually exclusive, then $P(E \cap F) = 0$. The following result now follows from the formula for the union of two events.

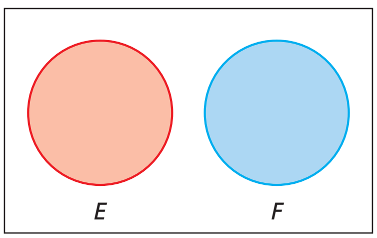


FIGURE 2

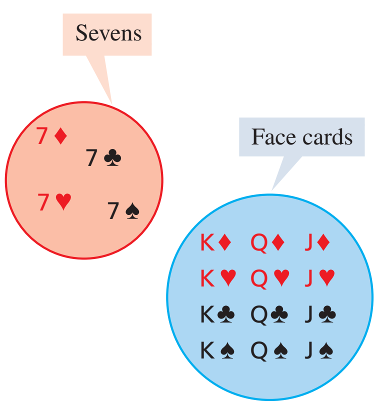
PROBABILITY OF THE UNION OF MUTUALLY EXCLUSIVE EVENTS

If E and F are mutually exclusive events, then

$$P(E \cup F) = P(E) + P(F)$$

EXAMPLE 7 | Finding the Probability of the Union of Mutually Exclusive Events

What is the probability that a card drawn at random from a standard 52-card deck is either a seven or a face card?



SOLUTION Let E denote the event "the card is a seven," and let F denote the event "the card is a face card." These events are mutually exclusive because a card cannot be at the same time a seven and a face card. By the above formula we have

$$P(E \cup F) = P(E) + P(F) = \frac{4}{52} + \frac{12}{52} = \frac{4}{13}$$

NOW TRY EXERCISES **23** AND **25**

Conditional Probability and the Intersection of Events

When we calculate probabilities, there sometimes is additional information that may alter the probability of an event. For example, suppose a person is chosen at random. What is the probability that the person has long hair? How does the probability change if we are given the additional information that the person chosen is a woman? In general, the probability of an event E given that another event F has occurred is expressed by writing

$$P(E \mid F) = \text{The probability of } E \text{ given } F$$

For example, suppose a die is rolled. Let E be the event of "getting a two" and let F be the event of "getting an even number." Then

$$P(E \mid F) = P(\text{The number is two given that the number is even})$$

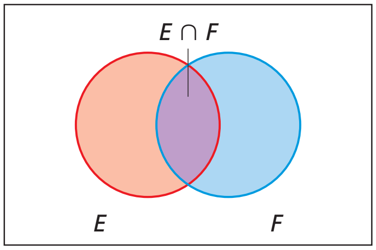


FIGURE 3

Since we know that the number is even, the possible outcomes are the three numbers 2, 4, and 6. So in this case the probability of a "two" is $P(E \mid F) = \frac{1}{3}$.

In general, if we know that F has occurred, then F serves as the sample space (see Figure 3). So $P(E \mid F)$ is determined by the number of outcomes in E that are also in F, that is, the number of outcomes in $E \cap F$.

CONDITIONAL PROBABILITY

Let E and F be events in a sample space S. The **conditional probability of E given that F occurs** is

$$P(E \mid F) = \frac{n(E \cap F)}{n(F)}$$

EXAMPLE 8 | Finding Conditional Probability

A mathematics class consists of 30 students; 12 of them study French, 8 study German, 3 study both of these languages, and the rest do not study a foreign language. If a student is chosen at random from this class, find the probability of each of the following events.

(a) The student studies French.

(b) The student studies French, given that he or she studies German.

(c) The student studies French, given that he or she studies a foreign language.

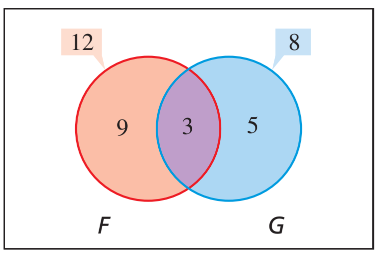


FIGURE 4

SOLUTION Let F denote the event "the student studies French," let G be the event "the student studies German," and let L be the event "the student studies a foreign language." It is helpful to organize the information in a Venn diagram, as in Figure 4.

(a) There are 30 students in the class, 12 of whom study French, so

$$P(F) = \frac{12}{30} = \frac{2}{5}$$

(b) We are asked to find $P(F \mid G)$, the probability that a student studies French given that the student studies German. Since eight students study German and three of these study French, it is clear that the required conditional probability is $\frac{3}{8}$. The formula for conditional probability confirms this:

$$P(F \mid G) = \frac{n(F \cap G)}{n(G)} = \frac{3}{8}$$

(c) From the Venn diagram in Figure 4 we see that the number of students who study a foreign language is $9 + 3 + 5 = 17$. Since 12 of these study French, we have

$$P(F \mid L) = \frac{n(F \cap L)}{n(L)} = \frac{12}{17}$$

NOW TRY EXERCISES **27** AND **29**

If we start with the expression for conditional probability and then divide numerator and denominator by $n(S)$, we get

$$P(E \mid F) = \frac{n(E \cap F)}{n(F)} = \frac{\dfrac{n(E \cap F)}{n(S)}}{\dfrac{n(F)}{n(S)}} = \frac{P(E \cap F)}{P(F)}$$

Multiplying both sides by $P(F)$ gives the following formula.

PROBABILITY OF THE INTERSECTION OF EVENTS

If E and F are events in a sample space S, then the probability of E *and* F is

$$P(E \cap F) = P(E)P(F \mid E)$$

EXAMPLE 9 | Finding the Probability of the Intersection of Events

Two cards are drawn, without replacement, from a 52-card deck. Find the probability of the following events.

(a) The first card drawn is an ace and the second is a king.

(b) The first card drawn is an ace and the second is also an ace.

SOLUTION Let E be the event "the first card is an ace," and let F be the event "the second card is a king."

(a) We are asked to find the probability of E *and* F, that is, $P(E \cap F)$. Now, $P(E) = \frac{4}{52}$. After an ace is drawn, 51 cards remain in the deck; of these, 4 are kings, so $P(F \mid E) = \frac{4}{51}$. By the above formula we have

$$P(E \cap F) = P(E)P(F \mid E) = \frac{4}{52} \times \frac{4}{51} \approx 0.0060$$

(b) Let E be the event "the first card is an ace," and let H be the event "the second card is an ace." The probability that the first card drawn is an ace is $P(E) = \frac{4}{52}$. After an ace is drawn, 51 cards remain; of these, 3 are aces, so $P(H \mid E) = \frac{3}{51}$. By the above formula we have

$$P(E \cap H) = P(E)P(H \mid E) = \frac{4}{52} \times \frac{3}{51} \approx 0.0045$$

NOW TRY EXERCISE **33** ■

When the occurrence of one event does not affect the probability of the occurrence of another event, we say that the events are **independent**. This means that the events E and F are independent if $P(E \mid F) = P(E)$ and $P(F \mid E) = P(F)$. For instance, if a fair coin is tossed, the probability of showing heads on the second toss is $\frac{1}{2}$, regardless of what was obtained on the first toss. So any two tosses of a coin are independent.

PROBABILITY OF THE INTERSECTION OF INDEPENDENT EVENTS

If E and F are independent events in a sample space S, then the probability of E *and* F is

$$P(E \cap F) = P(E)P(F)$$

EXAMPLE 10 | Finding the Probability of Independent Events

A jar contains five red balls and four black balls. A ball is drawn at random from the jar and then replaced; then another ball is picked. What is the probability that both balls are red?

SOLUTION Let E be the event "the first ball drawn is red," and let F be the event "the second ball drawn is red." Since we replace the first ball before drawing the second, the events E and F are independent. Now, the probability that the first ball is red is $\frac{5}{9}$. The probability that the second is red is also $\frac{5}{9}$. Thus the probability that both balls are red is

$$P(E \cap F) = P(E)P(F) = \frac{5}{9} \times \frac{5}{9} \approx 0.31$$

NOW TRY EXERCISE **37**

14.2 EXERCISES

CONCEPTS

1. The set of all possible outcomes of an experiment is called the ________ ________. A subset of the sample space is called an ________. The sample space for the experiment of tossing two coins is $S = \{HH, \_\_, \_\_, \_\_\}$. The event "getting at least one head" is $E = \{HH, \_\_, \_\_\}$. The probability of getting at least one head is

$$P(E) = \frac{n(\_\_)}{n(\_\_)} = \_\_\_.$$

2. Let E and F be events in a sample space S.
   (a) The probability of E *or* F occurring is $P(E \cup F) =$ ____________.
   (b) If the events E and F have no outcome in common (that is, the intersection of E and F is empty), then the events are called ________ ________. So in drawing a card from a deck, the event E, "getting a heart," and the event F, "getting a spade," are ________ ________.
   (c) If E and F are mutually exclusive, then the probability of E *or* F is $P(E \cup F) =$ ____________.

3. The conditional probability of E given that F occurs is $P(E \mid F) =$ ________. So in tossing a die the conditional probability of the event E, "getting a six," given that the event F, "getting an even number," has occurred is $P(E \mid F) =$ ________.

4. Let E and F be events in a sample space S.
   (a) The probability of E *and* F occurring is $P(E \cap F) =$ ____________.
   (b) If the occurrence of E does not affect the probability of the occurrence F, then the events are called ________. So in tossing a coin twice, the event E, "getting heads on the first toss," and the event F, "getting heads on the second toss," are ________.
   (c) If E and F are independent events, then the probability of E *and* F is $P(E \cap F) =$ ________.

SKILLS

5. An experiment consists of rolling a die. List the elements in the following sets.
   (a) The sample space
   (b) The event "getting an even number"
   (c) The event "getting a number greater than 4"

6. An experiment consists of tossing a coin and drawing a card from a deck.
   (a) How many elements does the sample space have?
   (b) List the elements in the event "getting heads and an ace."
   (c) List the elements in the event "getting tails and a face card."
   (d) List the elements in the event "getting heads and a spade."

Exercises 7–20 are about finding probability by counting.

7. An experiment consists of tossing a coin twice.
   (a) Find the sample space.
   (b) Find the probability of getting heads exactly two times.
   (c) Find the probability of getting heads at least one time.
   (d) Find the probability of getting heads exactly one time.

8. An experiment consists of tossing a coin and rolling a die.
   (a) Find the sample space.
   (b) Find the probability of getting heads and an even number.
   (c) Find the probability of getting heads and a number greater than 4.
   (d) Find the probability of getting tails and an odd number.

9–10 ■ A die is rolled. Find the probability of the given event.

9. **(a)** The number showing is a six.
   (b) The number showing is an even number.
   (c) The number showing is greater than five.

10. **(a)** The number showing is a two or a three.
   (b) The number showing is an odd number.
   (c) The number showing is a number divisible by 3.

11–12 ■ A card is drawn randomly from a standard 52-card deck. Find the probability of the given event.

11. **(a)** The card drawn is a king.
(b) The card drawn is a face card.
(c) The card drawn is not a face card.

12. **(a)** The card drawn is a heart.
(b) The card drawn is either a heart or a spade.
(c) The card drawn is a heart, a diamond, or a spade.

13–14 ■ A ball is drawn randomly from a jar that contains five red balls, two white balls, and one yellow ball. Find the probability of the given event.

13. **(a)** A red ball is drawn.
(b) The ball drawn is not yellow.
(c) A black ball is drawn.

14. **(a)** Neither a white nor yellow ball is drawn.
(b) A red, white, or yellow ball is drawn.
(c) The ball that is drawn is not white.

15. A poker hand, consisting of five cards, is dealt from a standard deck of 52 cards. Find the probability that the hand contains the cards described.
(a) Five hearts
(b) Five cards of the same suit
(c) Five face cards
(d) An ace, king, queen, jack, and a ten, all of the same suit (royal flush)

16. Three CDs are picked at random from a collection of 12 CDs of which four are defective. Find the probability of the following.
(a) All three CDs are defective.
(b) All three CDs are functioning properly.

17. Two balls are picked at random from a jar that contains three red and five white balls. Find the probability of the following events.
(a) Both balls are red.
(b) Both balls are white.

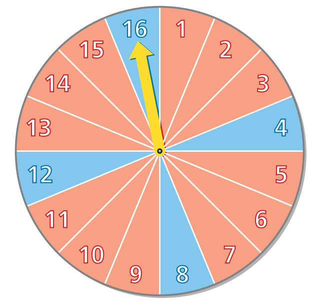
18. A letter is chosen at random from the word *EXTRATERRESTRIAL*. Find the probability of the given event.
(a) The letter *T* is chosen.
(b) The letter chosen is a vowel.
(c) The letter chosen is a consonant.

19. A five-card poker hand is drawn from a standard 52-card deck. Find the probability of the following events.
(a) At least one card is a spade.
(b) At least one card is a face card.

20. A pair of dice is rolled, and the numbers showing are observed.
(a) List the sample space of this experiment.
(b) Find the probability of getting a sum of 7.
(c) Find the probability of getting a sum of 9.
(d) Find the probability that the two dice show doubles (the same number).
(e) Find the probability that the two dice show different numbers.
(f) Find the probability of getting a sum of 9 or higher.

Exercises 21–26 are about the probability of the union of events.

21–22 ■ Refer to the spinner shown in the figure. Find the probability of the given event.



21. **(a)** The spinner stops on red.
(b) The spinner stops on an even number.
(c) The spinner stops on red or an even number.

22. **(a)** The spinner stops on blue.
(b) The spinner stops on an odd number.
(c) The spinner stops on blue or an odd number.

23–24 ■ A die is rolled, and the number showing is observed. Determine whether the events E and F are mutually exclusive. Then find the probability of the event $E \cup F$.

23. **(a)** E: The number is even.
F: The number is odd.
(b) E: The number is even.
F: The number is greater than 4.

24. **(a)** E: The number is greater than 3.
F: The number is less than 5.
(b) E: The number is divisible by 3.
F: The number is less than 3.

25–26 ■ A card is drawn at random from a standard 52-card deck. Determine whether the events E and F are mutually exclusive. Then find the probability of the event $E \cup F$.

25. **(a)** E: The card is a face card.
F: The card is a spade.
(b) E: The card is a heart.
F: The card is a spade.

26. **(a)** E: The card is a club.
F: The card is a king.
(b) E: The card is an ace.
F: The card is a spade.

Exercises 27–32 are about conditional probability.

27. A die is rolled. Find the given conditional probability.
(a) A "five" shows, given that the number showing is greater than 3.
(b) A "three" shows, given that the number showing is odd.

28. A card is drawn from a deck. Find the following conditional probability.
(a) The card is a queen, given that it is a face card.
(b) The card is a king, given that it is a spade.
(c) The card is a spade, given that it is a king.

29–30 ■ Refer to the spinner in Exercises 21–22.

29. Find the probability that the spinner has stopped on an even number, given that it has stopped on red.

30. Find the probability that the spinner has stopped on a number divisible by 3, given that it has stopped on blue.

31–32 ■ A jar contains five red balls numbered 1 to 5, and seven green balls numbered 1 to 7.

31. A ball is drawn at random from the jar. Find the following conditional probabilities.
(a) The ball is red, given that it is numbered 3.
(b) The ball is green, given that is numbered 7.
(c) The ball is red, given that it has an even number.
(d) The ball has an even number, given that it is red.

32. Two balls are drawn at random from the jar. Find the following conditional probabilities.
(a) The second ball drawn is red, given that the first is red.
(b) The second ball drawn is red, given that the first is green.
(c) The second ball drawn is even-numbered, given that the first is odd-numbered.
(d) The second ball drawn is even-numbered, given that the first is even-numbered.

Exercises 33–40 are about the probability of the intersection of events.

33. A jar contains seven black balls and three white balls. Two balls are drawn, without replacement, from the jar. Find the probability of the following events.
(a) The first ball drawn is black, and the second is white.
(b) The first ball drawn is black, and the second is black.

34. A drawer contains an unorganized collection of 18 socks. Three pairs are red, two pairs are white, and four pairs are black.
(a) If one sock is drawn at random from the drawer, what is the probability that it is red?
(b) Once a sock is drawn and discovered to be red, what is the probability of drawing another red sock to make a matching pair?
(c) If two socks are drawn from the drawer at the same time, what is the probability that both are red?

35. Two cards are drawn from a deck without replacement. Find the probability of the following events.
(a) The first is an ace and the second a king?
(b) Both cards are aces?

36. A die is rolled twice. Let E and F be the following events:

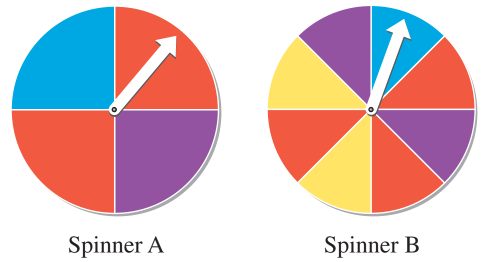
E: The first roll shows a "six."
F: The second roll shows a "six."

(a) Are the events E and F independent?
(b) Find the probability of showing a "six" on both rolls.

37. A die is rolled twice. What is the probability of getting a "one" on the first roll and an even number on the second roll?

38. A coin is tossed and a die is rolled.
(a) Are the events "tails" and "even number" independent?
(b) Find the probability of getting a tail and an even number.

39–40 ■ Spinners A and B shown in the figure are spun at the same time.



39. (a) Are the events "spinner A stops on red" and "spinner B stops on yellow" independent?
(b) Find the probability that spinner A stops on red and spinner B stops on yellow

40. (a) Find the probability that both spinners stop on purple.
(b) Find the probability that both spinners stop on blue.

APPLICATIONS

41. Four Siblings A couple intends to have four children. Assume that having a boy and having a girl are equally likely events.
(a) List the sample space of this experiment.
(b) Find the probability that the couple will have only boys.
(c) Find the probability that the couple will have two boys and two girls.
(d) Find the probability that the couple will have four children of the same gender
(e) Find the probability that the couple will have at least two girls.

42. Bridge Hands What is the probability that a 13-card bridge hand consists of all cards from the same suit?

43. Roulette An American roulette wheel has 38 slots; two slots are numbered 0 and 00, and the remaining slots are numbered from 1 to 36. Find the probability that the ball lands in an odd-numbered slot.

44. Making Words A toddler has wooden blocks showing the letters *C, E, F, H, N,* and *R.* Find the probability that the child arranges the letters in the indicated order.
(a) In the order *FRENCH*
(b) In alphabetical order

45. Lottery In the 6/49 lottery game, a player selects six numbers from 1 to 49. What is the probability of picking the six winning numbers?

46. An Unlikely Event The president of a large company selects six employees to receive a special bonus. He claims that the six employees are chosen randomly from among the 30 employees, of whom 19 are women and 11 are men. What is the probability that no woman is chosen?

47. Guessing on a Test An exam has ten true-false questions. A student who has not studied answers all ten questions by just guessing. Find the probability that the student correctly answers the given number of questions.
(a) All ten questions
(b) Exactly seven questions

48. Quality Control To control the quality of their product, the Bright-Light Company inspects three light bulbs out of each batch of ten bulbs manufactured. If a defective bulb is found, the batch is discarded. Suppose a batch contains two defective bulbs. What is the probability that the batch will be discarded?

49. Monkeys Typing Shakespeare An often-quoted example of an event of extremely low probability is that a monkey types Shakespeare's entire play *Hamlet* by randomly striking keys on a typewriter. Assume that the typewriter has 48 keys (including the space bar) and that the monkey is equally likely to hit any key.

(a) Find the probability that such a monkey will actually correctly type just the title of the play as his first word.

(b) What is the probability that the monkey will type the phrase "To be or not to be" as his first words?

50. Making Words A monkey is trained to arrange wooden blocks in a straight line. He is then given six blocks showing the letters *A, E, H, L, M, T.*

(a) What is the probability that he will arrange them to spell the word *HAMLET*?

(b) What is the probability that he will arrange them to spell the word *HAMLET* three consecutive times?

51. Making Words A toddler has eight wooden blocks showing the letters *A, E, I, G, L, N, T,* and *R.* What is the probability that the child will arrange the letters to spell one of the words *TRIANGLE* or *INTEGRAL*?

52. Horse Race Eight horses are entered in a race. You randomly predict a particular order for the horses to complete the race. What is the probability that your prediction is correct?

53. Genetics Many genetic traits are controlled by two genes, one dominant and one recessive. In Gregor Mendel's original experiments with peas, the genes controlling the height of the plant are denoted by T (tall) and t (short). The gene T is dominant, so a plant with the genotype (genetic makeup) TT or Tt is tall, whereas one with genotype tt is short. By a statistical analysis of the offspring in his experiments, Mendel concluded that offspring inherit one gene from each parent and that each possible combination of the two genes is equally likely. If each parent has the genotype Tt, then the following chart gives the possible genotypes of the offspring:

		Parent 2	
		T	**t**
Parent 1	**T**	TT	Tt
	t	Tt	tt

Find the probability that a given offspring of these parents will be

(a) tall **(b)** short

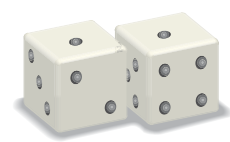
54. Genetics Refer to Exercise 53. Make a chart of the possible genotypes of the offspring if one parent has genotype Tt and the other has tt. Find the probability that a given offspring will be

(a) tall **(b)** short

55. Roulette An American roulette wheel has 38 slots. Two of the slots are numbered 0 and 00, and the rest are numbered from 1 to 36. A player places a bet on a number between 1 and 36 and wins if a ball thrown into the spinning roulette wheel lands in the slot with the same number. Find the probability of winning on two consecutive spins of the roulette wheel.

56. Choosing a Committee A committee of five is chosen randomly from a group of six males and eight females. What is the probability that the committee includes either all males or all females?

57. Snake Eyes What is the probability of rolling snake eyes ("double ones") three times in a row?



58. Lottery In the 6/49 lottery game a player selects six numbers from 1 to 49. What is the probability of selecting at least five of the six winning numbers?

59. Marbles in a Jar A jar contains six red marbles numbered 1 to 6 and ten blue marbles numbered 1 to 10. A marble is drawn at random from the jar. Find the probability that the given event occurs.

(a) The marble is red.

(b) The marble is odd-numbered.

(c) The marble is red or odd-numbered.

(d) The marble is blue or even-numbered.

60. Lottery In the 6/49 lottery game, a player selects six numbers from 1 to 49 and wins if he or she selects the winning six numbers. What is the probability of winning the lottery two times in a row?

61. Balls in a Jar Jar A contains three red balls and four white balls. Jar B contains five red balls and two white balls. Which one of the following ways of randomly selecting balls gives the greatest probability of drawing two red balls?

(i) Draw two balls from jar B.

(ii) Draw one ball from each jar.

(iii) Put all the balls in one jar, and then draw two balls.

62. Slot Machine A slot machine has three wheels. Each wheel has 11 positions: a bar and the digits 0, 1, 2, . . . , 9. When the handle is pulled, the three wheels spin independently before coming to rest. Find the probability that the wheels stop on the following positions.

(a) Three bars

(b) The same number on each wheel
(c) At least one bar.

63. Combination Lock A student has locked her locker with a combination lock, showing numbers from 1 to 40, but she has forgotten the three-number combination that opens the lock. She remembers that all three numbers in the combination are different. To open the lock, she decides to try all possible combinations. If she can try ten different combinations every minute, what is the probability that she will open the lock within one hour?

64. Committee Membership A mathematics department consists of ten men and eight women. Six mathematics faculty members are to be selected at random for the curriculum committee.

(a) What is the probability that two women and four men are selected?
(b) What is the probability that two or fewer women are selected?
(c) What is the probability that more than two women are selected?

65. Class Photo Twenty students are arranged randomly in a row for a class picture. Paul wants to stand next to Phyllis. Find the probability that he gets his wish.

DISCOVERY ■ DISCUSSION ■ WRITING

66. Oldest Son A family with two children is randomly selected. Assume that the events of having a boy or a girl are equally likely. Find the following probabilities.

(a) The family has two boys given that the oldest child is a boy.
(b) The family has two boys given that one of the children is a boy.

DISCOVERY PROJECT **Small Samples, Big Results**

In this project we perform several experiments that show how we can obtain information about a big population from a small sample. You can find the project at the book companion website: **www.stewartmath.com**

14.3 BINOMIAL PROBABILITY

Binomial Probability ▶ The Binomial Distribution

In this section we study a special kind of probability that plays a crucial role in modeling many real-world situations.

▼ Binomial Probability

Calculating the probability of independent events is studied on page 907.

A coin is weighted so that the probability of heads is 0.6. What is the probability of getting exactly two heads in five tosses of this coin? Since the tosses are independent, the probability of getting two heads followed by three tails is

$$\underbrace{0.6}_{\text{Heads}} \times \underbrace{0.6}_{\text{Heads}} \times \underbrace{0.4}_{\text{Tails}} \times \underbrace{0.4}_{\text{Tails}} \times \underbrace{0.4}_{\text{Tails}} = (0.6)^2(0.4)^3$$

But this is not the only way we can get exactly two heads. The two heads can occur, for example, on the second toss and the last toss. In this case the probability is

$$\underbrace{0.4}_{\text{Tails}} \times \underbrace{0.6}_{\text{Heads}} \times \underbrace{0.4}_{\text{Tails}} \times \underbrace{0.4}_{\text{Tails}} \times \underbrace{0.6}_{\text{Heads}} = (0.6)^2(0.4)^3$$

In fact, the two heads could occur on any two of the five tosses. Thus there are $C(5, 2)$ ways in which this can happen, each with probability $(0.6)^2(0.4)^3$. It follows that

$$P(\text{exactly 2 heads in 5 tosses}) = C(5, 2)(0.6)^2(0.4)^3 \approx 0.023$$

The probability that we have just calculated is an example of a binomial probability. In general, a **binomial experiment** is one in which there are two outcomes, which are called "success" and "failure." In the coin-tossing experiment described above, "success" is getting "heads," and "failure" is getting "tails." The following box tells us how to calculate the probabilities associated with binomial experiments when we perform them many times.

The name "binomial probability" is appropriate because $C(n, r)$ is the same as the binomial coefficient $\binom{n}{r}$ (see Exercise 86, page 900).

BINOMIAL PROBABILITY

An experiment has two possible outcomes called "success" and "failure," with $P(\text{success}) = p$ and $P(\text{failure}) = 1 - p$. The probability of getting exactly r successes in n independent trials of the experiment is

$$P(r \text{ successes in } n \text{ trials}) = C(n, r)p^r(1 - p)^{n-r}$$

EXAMPLE 1 | Binomial Probability

A fair die is rolled 10 times. Find the probability of each event.

(a) Exactly 2 sixes.

(b) At most 1 six.

(c) At least 2 sixes.

SOLUTION Let's call "getting a six" success and "not getting a six" failure. So $P(\text{success}) = \frac{1}{6}$ and $P(\text{failure}) = \frac{5}{6}$. Since each roll of the die is independent of the other rolls, we can use the formula for binomial probability with $n = 10$ and $p = \frac{1}{6}$.

(a) $P(\text{exactly 2 sixes}) = C(10, 2)(\frac{1}{6})^2(\frac{5}{6})^8 \approx 0.29$

(b) The statement "at most 1 six" means 0 sixes or 1 six. So

$P(\text{at most one six})$

$$= P(\text{0 sixes or 1 six}) \qquad \text{Meaning of "at most"}$$

$$= P(\text{0 sixes}) + P(\text{1 six}) \qquad P(A \text{ or } B) = P(A) + P(B)$$

$$= C(10, 0)\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{10} + C(10, 1)\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^9 \qquad \text{Binomial probability}$$

$$\approx 0.1615 + 0.3230 \qquad \text{Calculator}$$

$$\approx 0.4845 \qquad \text{Calculator}$$

(c) The statement "at least two sixes" means two or more sixes. Instead of adding the probabilities of getting 2, 3, 4, 5, 6, 7, 8, 9, or 10 sixes (which is a lot of work), it's easier to find the probability of the complement of this event. The complement of the event "two or more sixes" is "0 or 1 six." So

$$P(\text{two or more sixes}) = 1 - P(\text{0 or 1 six}) \qquad P(E) = 1 - P(E')$$

$$= 1 - 0.4845 \qquad \text{From part (b)}$$

$$= 0.5155 \qquad \text{Calculator}$$

NOW TRY EXERCISES **3** AND **21** ■

EXAMPLE 2 | Binomial Probability

Patients infected with a certain virus have a 40% chance of surviving. There are 10 patients in a hospital who have acquired this virus. Find the probability that 7 or more of the patients survive.

SOLUTION Let's call the event "patient survives" success and the event "patient dies" failure. We are given that the probability of success is $p = 0.4$, so the probability of failure is $1 - p = 1 - 0.4 = 0.6$. We need to calculate the probability of 7, 8, 9, or 10 successes in 10 trials.

$$P(7 \text{ out of } 10 \text{ recover}) = C(10, 7)(0.4)^7(0.6)^3 \approx 0.04247$$
$$P(8 \text{ out of } 10 \text{ recover}) = C(10, 8)(0.4)^8(0.6)^2 \approx 0.01062$$
$$P(9 \text{ out of } 10 \text{ recover}) = C(10, 9)(0.4)^9(0.6)^1 \approx 0.00157$$
$$P(10 \text{ out of } 10 \text{ recover}) = C(10, 10)(0.4)^{10}(0.6)^0 \approx 0.00010$$

Adding the probabilities, we find that

$$P(7 \text{ or more recover}) \approx 0.05476$$

There is about a 1 in 20 chance that 7 or more patients recover.

NOW TRY EXERCISE **35**

The Binomial Distribution

We can describe how the probabilities of an experiment are "distributed" among all the outcomes of an experiment by making a table of values. The function that assigns to each outcome its corresponding probability is called a **probability distribution**. A bar graph of a probability distribution in which the width of each bar is 1 is called a **probability histogram**. The next example illustrates these concepts.

EXAMPLE 3 | Probability Distributions

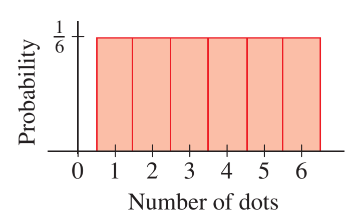
Make a table of the probability distribution for the experiment of rolling a fair die and observing the number of dots. Draw a histogram of the distribution.

SOLUTION When rolling a fair die each face has probability $1/6$ of showing. The probability distribution is shown in the following table. To draw a histogram, we draw bars of width 1 and height $\frac{1}{6}$ corresponding to each outcome.

Probability Distribution

Outcome (dots)	Probability
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Probability Histogram



NOW TRY EXERCISE **15**

A probability distribution in which all outcomes have the same probability is called a **uniform distribution**. The rolling-a-die experiment in Example 3 is a uniform distribution. The probability distribution of a binomial experiment is called a **binomial distribution**.

EXAMPLE 4 | A Binomial Distribution

A fair coin is tossed eight times, and the number of heads is observed. Make a table of the probability distribution, and draw a histogram. What is the number of heads that is most likely to show up?

SOLUTION This is a binomial experiment with $n = 8$ and $p = \frac{1}{2}$, so $1 - p = \frac{1}{2}$ as well. We need to calculate the probability of getting 0 heads, 1 head, 2 heads, 3 heads, and so on. For example, to calculate the probability of 3 heads we have

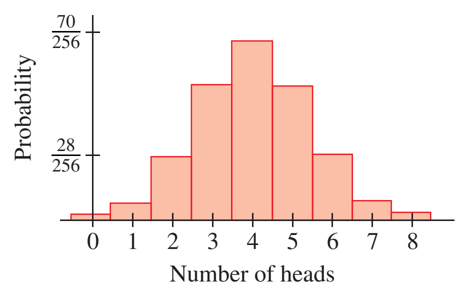
$$P(3\text{ heads}) = C(8, 3)\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5 = \frac{28}{256}$$

The other entries in the following table are calculated similarly. We draw the histogram by making a bar for each outcome with width 1 and height equal to the corresponding probability.

Probability Distribution

Outcome (heads)	Probability
0	$\frac{1}{256}$
1	$\frac{8}{256}$
2	$\frac{28}{256}$
3	$\frac{56}{256}$
4	$\frac{70}{256}$
5	$\frac{56}{256}$
6	$\frac{28}{256}$
7	$\frac{8}{256}$
8	$\frac{1}{256}$

Probability Histogram



NOW TRY EXERCISE **17**

Notice that the sum of the probabilities in a probability distribution is 1, because the sum is the probability of the occurrence of *any* outcome in the sample space (this is the certain event).

14.3 EXERCISES

CONCEPTS

1. A binomial experiment is one in which there are exactly ________ outcomes. One outcome is called ________, and the other is called ________.

2. If a binomial experiment has probability p of success, then the probability of failure is ________. The probability of getting exactly r successes in n trials of this experiment is $C(\_\_\_\_, \_\_\_\_)p^{\square}(1 - p)^{\square}$.

SKILLS

3–14 ■ Five independent trials of a binomial experiment with probability of success $p = 0.7$ are performed. Find the probability of each event.

3. Exactly two successes

4. Exactly three successes

5. No successes

6. All successes

7. Exactly one success

8. Exactly one failure

9. At least four successes

10. At least three successes

11. At most one failure

12. At most two failures

13. At least two successes

14. At most three failures

15–16 ■ An experiment is described. **(a)** Complete the table of the probability distribution. **(b)** Draw a probability histogram.

15. A jar contains five balls numbered 1 to 5. A ball is drawn at random, and the number of the ball is observed.

Outcome	Probability
1	0.2
2	
3	
4	
5	

16. A jar contains five balls numbered 1, three balls numbered 2, one ball numbered 3, and one ball numbered 4. A ball is drawn at random and the number of the ball is observed.

Outcome	Probability
1	0.5
2	
3	
4	
5	

17–20 ■ A binomial experiment with probability of success p is performed n times. **(a)** Make a table of the probability distribution. **(b)** Draw a probability histogram.

17. $n = 4, \quad p = 0.5$

18. $n = 5, \quad p = 0.4$

19. $n = 7, \quad p = 0.2$

20. $n = 6, \quad p = 0.9$

APPLICATIONS

21. Rolling Dice Six dice are rolled. Find the probability that two of them show a four.

22. Archery An archer hits his target 80% of the time. If he shoots seven arrows, what is the probability of each event?
(a) He never hits the target.
(b) He hits the target each time.
(c) He hits the target more than once.
(d) He hits the target at least five times.

23. Television Ratings According to a ratings survey, 40% of the households in a certain city tune in to the local evening TV news. If ten households are visited at random, what is the probability that four of them will have their television tuned to the local news?

24. Spread of Disease Health authorities estimate that 10% of the raccoons in a certain rural county are carriers of rabies. A dog is bitten by four different raccoons in this county. What is the probability that he was bitten by at least one rabies carrier?

25. Blood Type About 45% of the populations of the United States and Canada have Type O blood.
(a) If a random sample of ten people is selected, what is the probability that exactly five have Type O blood?
(b) What is the probability that at least three of the random sample of ten have Type O blood?

26. Handedness A psychologist needs 12 left-handed subjects for an experiment, and she interviews 15 potential subjects. About 10% of the population is left-handed.
(a) What is the probability that exactly 12 of the potential subjects are left-handed?
(b) What is the probability that 12 or more are left-handed?

27. Germination Rates A certain brand of tomato seeds has a 0.75 probability of germinating. To increase the chance that at least one tomato plant per seed hill germinates, a gardener plants four seeds in each hill.
(a) What is the probability that at least one seed germinates in a given hill?
(b) What is the probability that two or more seeds will germinate in a given hill?
(c) What is the probability that all four seeds germinate in a given hill?

28. Genders of Children Assume that for any given live human birth, the chances that the child is a boy or a girl are equally likely.
(a) What is the probability that in a family of five children a majority are boys?
(b) What is the probability that in a family of seven children a majority are girls?

29. Genders of Children The ratio of male to female births is in fact not exactly one to one. The probability that a newborn turns out to be a male is about 0.52. A family has ten children.
(a) What is the probability that all ten children are boys?
(b) What is the probability all are girls?
(c) What is the probability that five are girls and five are boys?

30. Education Level In a certain county 20% of the population has a college degree. A jury consisting of 12 people is selected at random from this county.
(a) What is the probability that exactly two of the jurors have a college degree?
(b) What is the probability that three or more of the jurors have a college degree?

31. Defective Light Bulbs The DimBulb Lighting Company manufactures light bulbs for appliances such as ovens and refrigerators. Typically, 0.5% of their bulbs are defective.

From a crate with 100 bulbs, three are tested. Find the probability that the given event occurs.
(a) All three bulbs are defective.
(b) One or more bulbs is defective.

32. **Quality Control** An assembly line that manufactures fuses for automotive use is checked every hour to ensure the quality of the finished product. Ten fuses are selected randomly, and if any one of the ten is found to be defective, the process is halted and the machines are recalibrated. Suppose that at a certain time 5% of the fuses being produced are actually defective. What is the probability that the assembly line is halted at that hour's quality check?

33. **Sick Leave** The probability that a given worker at Dyno Nutrition will call in sick on a Monday is 0.04. The packaging department has eight workers. What is the probability that two or more packaging workers will call in sick next Monday?

34. **Political Surveys** In a certain county, 60% of the voters are in favor of an upcoming school bond initiative. If five voters are interviewed at random, what is the probability that exactly three of them will favor the initiative?

35. **Pharmaceuticals** A drug that is used to prevent motion sickness is found to be effective about 75% of the time. Six friends, all prone to seasickness, go on a sailing cruise, and all take the drug. Find the probability of each event.
(a) None of the friends gets seasick.
(b) All of the friends get seasick.
(c) Exactly three get seasick.
(d) At least two get seasick.

36. **Reliability of a Machine** A machine that is used in a manufacturing process has four separate components, each of which has a 0.01 probability of failing on any given day. If any component fails, the entire machine breaks down. Find the probability that on a given day the indicated event occurs.
(a) The machine breaks down.
(b) The machine does not break down.
(c) Only one component does not fail.

37. **Genetics** Huntington's disease is a hereditary ailment caused by a recessive gene. If both parents carry the gene but do not have the disease, there is a 0.25 probability that an offspring will fall victim to the condition. A newlywed couple find through genetic testing that they both carry the gene (but do not have the disease). If they intend to have four children, find the probability of each event.
(a) At least one child gets the disease.
(b) At least three of the children get the disease.

38. **Selecting Cards** Three cards are randomly selected from a standard 52-card deck, one at a time, with each card replaced in the deck before the next one is picked. Find the probability of each event.
(a) All three cards are hearts.
(b) Exactly two of the cards are spades.
(c) None of the cards is a diamond.
(d) At least one of the cards is a club.

39. **Smokers and Nonsmokers** The participants at a mathematics conference are housed dormitory-style, five to a room. Because of an oversight, conference organizers forget to ask whether the participants are smokers. In fact, it turns out that 30% are smokers. Find the probability that Fred, a nonsmoking conference participant, will be housed with:
(a) Exactly one smoker.
(b) One or more smokers.

40. **Telephone Marketing** A mortgage company advertises its rates by making unsolicited telephone calls to random numbers. About 2% of the calls reach consumers who are interested in the company's services. A telephone consultant can make 100 calls per evening shift.
(a) What is the probability that two or more calls will reach an interested party in one shift?
(b) How many calls does a consultant need to make to ensure at least a 0.5 probability of reaching one or more interested parties? [*Hint:* Use trial and error.]

41. **Effectiveness of a Drug** A certain disease has a mortality rate of 60%. A new drug is tested for its effectiveness against this disease. Ten patients are given the drug, and eight of them recover.
(a) Find the probability that eight or more of the patients would have recovered without the drug.
(b) Does the drug appear to be effective? (Consider the drug effective if the probability in part (a) is 0.05 or less.)

42. **Hitting a Target** An archer normally hits the target with probability of 0.6. She hires a new coach for a series of special lessons. After the lessons she hits the target in five out of eight attempts.
(a) Find the probability that she would have hit five or more out of the eight attempts before her lessons with the new coach.
(b) Did the new coaching appear to make a difference? (Consider the coaching effective if the probability in part (a) is 0.05 or less.)

DISCOVERY ■ DISCUSSION ■ WRITING

43. **Most Likely Outcome for *n* Tosses of a Coin** A balanced coin is tossed n times. In this exercise we investigate the following question: What is the number of heads that has the greatest probability of occurring? Note that for a balanced coin the probability of heads is $p = 0.5$.

(a) Suppose $n = 8$. Draw a probability histogram for the resulting binomial distribution. What number of heads has the greatest probability of occurring? If $n = 100$, what number of heads has the greatest probability of occurring?

(b) Suppose $n = 9$. Draw a probability histogram for the resulting binomial distribution. What number of heads has the greatest probability of occurring? If $n = 101$, what number of heads has the greatest probability of occurring?

14.4 Expected Value

| Expected Value ▶ What Is a Fair Game?

In this section we study an important application of probability called *expected value.*

▼ Expected Value

Suppose that a coin has probability 0.8 of showing heads. If the coin is tossed many times, we would *expect* to get heads about 80% of the time. Now, suppose that you get a payout of one dollar for each head. If you play this game many times, you would expect *on average* to gain \$0.80 per game:

$$\begin{pmatrix}\text{Expected payout}\\ \text{per game}\end{pmatrix} = \begin{pmatrix}\text{Amount of payout}\\ \text{per game}\end{pmatrix} \times \begin{pmatrix}\text{Probability of payout}\\ \text{per game}\end{pmatrix}$$

$$= \$1.00 \times 0.80 = \$0.80$$

The reasoning in this example motivates the following definition.

DEFINITION OF EXPECTED VALUE

A game gives payouts $a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, \ldots, p_n$. The **expected value** (or **expectation**) E of this game is

$$E = a_1 p_1 + a_2 p_2 + \cdots + a_n p_n$$

EXAMPLE 1 | Finding Expected Value

A die is rolled, and you receive \$1 for each point that shows. What is your expectation?

SOLUTION Each face of the die has probability $\frac{1}{6}$ of showing. So you get \$1 with probability $\frac{1}{6}$, \$2 with probability $\frac{1}{6}$, \$3 with probability $\frac{1}{6}$, and so on. Thus the expected value is

$$E = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

So if you play this game many times, you will make, on average, \$3.50 per game.

NOW TRY EXERCISE 3 ■

EXAMPLE 2 | Finding Expected Value

In Monte Carlo the game of roulette is played on a wheel with slots numbered 0, 1, 2, . . . , 36. The wheel is spun, and a ball dropped in the wheel is equally likely to end up in any

one of the slots. To play the game, you bet \$1 on any number. (For example, you may bet \$1 on number 23.) If the ball stops in your slot, you get \$36 (the \$1 you bet plus \$35). Find the expected value of this game.

SOLUTION The gambler can gain \$35 with probability $\frac{1}{37}$ and can lose \$1 with probability $\frac{36}{37}$. So the gambler's expected value is

$$E = (35)\frac{1}{37} + (-1)\frac{36}{37} \approx -0.027$$

In other words, if you play this game many times, you would expect to lose 2.7 cents on every dollar you bet (on average). Consequently, the house expects to gain 2.7 cents on every dollar that is bet.

NOW TRY EXERCISE **13** ■

EXAMPLE 3 | Expected Number

At any given time, the express checkout lane at a small supermarket has three shoppers in line with probability 0.2, two shoppers with probability 0.5, one shopper with probability 0.2, and no shoppers with probability 0.1. If you go to this market, how many shoppers would you expect to find waiting in the express checkout lane?

SOLUTION The "payouts" here are the number of shoppers waiting in line. To find the expected number of shoppers waiting in line, we multiply each "payout" by its probability and add the results.

$$E = 3(0.2) + 2(0.5) + 1(0.2) + 0(0.1) = 1.8$$

So on average, you would expect 1.8 shoppers waiting in the express lane.

NOW TRY EXERCISE **21** ■

What Is a Fair Game?

A **fair game** is game with expected value zero. So if you play a fair game many times, you would expect, on average, to break even.

EXAMPLE 4 | A Fair Game?

Suppose that you play the following game. A card is drawn from a deck. If the card is an ace, you get a payout of \$10. If the card is not an ace, you have to pay \$1.

(a) Is this a fair game?

(b) If the game is not fair, find the payout amount that would make this game a fair game.

MATHEMATICS IN THE MODERN WORLD

Fair Voting Methods

The methods of mathematics have recently been applied to problems in the social sciences. For example, how do we find fair voting methods? You may ask, "What is the problem with how we vote in elections?" Well, suppose candidates A, B, and C are running for president. The final vote tally is as follows: A gets 40%, B gets 39%, and C gets 21%. So candidate A wins. But 60% of the voters *didn't* want A. Moreover, suppose you voted for C, but you dislike A so much that you would have been willing to change your vote to B to avoid having A win. Suppose most of the voters who voted for C feel the same way you do. Then we have a situation in which most of the voters prefer B over A, but A wins. Is that fair?

In the 1950s Kenneth Arrow showed mathematically that no democratic method of voting can be completely fair; he later won a Nobel Prize for his work. Mathematicians continue to work on finding fairer voting systems. The system that is most often used in federal, state, and local elections is called *plurality voting* (the candidate with the most votes wins). Other systems include *majority voting* (if no candidate gets a majority, a runoff is held between the top two vote-getters), *approval voting* (each voter can vote for as many candidates as he or she approves of), *preference voting* (each voter orders the candidates according to his or her preference), and *cumulative voting* (each voter gets as many votes as there are candidates and can give all of his or her votes to one candidate or distribute them among the candidates as he or she sees fit). This last system is often used to select corporate boards of directors. Each system of voting has both advantages and disadvantages.

SOLUTION

(a) In this game you get a payout of \$10 if an ace is drawn (probability $\frac{4}{52}$), and you lose \$1 if any other card is drawn (probability $\frac{48}{52}$). So the expected value is

$$E = 10\left(\frac{4}{52}\right) - 1\left(\frac{48}{52}\right) = -\frac{8}{52}$$

Since the expected value is not zero, the game is not fair. If you play this game many times, you would expect to lose, on average, $\frac{8}{52} \approx \$0.15$ per game.

(b) We want to find the payout x that makes the expected value 0.

$$E = x\left(\frac{4}{52}\right) - 1\left(\frac{48}{52}\right) = 0$$

Solving this equation we get $x = 12$. So a payout of \$12 for an ace would make this a fair game.

NOW TRY EXERCISE **25**

Games of chance in casinos are never fair; the gambler always has a negative expected value (as in Examples 2 and 4(a)). This makes gambling profitable for the casino and unprofitable for the gambler.

14.4 EXERCISES

CONCEPTS

1. If a game gives payoffs of \$10 and \$100 with probabilities 0.9 and 0.1, respectively, then the expected value of this game is

$$E = \_\_\_\_\_ \times 0.9 + \_\_\_\_\_ \times 0.1 = \_\_\_\_\_\_$$

2. If you played the game in Exercise 1 many times, then you would expect your average payoff per game to be about

\$__________.

SKILLS

3–12 ■ Find the expected value (or expectation) of the games described.

3. Mike wins \$2 if a coin toss shows heads and \$1 if it shows tails.

4. Jane wins \$10 if a die roll shows a six, and she loses \$1 otherwise.

5. The game consists of drawing a card from a deck. You win \$100 if you draw the ace of spades or lose \$1 if you draw any other card.

6. Tim wins \$3 if a coin toss shows heads or \$2 if it shows tails.

7. Carol wins \$3 if a die roll shows a six, and she wins \$0.50 otherwise.

8. A coin is tossed twice. Albert wins \$2 for each heads and must pay \$1 for each tails.

9. A die is rolled. Tom wins \$2 if the die shows an even number, and he pays \$2 otherwise.

10. A card is drawn from a deck. You win \$104 if the card is an ace, \$26 if it is a face card, and \$13 if it is the 8 of clubs.

11. A bag contains two silver dollars and eight slugs. You pay 50 cents to reach into the bag and take a coin, which you get to keep.

12. A bag contains eight white balls and two black balls. John picks two balls at random from the bag, and he wins \$5 if he does not pick a black ball.

APPLICATIONS

13. **Roulette** In the game of roulette as played in Las Vegas, the wheel has 38 slots. Two slots are numbered 0 and 00, and the rest are numbered 1 to 36. A \$1 bet on any number wins \$36 (\$35 plus the \$1 bet). Find the expected value of this game.

14. **Sweepstakes** A sweepstakes offers a first prize of \$1,000,000, second prize of \$100,000, and third prize of \$10,000. Suppose that two million people enter the contest and three names are drawn randomly for the three prizes.
    (a) Find the expected winnings for a person participating in this contest.
    (b) Is it worth paying a dollar to enter this sweepstakes?

15. **A Game of Chance** A box contains 100 envelopes. Ten envelopes contain \$10 each, ten contain \$5 each, two are "unlucky," and the rest are empty. A player draws an envelope from the box and keeps whatever is in it. If a person draws an unlucky envelope, however, he must pay \$100. What is the expectation of a person playing this game?

16. Combination Lock A safe containing $1,000,000 is locked with a combination lock. You pay $1 for one guess at the six-digit combination. If you open the lock, you get to keep the million dollars. What is your expectation?

17. Gambling on Stocks An investor buys 1000 shares of a risky stock for $5 a share. She estimates that the probability that the stock will rise in value to $20 a share is 0.1 and the probability that it will fall to $1 a share is 0.9. If the only criterion for her decision to buy this stock was the expected value of her profit, did she make a wise investment?

18. Slot Machine A slot machine has three wheels, and each wheel has 11 positions: the digits 0, 1, 2, . . . , 9 and the picture of a watermelon. When a quarter is placed in the machine and the handle is pulled, the three wheels spin independently and come to rest. When three watermelons show, the machine pays the player $5; otherwise, nothing is paid. What is the expected value of this game?

19. Lottery In a 6/49 lottery game, a player pays $1 and selects six numbers from 1 to 49. Any player who has chosen the six winning numbers wins $1,000,000. Assuming that this is the only way to win, what is the expected value of this game?

20. Lightning Insurance An insurance company has determined that in a certain region the probability of lightning striking a house in a given year is about 0.0003, and the average cost of repairs of lightning damage is $7500 per incident. The company charges $25 per year for lightning insurance.

(a) Find the company's expected value for each lightning insurance policy.

(b) If the company has 450,000 lightning damage policies, what is the company's expected yearly income from lightning insurance?

21. Expected Number During the school year, a college student watches TV for two hours a week with probability 0.15, three hours with probability 0.45, four hours with probability 0.30, and five hours with probability 0.10. Find the expected number of hours of TV that he watches per week.

22. Expected Number In a large liberal arts college 5% of the students are studying three foreign languages, 15% are studying two foreign languages, 45% are studying one foreign language, and 35% are not studying a foreign language. If a student is selected at random, find the expected number of foreign languages that he or she is studying.

23. Expected Number A student goes to swim practice several times a week. In any given week the probability that he swims three times is 0.30, two times is 0.45, one time is 0.15, and no times is 0.10. Find the expected number of times the student goes to practice in any given week.

24. Expected Number Consider families with three children, and assume that the probability of having a girl is $\frac{1}{2}$.

(a) Complete the table for the probabilities of having 0, 1, 2, or 3 girls in a family of three children.

(b) Find the expected number of girls in a family of three children.

Number of girls	Probability
0	$\frac{1}{8}$
1	
2	
3	

25–30 ■ A game of chance is described. **(a)** Is the game fair? **(b)** If the game is not fair, find the payout amount that would make the game fair.

25. A Fair Game? A card is drawn from a deck. If the card is the ace of spades you get a payout of $12. If the card is not an ace you have to pay $0.50.

26. A Fair Game? A die is rolled. You get $20 if a one or a six shows; otherwise, you pay $10.

27. A Fair Game? A pair of dice is rolled. You get $30 if two ones show; otherwise, you pay $2.

28. A Fair Game? A die is rolled and a coin is tossed. If the result is a "six" and "heads," you get $10. For any other result you pay $1.

29. A Fair Game? A card is drawn from a deck, a die is rolled, and a coin is tossed. If the result is the "ace of spades," a "six," and "heads," you get $600. For any other result you pay $1.

30. A Fair Game? A bag contains two silver dollars and six slugs. A game consists of reaching into the bag and drawing a coin, which you get to keep. If you draw a slug, you pay $0.50.

DISCOVERY ■ DISCUSSION ■ WRITING

31. The Expected Value of a Sweepstakes Contest A magazine clearinghouse holds a sweepstakes contest to sell subscriptions. If you return the winning number, you win $1,000,000. You have a 1-in-20-million chance of winning, but your only cost to enter the contest is a first-class stamp to mail the entry. Use the current price of a first-class stamp to calculate your expected net winnings if you enter this contest. Is it worth entering the contest?

14.5 Descriptive Statistics (Numerical)

Measures of Central Tendency: Mean, Median, Mode ▶ Organizing Data: Frequency Tables and Stemplots ▶ Measures of Spread: Variance and Standard Deviation ▶ The Five-Number Summary: Box Plots

In *Focus on Modeling* on pages 130–139 we studied **two-variable data**. For example, the data on page 133 measure *two* quantities: asbestos level *and* number of cancer tumors for each individual rat.

Our world is filled with data: weather data, economic data, health data, medical data, political data, census data, and many more. The things that a data set describes are called **individuals**. So individuals could be people, trees, buildings, etc. The property of the individuals that is described by the data is called a **variable**. In this chapter we study **one-variable data**, in which only one property of each individual is measured. For example, the individuals of a data set could be cows, and the variable could be milk production in gallons for each cow.

Data usually consist of thousands or even millions of numbers. The first goal of statistics is to describe such huge sets of data in simpler terms. A **summary statistic** is a single number that summarizes some property of the data. For example, the *average* of a data set is a single number that tells us where the "center" of the data lies. In this section we study several summary statistics.

▼ Measures of Central Tendency: Mean, Median, Mode

One way to make sense of data is to find a "typical" number or the "center" of the data. Any such number is called a **measure of central tendency**. One measure of central tendency is the average (or the mean).

The mean is a summary statistic.

THE MEAN

Let $x_1, x_2, \ldots, x_n$ be n data points. The **mean** (or **average**), denoted by $\bar{x}$, is the sum divided by n:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

EXAMPLE 1 | The Mean Age of Preschoolers

The ages (in years) of the students in a certain class of preschoolers are listed below. Find the average age of a preschooler in this class.

2 2 2 3 3 3 4 4 4 4 4 5

SOLUTION Since the number of children in the class is 12, we find the average by adding up the ages of the children and dividing by 12:

$$\bar{x} = \frac{2+2+2+3+3+3+4+4+4+4+4+5}{12} = \frac{40}{12} \approx 3.3$$

So the average age of a preschooler in this class is about 3.3 years.

NOW TRY EXERCISES **5** AND **25** ■

Another measure of central tendency is the *median,* which is the middle number of an ordered list of numbers—there are as many data points above the median as there are below it.

The median is a summary statistic.

THE MEDIAN

Let $x_1, x_2, \ldots, x_n$ be n data points, written in increasing order.

- If n is odd, the **median** is the middle number.
- If n is even, the **median** is the average of the two middle numbers.

EXAMPLE 2 | Mean and Median Income

The yearly incomes (in thousands of dollars) of five college graduates are listed below.

280 56 59 62 51

(a) Find the mean income.

(b) Find the median income.

(c) Which do you think is a better indicator of central tendency in this case: the mean or the median?

SOLUTION

(a) We add the incomes and divide by 5:

$$\bar{x} = \frac{280 + 56 + 59 + 62 + 51}{5} \approx 102$$

So the average income of the five graduates is about $102,000.

(b) To find the median income, we first put the list of incomes in increasing order:

51 56 59 62 280

Since there are five numbers in this list and 5 is odd, the median is the middle number, 59. So the median income of the five graduates is $59,000.

(c) We see from parts (a) and (b) that the mean and the median are different. The average income of the five graduates is $102,000, but this is not the typical income. In fact, not one of the graduates had an income even close to $102,000. However, the median income of $59,000 is a more typical income for these five graduates. So in this case the median income is a better description of the central tendency than the mean is.

```
1-Var Stats
  x̄=102
   ⋮
  Med=59
   ⋮
```

FIGURE 1

Graphing Calculator Note To obtain summary statistics on the TI-83 graphing calculator, follow the steps on page 929. The calculator output for the data in this example is shown in Figure 1.

NOW TRY EXERCISES **9** AND **27**

If a data set includes a number that is "far out" or far away from the rest of the data, that data point is called an **outlier**. In Example 2, $280,000 is an outlier. In general, when a data set has outliers, the median is a better indicator of central tendency than the mean.

EXAMPLE 3 | Effect of Outliers on the Mean and Median

The table in the margin gives the selling prices of houses sold in 2007 in a small coastal California town.

Address	Selling price ($)
First Street	210,000
Elm Street	260,000
Ocean Avenue	1,800,000
Desert Lane	245,000
Cactus Lane	230,000
Jacaranda Avenue	255,000

(a) Find the mean house price.

(b) Find the median house price.

(c) The house on Ocean Avenue is resold the same year for $2,400,000. Find the mean and median house price for the adjusted data.

(d) Comment on the effects of outliers on the mean and median.

Ricardo Miguel/Shutterstock.com

The National Association of Realtors publishes yearly data on the *median* price of a house, not the average price. This is because the few houses that cost tens of millions of dollars are outliers. These outliers have a large effect on the average price but not on the median price, as we see from Example 3.

SOLUTION

(a) Writing the prices in thousands of dollars, we have

$$\bar{x} = \frac{210 + 260 + 1{,}800 + 245 + 230 + 255}{6} = 500$$

So the average house price is \$500,000.

(b) To find the median, we first arrange the prices in order.

$$210 \quad 230 \quad \underbrace{245 \quad 255}_{\text{Middle two numbers}} \quad 260 \quad 1800$$

Since there is an even number of prices, the median is the average of the middle two numbers:

$$\text{median} = \frac{245 + 255}{2} = 250$$

So the median house price is \$250,000.

(c) For the revised data

$$\bar{x} = \frac{210 + 260 + 2{,}400 + 245 + 230 + 255}{6} = 600$$

So the mean price for the revised data is \$600,000. But the median price is unchanged at \$250,000.

(d) The \$1,800,000 price is an outlier. Because of this outlier the average price of \$500,000 is not a typical price (the majority of houses sold for prices in the 200,000s). The median of \$250,000 is a more typical price. We observe that changing the value of an outlier affects the average but not the median (as long as the change in the outlier keeps the outlier on the same side of the median).

NOW TRY EXERCISE **31**

The *mode* of a data set is a summary statistic that is usually less informative than the mean or median, but has the advantage of not being limited to numerical data.

The mode is a summary statistic.

THE MODE

The **mode** of a data set is the element that appears most often in the data.

For example, the mode of the (non-numerical) data consisting of Korean last names is Kim. The mode of the data set 1, 1, 2, 2, 2, 3, 5, 8 is the number 2. The data set 1, 2, 2, 3, 5, 5, 8 has two modes: 2 and 5. Data sets with two modes are called **bimodal**. The data set 1, 2, 4, 5, 7, 8 has **no mode**.

Organizing Data: Frequency Tables and Stemplots

Sometimes listing the data in a special way can help us get useful information about the data. Two such methods are *frequency tables* and *stemplots.*

A **frequency table** for a set of data is a table that includes each different data point and the number of times that point occurs in the data. The mode is most easily determined from a frequency table.

Frequency Table

Score	Frequency
5	16
4	8
3	5
2	5
1	3
0	3

EXAMPLE 4 | Using a Frequency Table

The scores obtained by the students in an algebra class on a five-question quiz are given in the frequency table in the margin. Find the mean, median, and mode of the scores.

SOLUTION The mode is 5, because more students got this score than any other score.

The total number of quizzes is $16 + 8 + 5 + 5 + 3 + 3 = 40$. To find the mean, we add all the scores and divide by 40. Note that the score 5 occurs 16 times, the score 4 occurs 8 times, and so on. So the mean score is

$$\frac{5(16) + 4(8) + 3(5) + 2(5) + 1(3) + 0(3)}{40} = 3.5$$

There are 40 students in this class. If we rank the scores from highest to lowest, the median score is the average of the 20th and 21st scores. Looking down the frequency column in the table, we see that these scores are each 4. So the median score is 4.

In Example 5 the mean, median, and mode are different from each other, and each gives us different information about the central tendency of the data.

Graphing Calculator Note To obtain the mean and median on the TI-83, follow the instructions on page 929 for calculating with data given by a frequency table.

NOW TRY EXERCISES **15** AND **35**

A **stemplot** (or a **stem-and-leaf plot**) organizes the data by using the digits of the numbers in the data. Each number in the data is written as a **stem** consisting of the leftmost digits and a **leaf** consisting of the rightmost digit. Numbers with the same stem are grouped together in a row, with the stem written only once. The next example illustrates the method.

EXAMPLE 5 | Making a Stem-and-Leaf Plot

Swim Times (s)

42.8	41.5	42.7	46.2
41.1	42.9	40.2	43.6
42.7	39.5	43.1	40.5
43.3	41.1	46.7	42.7

A county swim club records the times (in seconds) shown in the margin for the 50-yard freestyle swim event.

(a) Make a stem-and-leaf plot of the data.

(b) Find the median and mode using the stemplot from part (a).

SOLUTION

(a) We use the first two digits for the stems and the digit after the decimal point for the leaves. For example, 39.5 has stem 39 and leaf 5.

Stem	Leaves
39	5
40	2 5
41	1 1 5
42	7 7 7 8 9
43	1 3 6
44	
45	
46	2 7

39 | 5 means 39.5

We include the stems that have no leaves, because they provide information about the data. In this case no swimmer had a time in the 44- or 45-second range.

(b) There are 16 data points. The stem-and-leaf plot displays the data in increasing order. So to find the median, we count the leaves starting at the smallest data point until we reach the middle two numbers (the eighth and ninth leaves). The corresponding data points are 42.7 and 42.8, so the median is

$$\text{median} = \frac{42.7 + 42.8}{2} = 42.75$$

The mode is the number that appears most often in the data. We can see from the stem-and-leaf plot that this number is 42.7.

NOW TRY EXERCISES **17** AND **37**

Stem	Leaves
0	5
1	7
2	9 9
3	
4	
5	5 6 6
6	2 8 9 9 9 9
7	3 4
8	3
9	8

2 | 8 means 28

EXAMPLE 6 | Reading a Stem-and-Leaf Plot

A stemplot of the ages of patients in a hospital ward is shown in the margin.

(a) How many patients are 56 years old?

(b) How many patients are in their 50s? In their 40s?

(c) What is the total number of patients in the ward?

(d) What age group had the most patients?

(e) What are the ages of the youngest and oldest patients?

(f) Find the mean, median, and mode.

SOLUTION From the statement "2 | 8 means 28" we see that the stem represents the tens digit and the leaves represent the units digit. Reading the stemplot starting at the lowest age, we get 5, 17, 29, 29, 55, 56, 56, and so on until we reach 98.

(a) Two patients are 56 years old.

(b) Counting the leaves that correspond to the stem 5, we see that there are three patients in their 50s. There are no patients in their 40s.

(c) Counting all the leaves, we see that there are 17 patients.

(d) The stem with the most leaves is 6, so the patients in their 60s are the largest group.

(e) The youngest patient is 5, and the oldest is 98.

(f) The mode is 69. There are 17 patients, so the median is the middle age (the ninth age), which is 68. To find the mean, we add all the ages and divide by 17:

$$\bar{x} = \frac{5 + 17 + 2(29) + 55 + 2(56) + 62 + 68 + 4(69) + 73 + 74 + 83 + 98}{17} \approx 57.7$$

So the mean age is approximately 57.7 years.

NOW TRY EXERCISES **19** AND **39**

Measures of Spread: Variance and Standard Deviation

Measures of central tendency identify the "center" or "typical value" of the data. **Measures of spread** (also called **measures of dispersion**) describe the spread or variability of the data around a central value. For example, each of the following data sets has mean 72, but it is clear that the first set of data has more variability than the second.

50, 58, 78, 81, 93   72, 71, 72, 72, 73

The most important measures of variability in statistics are the *variance* and *standard deviation.*

The standard deviation and variance are summary statistics.

The symbol σ is the lower case Greek letter "sigma."

THE VARIANCE AND STANDARD DEVIATION

Let $x_1, x_2, \ldots, x_N$ be N data points, and let $\bar{x}$ be their mean. The **standard deviation** of the data is

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2}$$

The **variance** is σ^2, the square of the standard deviation.

EXAMPLE 7 | Calculating Standard Deviation

Two machines are used in filling 16-ounce soda bottles. To test how consistently each machine fills the bottles, a sample of 20 bottles from the output of each machine is selected. Find the standard deviation for each machine. Which machine is more consistent in filling the bottles?

Soda Machine I (x)				**Soda Machine II (y)**			
16.2	16.1	15.9	16.0	16.1	16.2	16.4	.15.8
16.4	16.1	16.1	16.1	16.2	15.6	16.3	16.3
15.8	15.9	16.1	15.8	16.8	15.4	16.0	15.6
16.0	16.0	16.2	15.9	15.8	16.2	16.2	16.1
16.3	16.1	16.0	16.0	16.3	15.8	16.0	15.9

SOLUTION The first step is to find the mean for each soda machine:

$$\bar{x} = \frac{16.2 + 16.1 + \cdots + 16.0}{20} = 16.05$$

$$\bar{y} = \frac{16.1 + 16.2 + \cdots + 15.9}{20} = 16.05$$

The standard deviations σ_x and σ_y for Soda Machines I and II, respectively, are

$$\sigma_x = \sqrt{\frac{(16.2 - 16.05)^2 + (16.1 - 16.05)^2 + \cdots + (16.0 - 16.05)^2}{20}} \approx 0.15$$

$$\sigma_y = \sqrt{\frac{(16.1 - 16.05)^2 + (16.2 - 16.05)^2 + \cdots + (15.9 - 16.05)^2}{20}} \approx 0.32$$

Soda Machine I is more consistent in filling the bottles because the standard deviation of the data from Machine I is much smaller than that of the data from Machine II.

```
1-Var Stats
  x̄=16.05
   ⋮
  σx=.15
   ⋮
```

FIGURE 2

Graphing Calculator Note Follow the steps on page 929 to obtain summary statistics on the TI-83 graphing calculator. The calculator output for Machine I is shown in Figure 2.

NOW TRY EXERCISES **21** AND **41**

▼ The Five-Number Summary: Box Plots

A simple indicator of the spread of data is the location of the **minimum** and **maximum** values. Other indicators of spread are the *quartiles.* Recall that the median divides a data set in half. The median of the lower half of the data is called the **first quartile**, Q_1. The median of the upper half of the data is called the **third quartile**, Q_3. The minimum, maximum, median, and first and third quartiles together give a good picture of the spread *and* center of the data.

THE FIVE-NUMBER SUMMARY

The **five-number summary** for a data set are the five numbers below, written in the indicated order.

minimum, Q_1, median, Q_3, maximum

Figure 3 shows how these five numbers are related for a data set listed in order. For the data in the figure, $Q_1 = (2 + 4)/2 = 3$ and $Q_3 = (16 + 20)/2 = 18$.

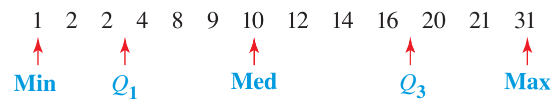


FIGURE 3 The five-number summary

The five-number summary allows us to quickly calculate the following two different indicators of spread. A simple indicator of spread is the **range**, which is the difference between the maximum and minimum values:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

This can be compared to the spread of the middle of the data as measured by the **interquartile range** (**IQR**), which is the difference between the third and first quartiles:

$$\text{IQR} = Q_3 - Q_1$$

For the data in Figure 3 the range is $41 - 1 = 40$, and the interquartile range is $\text{IQR} = 18 - 3 = 15$.

EXAMPLE 8 | Using a Five-Number Summary

The lists in the margin give HDL cholesterol levels ("good cholesterol") for 17 male and 13 females who participated in a health survey. Find the five-number summary for each group. Compare the center and the spread in cholesterol levels for the two groups.

Males

32	52	30	42	44
52	55	29	29	31
42	46	41	34	50
36	30			

Females

58	60	48	55	56
46	54	50	49	51
53	54	53		

SOLUTION For each group we first arrange the data in ascending order and then find the median. Next we find Q_1, the median of the lower half of the data, and Q_3, the median of the upper half of the data. You can check that the five-number summaries are as follows:

Males:	29, 30.5, 41, 48, 55
Females:	46, 49.5, 53, 55.5, 60

The range and interquartile range are calculated as follows:

Males:	Range $= 55 - 29 = 26$	$\text{IQR} = 48 - 30.5 = 17.5$
Females:	Range $= 60 - 46 = 14$	$\text{IQR} = 55.5 - 49.5 = 6.0$

We first observe that the median HDL level for females, 53, is significantly higher than that of the males, 41. Moreover, female HDL levels show considerably less variability than the males, as measured by the IQR.

Graphing Calculator Note To obtain the five-number summary on the TI-83 calculator, follow the instructions on page 929. The calculator output for the male HDL data is shown in Figure 4.

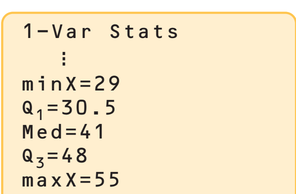


FIGURE 4

NOW TRY EXERCISES **23(a)** AND **43** ■

A **box plot** (also called a **box-and-whisker plot**) is a method for graphically displaying a five-number summary. The plot consists of a rectangle whose left- and right-hand sides correspond to Q_1 and Q_3, respectively. The box is divided by a line segment at the location of the *median*. The **whiskers** are line segments that extend from both sides of the box to the location of the *maximum* and *minimum* values. Figure 5 shows a box plot for the data in Figure 3.

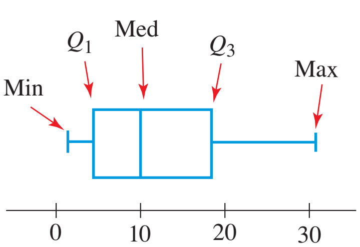


FIGURE 5

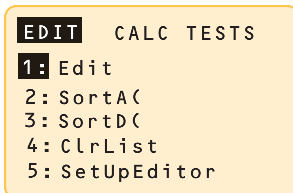
(*Text continues on page 930.*)

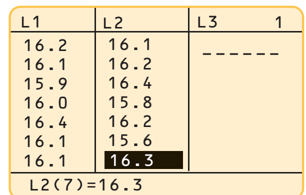
SUMMARY STATISTICS
On a Graphing Calculator

The following steps show how to obtain summary statistics on the TI-83 or TI-84 calculators.

Step 1 Enter the Data

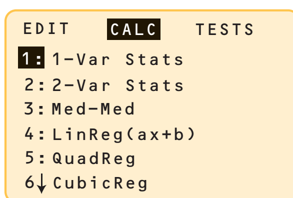
Press the STAT key. Choose **EDIT**, then **1:Edit**, then press ENTER. Now enter the data in one of the columns labeled **L1**, **L2**, **L3**, ... Enter different data sets in separate columns as shown below. If the data are given in a frequency table, enter the data in **L1** and the corresponding frequencies in **L2**.

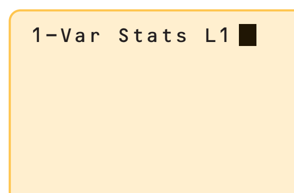




Step 2 Select Summary Statistics

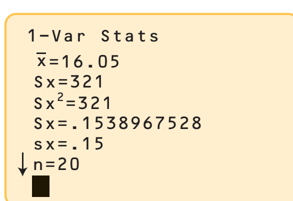
Press the STAT key, choose **CALC** and **1:1-Var Stats**, then press ENTER. Next, choose the column in which the data have been entered. To select **L1**, press 2nd 1; the result is shown below. If the data are entered in **L1** and the corresponding frequencies are entered in **L2**, select **L1**, **L2** (separated by a comma).

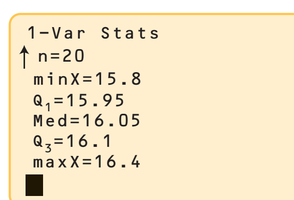




Step 3 Obtain Results

Press ENTER to obtain the summary statistics for the data, as shown below. Scroll down to see more of the statistics.





EXAMPLE 9 | Box Plots

Draw box plots for the two data sets in Example 8. Use the box plots to compare the central tendency and the spread of the two data sets.

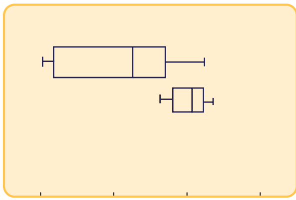


FIGURE 6

SOLUTION The box plots are drawn by using a graphing calculator in Figure 6. From the box plots we can visually compare the relative locations of the five number summaries for each data set.

The median for females is considerably to the right of the median for males, so the median HDL is higher for females. The horizontal length of the box is the interquartile range. From Figure 6 we see that the IQR is quite a bit wider for males, indicating more variability in the male data. The ends of the whiskers indicate the minimum and maximum values. Again, we see that the male data have a much larger range and that the female data are toward the top end of the male data.

Graphing Calculator Note To obtain box plots on the TI-83, follow the instructions on page 939 in Section 14.6.

NOW TRY EXERCISES **23(b)** AND **43** ■

When we are working with quartiles, the median of the data is also called the **second quartile**, Q_2. So the three quartiles Q_1, Q_2, and Q_3 separate the data into four parts. Statisticians also use *quintiles, deciles,* and *percentiles.* The familiar **percentiles** are numbers that divide the data into 100 parts, each with the same number of data points. Standardized test scores are often reported as percentiles. So a score in the 99th percentile is a score in the top one-hundredth part of the scores on that test.

14.5 EXERCISES

CONCEPTS

1. If we measure one property of the individuals in a group (height, for example) we obtain _____-variable data. What is meant by a measure of *central tendency* for one-variable data? Measures of central tendency for one-variable data include the __________, the __________, and the __________

2. (a) To find the average of a list of n numbers, we first _____ all the numbers and then divide by _____.

(b) The mode of a data set is the data point that appears _______ often.

3. To find the median of a list of n numbers, we first put the numbers in _____________.

(a) If n is odd, the median is the _______ number in the list.

(b) If n is even, the median is the average of the two _______ numbers in the list.

4. (a) What is meant by a *measure of spread* for a data set? Measures of spread for a data set include the standard ___________.

(b) The numbers in the *five-number summary* of a data set are the ____________, the ____________, the ____________, the ____________, and the ____________.

SKILLS

5–10 ■ A data set is given. **(a)** Find the median of the data. **(b)** Find the average of the data. How many data points are greater than the average?

5. 113 21 16 18 19 29 22

6. 130 510 116 132 140 132 121

7. 69 71 74 73 72 73 69 69

8. 91 87 84 82 87 84 93 82

9. 39 28 57 11 45 72 38
28 41 43 60 59 24 17

10. 109 110 97 130 112 102 95
81 14 200 117 110 119

11–12 ■ A frequency table for a data set is given. Find the mean, median, and mode.

11. **Frequency Table**

x	Frequency
16	13
17	5
18	12
19	0
20	2

12. **Frequency Table**

x	Frequency
4	1
5	0
6	2
7	10
8	12
9	8
10	5

13–16 ■ A data set is given. **(a)** Organize the data into a frequency table. **(b)** Find the mean, median, and mode of the data.

13. 11 9 6 8 11 9 8 11

14. 20 21 17 21 20 20 19 20

15. 52 49 47 38 52 49
52 52 47 52 49 52
38 41 41 49 52 52

16. 17 19 20 13 21 15 15 17
17 15 19 19 15 14 20 14
14 17 17 13 17 15 20 19

17–18 ■ A stem-and-leaf plot of a data set is given. **(a)** What is the total number of data points? **(b)** Find the median and mode of the data.

17.

Stem	Leaves
0	3 4
1	0 1 1 5
2	3
3	4 4
4	
5	5 6 6 6 9
6	2 3 7 7 8 8
7	4 5
8	2 3 3
9	1 4 5 5

1 | 2 means 12

18.

Stem	Leaves
3.1	4 4 5 6 9
3.2	0 0 3 5 5 7
3.3	9
3.4	2 3 3 5 8 8
3.5	1 1 1 3

3.1 | 2 means 3.12

19–20 ■ A data set is given. **(a)** Make a stem-and-leaf plot of the data. **(b)** Find the median and mode of the data using the stemplot of part (a).

19. 13 21 41 42 41
40 32 33 35 33
41 41 35 22 29

20. 1.10 1.17 1.22 1.31 1.18 1.17 1.35
1.24 1.26 1.23 1.17 1.46 1.47 1.40
1.16 1.47 1.31 1.36 1.34 1.22 1.25

21–22 ■ A data set is given. **(a)** Find the mean. **(b)** Find the standard deviation.

21. 28.6 30.2 29.5 30.9 31.2
29.5 30.2 30.8 28.2 29.6
32.1 31.6 30.2 28.4 29.0

22. 8 8 11 6 5
6 8 9 6 10
9 9 6 8 4
3 7 8 10 9

23–24 ■ A data set is given. **(a)** Find the five-number summary for the data set. **(b)** Draw a box plot for the data.

23. 13 15 17 18 16
15 13 12 17 21
29 28 19 26 25
22 25

24. 62 70 95 61 72 73 89
82 85 68 99 93 64 71
79 78 65 85 86 72 92
94 97 83 85 70

APPLICATIONS

25. Basketball Stats The data set below shows the number of points scored by Kobe Bryant of the Los Angeles Lakers in each basketball game in which he played in February 2007. What is the average number of points per game that Kobe Bryant scored in that month?

46 30 6 11 36
33 31 29 23 41
17 21 30 21 33

26. Lion Prides A wildlife biologist in the southern Sahara Desert of Africa records the number of lions in the different prides in her area. What is the average number of lions in a pride?

18 8 14 15 16 12 17

27. Apgar Score Doctors use the Apgar score to assess the health of a newborn baby immediately after the child is born. The data set below shows the Apgar scores of babies born in Memorial Hospital on March 11, 2007. Find the average and the median Apgar scores for these babies.

9 8 10 3 5 8 10

28. Weights of Sextuplets The Hanselman sextuplets were born three months premature on February 26, 2004, in Akron, Ohio, and they have all survived to this date. The data set below shows the birth weights of each child in kilograms. What are the average and the median birth weights of these sextuplets?

2.6 1.6 2.4 2.5 2.5 2.1

29. Quiz Average The data set below shows Jordan's scores on her first five algebra quizzes.

(a) What is her average quiz score?

(b) Jordan receives a score of 10 on her sixth quiz, so now what is her average quiz score?

9 9 6 7 7

30. **Quiz Average** The data set below shows Chad's scores on his first four geography quizzes.
(a) What is his average quiz score?
(b) Chad receives a score of 5 on his fifth quiz, so now what is his average quiz score?

7 9 8 9

31. **Investment Seminar** The organizer of an investment seminar surveys the participants on their yearly income. The data set below shows the yearly income (in dollars) of the participants.
(a) Find the average and median income of the participants.
(b) How many participants have an income above the average?
(c) A new participant joins the seminar, and her yearly income is $500,000. Now what are the average and median incomes? In this case, do you think that the average or the median is the better indicator of central tendency?
(d) Comment on the effects of outliers on the mean and median.

56,000 58,000 48,000 45,000
59,000 72,000 63,000

32. **Dairy Farming** A dairy farmer in Illinois records the weights (in pounds) of all his cows. The table below shows his data.
(a) Find the average and median weight of the cows on the farm.
(b) How many cows have a weight above the average?
(c) The farmer purchases a calf that weighs only 420 pounds. Now what are the average and median weights of the cows on the farm? In this case, do you think that the average or the median is the better indicator of central tendency?
(d) Comment on the effects of outliers on the mean and median.

880 970 930 890 980 920 900

33. **Home Sales** A realty agency in Albuquerque, New Mexico, records the prices (in dollars) of homes sold in one neighborhood.
(a) What are the average and median home prices in this neighborhood? Which do you think is the better indicator of central tendency?
(b) Is the house that sold for $329,000 above or below the average price? The median price?
(c) After these data were gathered, another home in the neighborhood sold for $2,860,000. Now what are the average and median home prices? In this case, do you think that the average or the median is the better indicator of central tendency?

299,000 329,000 355,000
316,000 330,000

34. **Home Sales** A realty agency in Napa Valley, California records the prices (in dollars) of homes sold in one neighborhood.
(a) What are the average and median home prices in this neighborhood? Which is the better indicator of central tendency?
(b) Is the house that sold for $2,319,000 above or below the average price? The median price?
(c) After these data were gathered, another home in the neighborhood sold for $300,000. Now what are the average and median home prices? Which do you think is the better indicator of central tendency in this case?

2,278,000 2,231,000 2,319,000 2,279,000
2,365,000 2,279,000 2,319,000

35. **Mice** The weights of newborn mice are given in the following frequency table. Find the mean, median, and mode of the weights.

Frequency Table

Weight (gm)	Frequency
0.6	2
0.7	5
0.8	9
0.9	10
1.0	16
1.1	15
1.2	18
1.3	8
1.4	6

36. **Goals** A statistician records the total number of goals in the last game of the season for each soccer team in a local club. Find the mean, median, and mode of the number of goals.

Frequency Table

Number	Frequency
0	4
1	5
2	3
3	2
4	3
5	2
6	1
7	2
8	1

37. **Basketball Team** The stemplot below shows the total number of points scored in the first game of the season for each player on a local basketball team.
(a) How many players scored 12 points?
(b) How many players scored more than ten points?
(c) What is the total number of players on the team?
(d) What is the total number of points scored?
(e) Find the mean, median, and mode.

Stem	Leaves
0	1 2 2 3 3 5 8
1	2 2 2 5
2	1
3	0

2 | 1 means 21

38. Weddell Seals The stemplot below shows weights (in pounds) of a sample of adult Weddell seals recorded in an area of Antarctica.

(a) How many seals weigh 962 pounds?
(b) How many seals weigh between 940 and 960 pounds?
(c) How many seals were weighed?
(d) Find the mean, median, and mode.

Stem	Leaves
90	2 7
91	
92	0 1 4 8
93	2 4 7 9 9
94	3 5 8 8 8 8 8 9
95	1 4 4 7 8
96	1 1 2 2 2 4 5 5 7 8 8 9 9
97	0 1 4 5 5 9
98	1 2 2 4 6 7 7
99	1 6

91 | 5 means 915

39. Blood-Clotting Test A sample of patients on the anticoagualant Warfarin is tested for their blood coagulation (clotting) rate with the INR test. The results of the test are listed below.

(a) Make a stem-and-leaf plot of the data.
(b) Find the median and mode using the stemplot from part (a).

INR

1.8	3.2	2.4	3.1
3.4	2.4	2.6	4.0
2.6	2.9	3.0	4.1
2.6	2.3	2.8	2.2

40. Gasoline Prices Listed below are the gasoline prices for regular unleaded gas in a one-month period at a local station.

(a) Make a stem-and-leaf plot of the data.
(b) Find the median and mode using the stemplot from part (a).

Gasoline Price ($)

3.01	3.03	3.10	3.05	3.02
3.02	3.25	3.09	3.10	3.09
3.08	3.02	3.41	3.43	3.05
3.22	3.26	3.04	3.23	3.07
3.09	3.10	3.09	3.24	3.12
3.31	3.35	3.36	3.19	3.17

41. Incomes and Gender Listed below are the yearly incomes of a sample of men and women in the state of New York. Find the mean and standard deviation for the two data sets. Does there appear to be a difference in variation between the two data sets?

Yearly Income of Men ($)

46,000	49,000	54,000	35,000
72,000	57,000	56,000	51,000
48,000	53,000	61,000	59,000

Yearly Income of Women ($)

48,000	46,000	32,000	58,000
33,000	59,000	61,000	45,000
39,000	65,000	62,000	41,000

42. Crop Rotation Crop rotation is the practice of planting dissimilar crops sequentially in one location to avoid the buildup of pathogens and pests and the depletion of nutrients in the soil. A researcher conducts an experiment on crop rotation of strawberries. He uses crop rotation on 12 garden plots and does not use crop rotation on 12 other garden plots. Listed below are the numbers of pounds of strawberries harvested from each plot in the final growing season. Find the mean and standard deviation for each data set. Does there appear to be a difference in variation between the two data sets?

Crop Rotation (lb)				No Crop Rotation (lb)			
37	42	44	46	29	35	34	27
41	39	29	31	31	36	41	40
37	41	42	45	37	29	31	36

43. Carbohydrates The daily carbohydrate intake (in grams) of 18 Asians and 20 Hispanics who participated in a health survey are listed below. Find the five-number summary, and draw a box plot for each group. Compare the center and the spread in carbohydrate intake levels for the two groups.

Asian				Hispanic			
344	285	244	382	310	330	289	287
310	265	272	310	370	305	314	267
291	286	288	305	273	325	347	364
410	350	230	273	268	272	329	340
286	291			289	295	349	278

44. Test Scores The standardized test scores for seventh graders from two different schools in the same neighborhood are listed below. One sample is taken from a school with less than 100 students, and the other sample is taken from a school with more than 400 students. Find the five-number summary, and draw a box plot for each group. Compare the center and the spread in test scores for the two groups.

Small School Test Scores				Large School Test Scores			
1429	1172	1396	1550	1509	1572	1190	950
1245	1162	1051	1276	1143	1432	1198	1171
1367	1301	1265	1156	1039	967	1013	1025
1274	1105	1050	1040	1127	1041	1053	1429
1187				1592	1450	1590	1229

DISCOVERY ■ DISCUSSION ■ WRITING

45. Measures of Central Tendency For each of the following situations, determine whether the mean, median, or mode gives the most appropriate measure of central tendency. Explain your answer.

(a) A botanist counts the number of buds on each rose bush in his collection of 50 floribunda rose bushes. Most bushes have about 30 buds, but a few are sickly and have very few buds.
(b) The heights of 30 kindergarten students in an elementary school are measured.
(c) A researcher records the blood type of each individual in a sample of 60 middle-aged males.

46. Measures of Central Tendency A researcher computes the mean, median, and mode for the weights of milking cows on a small dairy farm. Determine which of the following statements *cannot* be true, and explain your answer.

(a) More than half of the cows have weights below the mode.
(b) More than half the cows have weights below the median.
(c) More than half the cows have weights below the mean.
(d) None of the cows have a weight exactly equal to the mean.
(e) None of the cows have a weight exactly equal to the mode.
(f) None of the cows have a weight exactly equal to the median.

14.6 Descriptive Statistics (Graphical)

Data in Categories ▶ Histograms and the Distribution of Data ▶ The Normal Distribution

In the preceding section we described data using summary statistics for central tendency and spread. In this section we describe data using graphs called histograms. Histograms allow us to see how the *entire* data set is *distributed* among its different values.

▼ Data in Categories

In Section 14.5 we studied **numerical data**—that is, data whose values are real numbers. For example, age, height, and cholesterol level are real numbers. But in many situations we obtain **categorical data**—that is, the data tell us only whether an individual does or does not belong to a particular category. For example, a survey of your class could include categorical data such as hair color (black, brown, blond, red), political affiliation (Democrat, Republican, Independent), and so on. In this case the data indicate the *number* of individuals in each category. Categorical data can be represented by *bar graphs* or *pie charts.*

A **bar graph** consists of vertical bars, one bar for each category. The height of each bar is proportional to the number of individuals in that category. So the y-axis has a numerical scale corresponding to the number (or the proportion) of the individuals in each category. The labels on the x-axis describe the different categories. See Figure 1(a).

A **pie chart** consists of a circle divided into sectors, one sector for each category. The central angle of each sector is proportional to the number of individuals in that category. Each sector is labeled with the name of the corresponding category. See Figure 1(b).

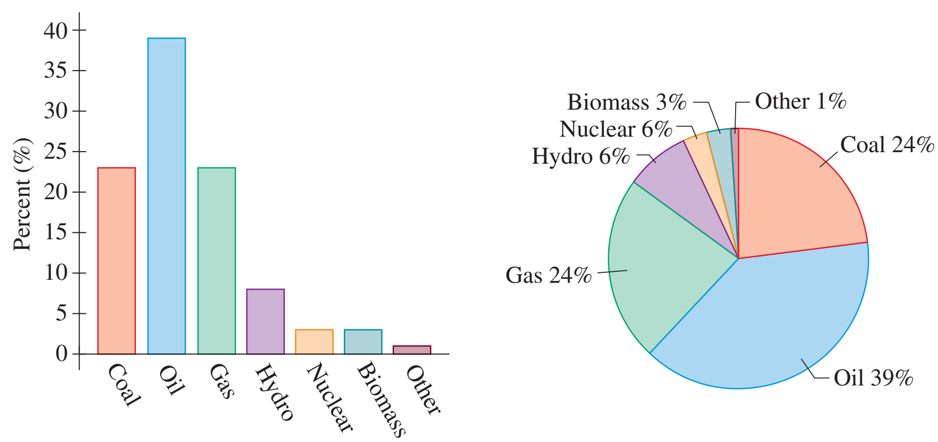


FIGURE 1 U.S. energy sources

(a) Bar graph (b) Pie chart

▼ Histograms and the Distribution of Data

For categorical data, bar graphs are an excellent method of visualizing how the data are distributed among the different categories. To visualize the distribution of one-variable numerical data, we first convert the data into categorical data. For example, "Age" is a numerical variable, but we can convert it into a categorical variable by placing individuals in age categories (for example, 0–10 years, 11–20 years, and so on) and recording the number of individuals in each category.

In general, to put numerical data into categories, we divide the range of the data into nonoverlapping contiguous intervals of equal lengths, called **bins**. To draw a **histogram** of the data, we first label the bins on the x-axis and then erect a rectangle on each bin; the height (and hence also the area) of each rectangle is proportional to the number of data points in that bin.

EXAMPLE 1 | Drawing a Histogram

52	68	69	70	78
78	79	81	83	88
88	88	88	89	89
89	92	92	95	95
95	98	98	98	99
99	101	101	103	106
108	108	109	109	113
115	115	119	128	139

Hospitals routinely measure patients' heart rates. The list in the margin gives the heart rate (in beats per minute) of 40 patients admitted to the ICU unit at a city hospital. Draw a histogram of the heart rate data. Use bins of size 10 starting at 50 beats per minute.

SOLUTION We first organize the data into bins of size 10, as in the frequency distribution table below. Next, we graph a bar over each interval; the height of each bar is the number of data points in that interval. For example, the height of the bar over the interval $60 \leq x < 70$ is 2. The histogram is shown in Figure 2.

Frequency Distribution

Interval	Frequency
$50 \leq x < 60$	1
$60 \leq x < 70$	2
$70 \leq x < 80$	4
$80 \leq x < 90$	9
$90 \leq x < 100$	10
$100 \leq x < 110$	8
$110 \leq x < 120$	4
$120 \leq x < 130$	1
$130 \leq x < 140$	1

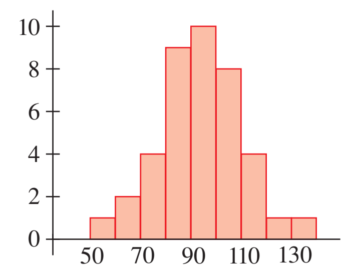


FIGURE 2 Histogram of heart rate data

Graphing Calculator Note To graph a histogram on a TI-83 calculator, follow the instructions on page 939.

NOW TRY EXERCISE **3**

A histogram gives a visual representation of how the data are **distributed** in the different bins. The histogram allows us to determine whether the data are **symmetric** about the mean, as in Figure 3. The heart rate data in Figure 2 are approximately symmetric. If the histogram has a long "tail" on the right, we say that the data are **skewed to the right**. Similarly, if there is a long tail on the left, the data are **skewed to the left**. Since the area of each bar in the histogram is proportional to the number of data points in that category, it follows that the **median** of the data is located at the x-value that divides the area of the histogram in half. Recall that extreme values have a large effect on the mean but not on the median (see Example 3 on page 923). So if the data are skewed to the right, the mean is to the right of the median; if the data are skewed to the left, the mean is to the left of the median.

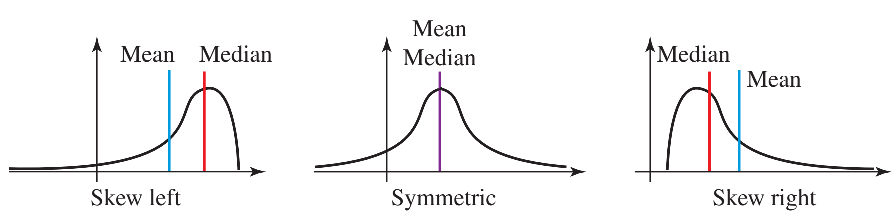
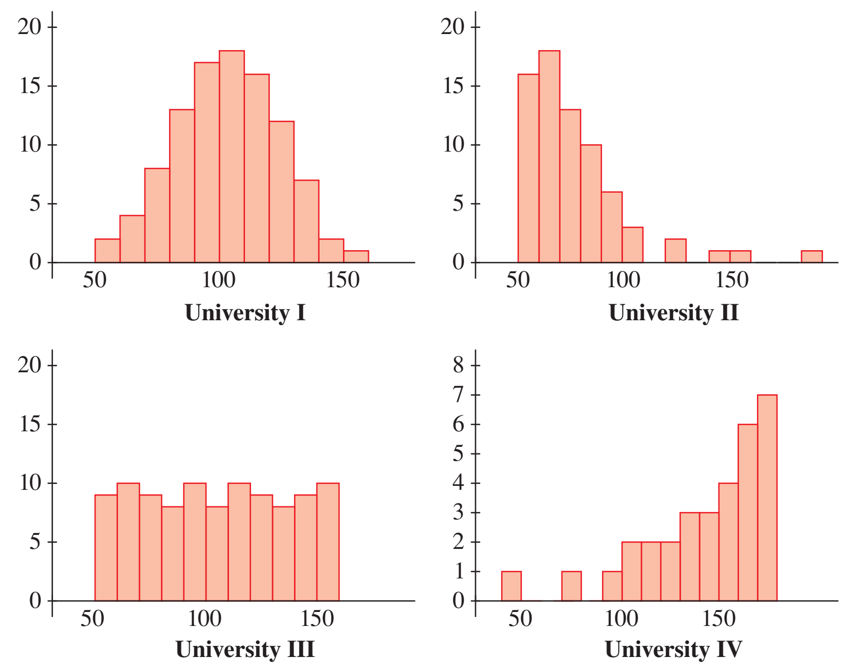


FIGURE 3 Symmetric and skew distributions

EXAMPLE 2 | Getting Information from a Histogram

Graduates of Universities I, II, III, and IV are surveyed about their yearly salaries. Histograms of the data are shown below, with the x-axis representing salaries in thousands of dollars. Are the data symmetric, skew, or neither? Determine the approximate location of the median, and determine the relative location of the mean.



SOLUTION

University I: The histogram shows that the data are more or less symmetric. The mean and median salaries are about \$100,000. Most of the salaries are close to the mean, with fewer salaries at the extremes.

University II: The histogram shows that the data are skewed to the right. The median is about \$70,000, because half the area of the histogram is below this salary and half is above. Since the data are skewed to the right, the mean is higher than \$70,000.

University III: The histogram shows that the data are more or less symmetric. The mean and median salaries are about \$100,000. But unlike University I, the salaries are more or less *uniformly* distributed, with approximately the same number of graduates in each salary category.

University IV: The histogram shows that the data are skewed to the left. The median is about \$150,000, because half the area of the histogram is below this salary and half is above. Since the data are skewed to the left, the mean is less than \$150,000.

NOW TRY EXERCISE **5**

The Normal Distribution

The binomial distribution we studied in Section 14.3 is also approximately normal.

Most real-world data are distributed in a special way called a *normal distribution*. For example, the heart rate data in Example 1 are approximately normally distributed. Figure 4 shows how a histogram of heart rate data with more data and smaller interval sizes fits a curve. The curve can be modeled by a function that we now define.

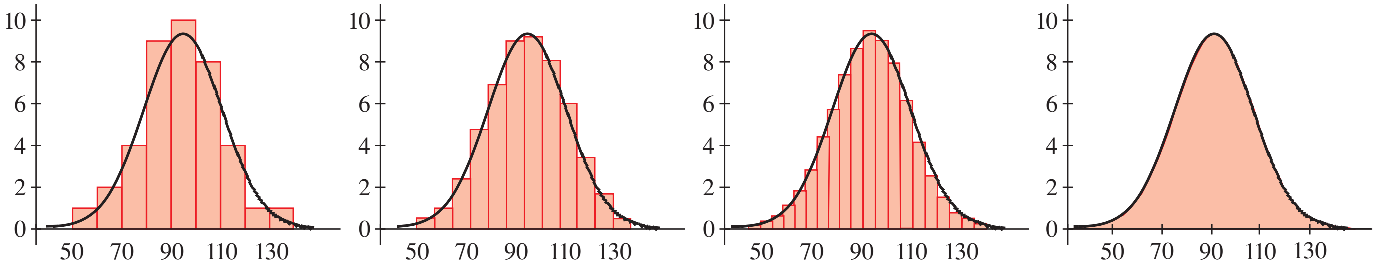


FIGURE 4 Histogram approximates normal curve

The **standard normal distribution** (or **standard normal curve**) is modeled by the function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

This distribution has mean 0 and standard deviation 1. **Normal distributions** with different means and standard deviations are modeled by transformations (shifting and stretching) of the above function. Specifically, the normal distribution with mean μ and standard deviation σ is modeled by the function

$$f(z) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

All normal distributions have the same general shape, called a **bell curve**. Graphs of several normal distributions are shown in Figure 5.

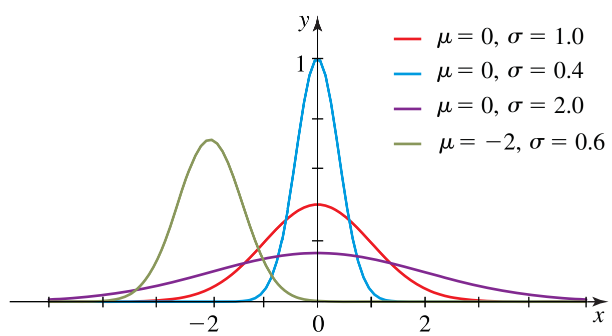


FIGURE 5 Normal curves

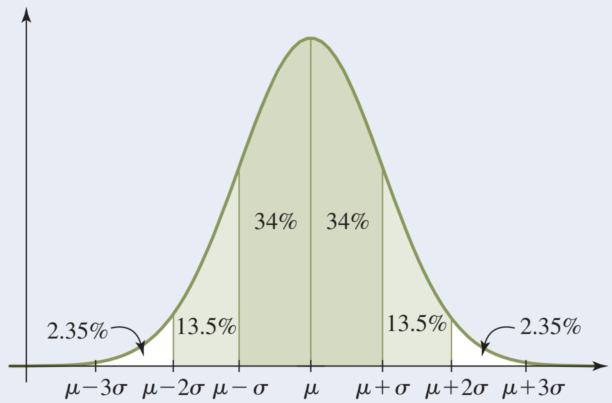
It can be shown by using calculus that for *any* data that are normally distributed about 68% of the data are within one standard deviation of the mean, about 95% are within two standard deviations of the mean, and about 99.7% are within three standard deviations of the mean. In fact, for normally distributed data the proportion of the data between any two values is completely determined by the mean and standard deviation of the data. This information is programmed into most graphing calculators.

NORMALLY DISTRIBUTED DATA

For normally distributed data with mean μ and standard deviation σ we have the following facts, called the **Empirical Rule**.

- Approximately 68% of the data are between $\mu - \sigma$ and $\mu + \sigma$.
- Approximately 95% of the data are between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data are between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The percentage of the data between other values can be obtained by using a graphing calculator.

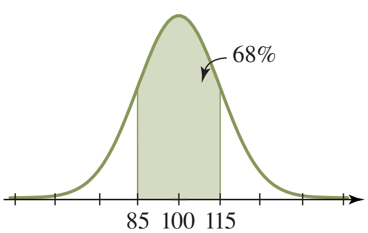


EXAMPLE 3 | Using the Normal Distribution (Empirical Rule)

IQ scores are normally distributed with mean 100 and standard deviation 15. Find the proportion of the population with IQ scores in the given interval. Also find the probability that a randomly selected individual has an IQ score in the given interval.

(a) Between 85 and 115 **(b)** At least 130 **(c)** At most 130

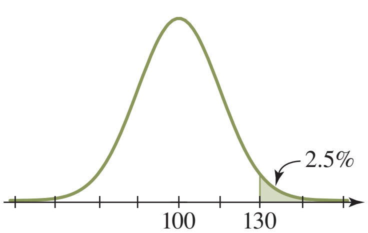
SOLUTION



(a) IQ scores between 85 and 115 are within one standard deviation of the mean:

$$100 - 15 = 85 \qquad \text{and} \qquad 100 + 15 = 115$$

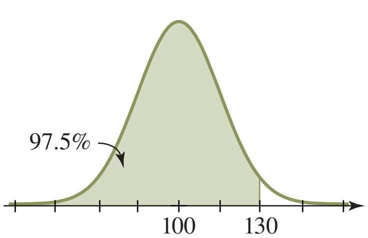
By the Empirical Rule, about 68% of the population have IQ scores between 85 and 115. So the probability that a randomly selected individual has an IQ between 85 and 115 is 0.68.



(b) IQ scores between 70 and 130 are within two standard deviations of the mean:

$$100 - 2(15) = 70 \qquad \text{and} \qquad 100 + 2(15) = 130$$

By the Empirical Rule, about 95% of the population have IQ scores between 70 and 130. The remaining 5% of the population have IQ scores above 130 or below 70. Since normally distributed data are symmetric about the mean, it follows that 2.5% have IQ scores above 130 (and 2.5% below 70). So the probability that a randomly selected individual has an IQ of at least 130 is 0.025.



(c) By part (b), 2.5% of the population have IQ scores above 130. It follows that the rest of the population have IQ scores below 130. Thus 97.5% of the population have IQ scores of 130 or below. So the probability that a randomly selected individual has an IQ of 130 at most is 0.975.

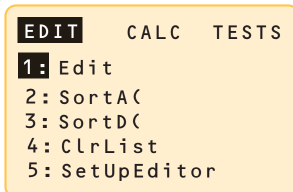
NOW TRY EXERCISE **13** ■

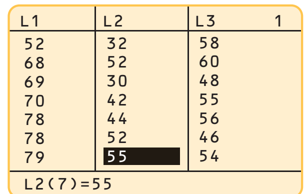
STATISTICAL GRAPHS
On a Graphing Calculator

The following steps show how to obtain histograms and box plots on the TI-83 or TI-84 calculators.

Step 1 Enter the Data

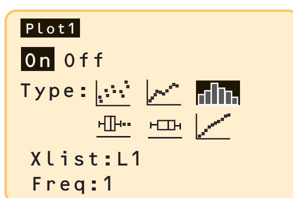
Press the STAT key. Choose EDIT, then 1:EDIT, then press ENTER. Now enter the data in one of the columns labled L1, L2, L3, ... If the data are from a frequency table, enter the frequencies in another column.

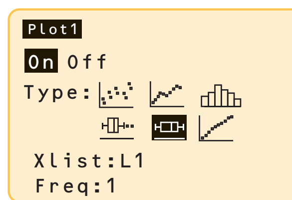




Step 2 Choose the Type of Graph

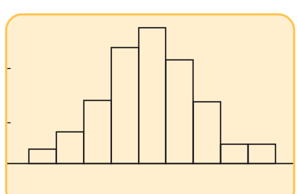
Press 2nd STAT PLOT. To draw a histogram or box plot, choose 1:Plot1 and press ENTER. Select ON, the type of plot, and the location of the data (if the data are in L1, press 2nd 1). Set Freq to 1 (or choose the column where you have entered the frequencies of the data).



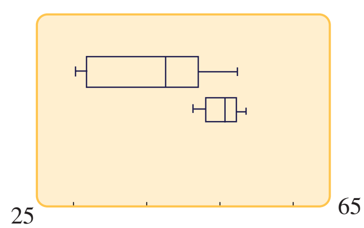


Step 3 Obtain the Graph(s)

To obtain the graph, press GRAPH, but first set the WINDOW. For a *histogram,* set Xscale to correspond to the desired width of each interval. For a *box plot*, choose Xmin and Xmax so that they encompass the range of the data.



Histogram



Box plot

For normally distributed data, the Empirical Rule gives the proportions of the data within one, two, or three standard deviations from the mean. To find the proportion in other intervals, we use a graphing calculator.

On the TI-83, go to the distribution menu `DISTR` by selecting `2nd` `VARS`, then choose `2:normalcdf(` and press `ENTER`. To find the proportion of the population between x_1 and x_2 we enter

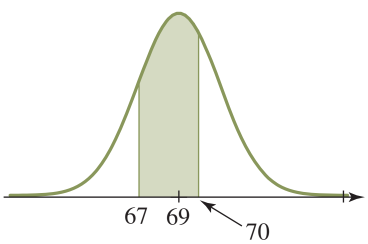
$$\texttt{normalcdf}(x_1,\ x_2,\ \mu,\ \sigma)$$

where μ and σ are the mean and standard deviation of the population.

EXAMPLE 4 | Using the Normal Distribution (Calculator)

The heights of adult males in the United States are normally distributed with mean 69 inches and standard deviation 2.8 inches. Find the proportion of the adult male population with height in the given interval. Also find the probability that a randomly selected individual has height in the given interval.

(a) Between 67 in. and 70 in. **(b)** At least 75 in.



SOLUTION The Empirical Rule does not apply to the required intervals, so we use a graphing calculator as described above.

(a) To find the proportion between 67 and 70 on the TI-83, we use the command

```
normalcdf(67, 70, 69, 2.8)
```

The result is 0.4019823224. We conclude that about 40% of the population is in the given range. The probability that a randomly selected adult male has height between 67 and 70 inches is 0.40.

(b) In this case the lower bound is $x_1 = 75$ inches, and there is no upper bound. So we enter a very large number for the upper bound, say, $x_2 = 1000$.

```
normalcdf(75, 1000, 69, 2.8)
```

The result is 0.0160622279. We conclude that about 1.6% of the population is in the given range. The probability that a randomly selected adult male has height 75 inches or more is 0.016.

NOW TRY EXERCISE **17** ■

14.6 EXERCISES

CONCEPTS

1. When we construct a *histogram* of data, the data must be divided into ________. To draw the histogram, we label the categories on the ____-axis and erect a rectangle on each category that has height proportional to the number of ________ in that category. A histogram gives a visual representation of how the data are ________.

2. Many real-world data are ________ distributed. For normally distributed data with mean μ and standard deviation σ, approximately ______ % of the data are between $\mu - \sigma$ and $\mu + \sigma$, and approximately ______ % of the data are between $\mu - 2\sigma$ and $\mu + 2\sigma$.

SKILLS

3–4 ■ A data set is given. Draw a histogram of the data, using bins (or intervals) of the given size.

3. Bin length is 10, starting at 0

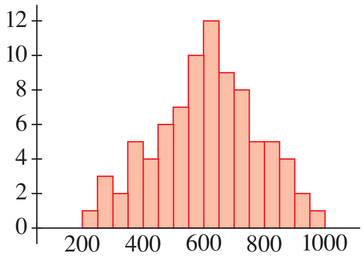
23	17	40	32	24	5	35	18	36	37
31	42	28	50	22	47	61	31	13	27
26	46	48	44	15	28	38	8	52	29

4. Bin length is 1, starting at 0

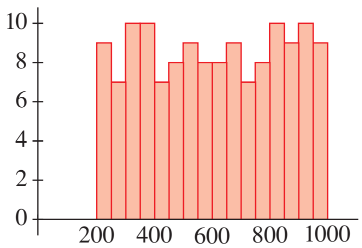
1.2	3.4	5.8	3.2	4.0	1.7	2.3	3.6
3.2	4.7	4.9	6.1	2.3	1.5	3.9	3.8
4.2	5.7	6.1	2.6	3.2	2.1	4.1	5.2
1.1	0.9	2.5	3.3	2.7	5.6	6.2	6.9

5–8 ■ A histogram of data is shown. **(a)** Are the data approximately symmetric, skew right, or skew left? **(b)** Determine the approximate location of the median and the relative location of the mean.

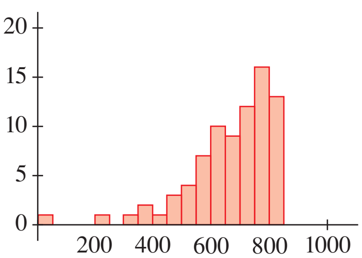
5.



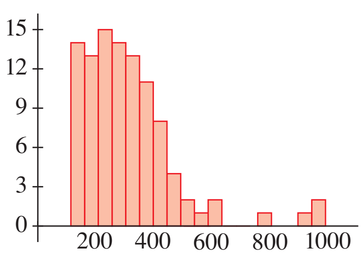
6.



7.



8.



9–12 ■ A data set is given. **(a)** Draw a histogram of the data, using bins (or intervals) of the given size. **(b)** Are the data symmetric, skew, or neither? **(c)** Calculate the median and the mean.

9. Bin length is 50, starting at 0

690	554	571	524	312	91	440
457	568	669	653	591	378	637
622	582	512	201	557	661	613
482	173	597	658	651	517	588
542	632	30	431	656	663	412
673	655	539	602	609	625	663

10. Bin length is 3, starting at 30

36	39	46	44	38	73	37	38
82	54	37	43	39	49	41	42
36	53	40	39	38	51	42	37
41	45	47	42	37	39	40	61

11. Bin length is 5, starting at 30

61	83	72	48	60	66	82	76	46	62
51	72	63	67	88	41	56	66	58	37
64	47	71	69	51	72	52	56	63	54
73	56	68	49	70	55	63	51	76	57
60	63	58	69	78	63	68	56	65	69

12. Bin length is 25, starting at 50

189	346	245	338	298	159	349	127
305	220	229	235	247	162	107	358
182	217	193	244	258	290	202	129
270	184	180	270	284	115	131	320
152	213	112	373	219	257	248	133
348	374	321	296	157	283	310	259

13–16 ■ A data set is normally distributed with mean 35 and standard deviation 9. Use the Empirical Rule to find the proportion of data points that lie in the given interval.

13. Between 26 and 44

14. Between 17 and 53

15. At most 17

16. At least 53

17–20 ■ A data set is normally distributed with mean 35 and standard deviation 9. Use a graphing calculator to find the proportion of data points that lie in the given interval.

17. Between 29 and 38

18. Between 15 and 40

19. At least 32

20. At most 21

APPLICATIONS

21. Multiple Sclerosis Multiple sclerosis (MS) is one of the most common neurological disorders causing disability in young adults. The list gives the age of onset for a sample of MS patients. Draw a histogram of the data, using bins of size 2 starting at 0.

23	52	32	27	28	30	13	37
33	29	25	31	26	30	5	29
41	27	30	34	28	19	31	29
32	28	33	28	31	29	29	30

22. Monarch Butterfly Monarch butterflies migrate thousands of miles each year to overwinter in a small mountainous area of Mexico. The list shows the wingspans (in centimeters) of a sample of monarch butterflies captured in their wintering grounds. Draw a histogram of the data, using bins of size 0.5 starting at 8.0.

8.6	10.9	9.8	12.4	10.6	10.7	10.9
10.0	10.2	10.3	9.4	9.1	11.2	9.7
11.6	12.0	10.4	9.5	10.9	10.5	9.9
11.1	10.5	11.3	11.7	11.3	10.3	11.5

23. Los Angeles Lakers Salaries Kobe Bryant of the Los Angeles Lakers basketball team earned a salary of 24.8 million dollars for the 2010–2011 season. The list gives the salaries (in millions of dollars) of each team member of the Los Angeles Lakers for the 2010–2011 season.

(a) Draw a histogram of the data, using bins (or intervals) of size 2 starting at 0.

(b) Are the data symmetric, skew, or neither?

(c) Calculate the median and the mean.

5.26	24.8	6.3	5.5	0.854	1.8	4.0
0.474	17.8	2.2	13.7	3.7	0.474	8.2

24. Hotel Room Rates The list gives the average daily room rate for various hotels in the San Francisco area.

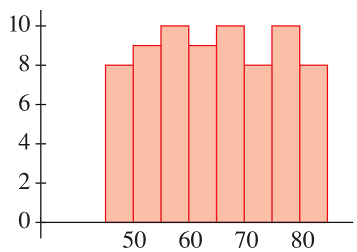
(a) Draw a histogram of the data, using bins (or intervals) of size 25, starting at 0.

(b) Are the data symmetric, skew, or neither?

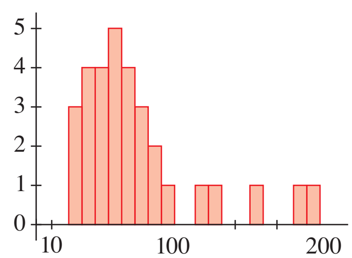
(c) Calculate the median and the mean.

129	192	105	89	152	159	227	111
159	185	179	90	109	59	169	109
215	87	119	69	94	109	118	355
127	59	160	67	149	188	269	75
99	94	242	129	159	215	155	475

25. Starting Salaries Graduates earning bachelor's degrees in engineering are surveyed about their starting salary. A histogram of the data is shown below, with the x-axis representing salaries in thousands of dollars. Are the data symmetric, skew, or neither? Determine the approximate location of the median and determine the relative location of the mean.



26. Savings A recent survey was taken to estimate the amount of savings for employees in a large advertising firm. A histogram of the data is shown below, with the x-axis representing savings in thousands of dollars. Are the data symmetric, skew, or neither? Determine the approximate location of the median and determine the relative location of the mean.



27. Cap Size A men's baseball cap is designed to fit heads with a circumference measuring from 21.25 in. to 24.25 in. The head circumferences of men are normally distributed with mean 22.75 in. and standard deviation 0.75 in. Use the Empirical Rule to find the proportion of men who will be able to wear this design of baseball cap.

28. Airline Seatbelts The seatbelts on a particular airplane are designed to fit people with waist sizes measuring up to 50 in. The waist size of the adult male population is normally distributed with mean 40.5 in. and standard deviation 4.75 in. Use the Empirical Rule to find the proportion of the adult male population who do not fit the airline's seatbelts.

29. Fashion Model The heights of adult females in the United States are normally distributed with mean 63.5 in. and standard deviation 2.5 in. Most fashion models have a height of 69 in. or more. Use a graphing calculator to find the proportion of women in the United States who are 69 in. or taller.

30. Rabbit Gestation The length of pregnancy (gestational period) of rabbits is normally distributed with mean 31.7 days and standard deviation 1.1 days. Use a graphing calculator to find the proportion of rabbit gestations that lie in the given interval.

(a) Between 30 and 33 days

(b) At most 34 days

31. Birth Weight Birth weight is a major determinant of infant mortality in the first year of life. The World Health Organization defines a low birth weight to be less than 2500 grams.

(a) The birth weights of infants born in the United States are normally distributed with mean 3250 grams and standard deviation 450 grams. Use a graphing calculator to find the probability that a randomly selected infant in the United States has a low birth weight.

(b) The birth weights of infants born in a particular country are normally distributed with mean 3000 grams and standard deviation 580 grams. Use a graphing calculator to find the probability that a randomly selected infant in this country has a low birth weight.

32. SAT Scores The SAT math scores of college applicants are normally distributed with mean 515 and standard deviation 100.

(a) A requirement for admission into a particular Ivy League university is an SAT math score of at least 690. What is the probability that a randomly selected college applicant has an SAT math score that meets the requirements for admission into that university?

(b) A requirement for admission into a particular state university is an SAT math score of at least 550. What is the probability that a randomly selected college applicant has an SAT math score that meets the requirements for admission into that university?

DISCOVERY PROJECT **Bias in Presenting Data**

In this project we learn to avoid misleading ways of presenting and interpreting data. You can find the project at the book companion website: **www.stewartmath.com**

14.7 INTRODUCTION TO STATISTICAL THINKING

The Key Role of Randomness ▶ Design of Experiments ▶ Sample Size and Margin of Error ▶ Two-Variable Data and Correlation

Statistical thinking:
A chef tastes a small sample of a stew to determine the quality of the entire stew.

Statistical pitfall:
A student decides that all cars of a certain model are defective on the basis of the problems he's having with his own car.

In **statistical thinking** we make judgments about an entire population based on a sample. A **population** consists of *all* the individuals in a group. A **sample** is any subset of the population. Everyone uses statistical thinking in everyday life but sometimes incorrectly. In this section we investigate key features of statistical thinking as well as some common **statistical pitfalls**.

Why do we use samples? It may appear that a **census** (that is, data obtained from the *entire* population) is a better guarantee of correct results. But a census is not necessarily more accurate. Imagine attempting to survey each of the over 300 million individuals in the United States. This would be an enormously costly and time-consuming task, and numerous data errors would inevitably occur. A carefully selected sample would most likely give more accurate results.

Data are collected in many contexts. In **survey sampling**, data are collected through responses to questionnaires. In **observational studies**, data are collected by observation (such as when bird-watchers make their traditional Christmas bird count). In **experimental studies**, data are collected by performing experiments.

▼ The Key Role of Randomness

zentilia/Shutterstock.com

The lottery method:
Balls are labeled with the lottery numbers. The balls are placed in a sphere and subjected to forced air. Any ball that is expelled from the sphere is considered randomly selected.

A **random sample** is one that is selected arbitrarily and without bias. In selecting a sample, the importance of randomness cannot be overemphasized. In statistical thinking *we expect a random sample to share approximately the same properties as the population.* Also, the larger the sample size, the more closely the sample properties approximate those of the population.

For example, to estimate the average height of adult males, it would be highly biased to select a sample from among NBA players. In general, **non-random samples** are generated when there is bias in the selection process. Such samples are **useless** for statistical purposes, because the sample is not representative of the population.

A **simple random sample** is one in which every individual in the population has the same probability of being selected. To satisfy this requirement, the sampling method that is used must be free of bias with respect to the property being measured. The "lottery method" of selection appears to produce a simple random sample (see the margin). **Computer-generated random numbers** (also called **pseudo-random numbers**) can be used in selecting random samples: We first assign a number to each individual in the population and then use a computer random number generator to select from the assigned numbers. The following are common types of sampling bias.

1. **Undercoverage bias** (or **exclusion bias**), in which part of the population is excluded from the sampling process.
2. **Response bias**, in which the wording of a questionnaire is not neutral but rather suggests or provokes a particular response.
3. **Nonresponse bias**, in which individuals with a common characteristic are unwilling (or neglect) to respond to a questionnaire. (Notice that this is not the opposite of response bias.)
4. **Self-selection bias** (or **voluntary response bias**), in which individuals select themselves (or volunteer) for the sample.

Many of these biases are a result of **convenience sampling**, in which individuals are sampled only because they are nearby or easily accessible.

AP Photo/Byron Rollins

"DEWEY DEFEATS TRUMAN"
This banner headline appeared in the *Chicago Tribune* on November 3, 1948, the morning after Election Day. The headline was wrong. In fact, Truman defeated Dewey by more than two million votes in the popular count. The conclusion of the *Chicago Tribune* was partly based on a telephone poll (a convenience sample). In 1948 few people had telephones; only the relatively wealthy could afford them. So the telephone poll suffered from undercoverage bias. A lot was learned from this embarrassing gaffe. Today, polling companies spend millions of dollars researching methods of selecting random samples. The methods they use are often kept as trade secrets.

EXAMPLE 1 | Bias in Sampling

A sample is selected as described. Is the sample a random sample? If not, identify the type of sampling bias.

(a) A political blog conducts an online poll to gauge the approval rating of a candidate for political office.

(b) To investigate people's preference for different brands of toothpaste, a polling firm selects a random sample of telephone numbers from the telephone directories of several large cities.

(c) To investigate the amount of energy drinks consumed by college freshmen, a student stands on a campus sidewalk and interviews willing participants.

(d) A mail survey asks people's opinion on the preservation of wilderness areas. Only a small percentage of the surveys were completed and returned, and those mostly favored an increase in wilderness areas.

(e) A survey designed by residents opposed to building a new mall in their neighborhood asks: "Do you support the building of the mall which will increase traffic and lower property values?"

SOLUTION

(a) The respondents select themselves to participate in the survey. So the sample suffers from a self-selection bias. The sample also suffers from undercoverage bias, since only visitors to the blog can participate.

(b) Many teenagers and an increasing number of adults rely solely on cell phones, which are not listed in telephone directories. So this sampling technique suffers from undercoverage bias, since teenagers would probably be underrepresented in the sample.

(c) This is a convenience sample. The student is surveying freshmen who are easily accessible. The sampling technique suffers from undercoverage bias (since not all freshmen are equally likely to pass by that particular sidewalk) as well as self-selection bias (since the surveyor depends on participants who stop by for the interview).

(d) It's possible that there is nonresponse bias. People who are not interested in the wilderness issue simply did not respond to the survey.

(e) This survey question is clearly designed to obtain a No response, so this sampling technique suffers from response bias.

NOW TRY EXERCISE **7**

In practice, it is difficult to ensure the selection of a random sample. Several methods have been developed for sampling in particular situations. Some of the most common are as follows.

1. **Systematic Sampling:** In systematic sampling, a sample is chosen systematically from a list. For example, we may pick every 100th name in a telephone book.
2. **Stratified Sampling:** In stratified sampling, the population is first divided into nonoverlapping groups (or strata), and then the sample is chosen proportionally from each group. For example, to pick a sample of registered voters, we may want to stratify the population into groups—white, African American, Hispanic, and other—and then randomly pick registered voters, choosing from each group a number proportional to the size of the group in the population.
3. **Cluster Sampling:** In cluster sampling, the population is divided into groups (or clusters) and then a random sample of *clusters* is selected. For example, to survey apartment dwellers in Los Angeles, we would first randomly select a collection of apartment *buildings* (the clusters) and then interview every resident in the selected buildings. This type of sampling reduces the enormous time and cost for the pollster in traveling from apartment to apartment.

EXAMPLE 2 | Sampling Methods

A sampling scenario is given. Classify the sampling method as systematic sampling, stratified sampling, or cluster sampling.

(a) To investigate student attitudes about the Big Bang Theory, a researcher first partitions students into class categories—freshmen, sophomores, juniors, and seniors—and then selects the random sample proportionally from each group.

(b) To gauge voter attitudes on an upcoming ballot measure, a pollster selects a sample by starting at an arbitrary name in a list of registered voters and then selecting every fiftieth voter on the list.

(c) To investigate college administrators' attitudes on grading practices, a researcher randomly selects a sample of universities, then interviews all top-level administrators at the selected universities.

SOLUTION

(a) This is stratified sampling. The strata from which the sample is selected are the class categories.

(b) This is systematic sampling since a sample is selected systematically (every fiftieth voter) from the voter list.

(c) This is cluster sampling. The clusters are the universities.

NOW TRY EXERCISE **13** ■

▼ Design of Experiments

In observational studies, the researcher has no control over the factors affecting the property being studied—the researcher is merely an observer. Extraneous or unintended variables that systematically affect the property being studied are called **confounding variables** (or **confounding factors** or **lurking variables**). Such variables are said to **confound** (or mix up) the results of the study. The following examples show how this can happen.

EXAMPLE 3 | Confounding Variables

Describe some possible confounding variables in each situation.

(a) To discover any relationship between heavy drinking and lung cancer, a researcher selects a random sample of heavy drinkers.

(b) To test a gasoline additive that claims to increase gas mileage, a researcher solicits volunteers to test the product in their cars.

(c) A new method of teaching mathematics is advertised to increase student scores on standardized tests. To test this claim, a researcher selects the mathematics classes at a local private school.

SOLUTION

(a) A possible confounding variable here is smoking. Many people who drink excessively also smoke. So the study may inadvertently measure the effects of smoking, not drinking.

(b) A possible confounding variable is the mechanical condition of the cars in the sample. The volunteers may own older or defective cars in which the additive is not effective.

(c) The students' mathematical aptitude (before the experiment) may be a confounding variable. If the students in the study normally score high on standardized tests, they probably would score high on the next standardized test with or without the new teaching method. The instructor's teaching skill may also be a confounding variable.

NOW TRY EXERCISE **17(a)** ■

Leah-Anne Thompson/Shutterstock.com

Testing the Salk Vaccine

From 1916 to the 1950s a polio epidemic was sweeping the United States. The crippling disease affected millions of Americans. In the 1950s Jonas Salk developed a polio vaccine. The vaccine had to be tested before it could be approved for general use. Several experimental designs involving first, second, and third graders were suggested. To test the vaccine on children, parental approval had to be obtained. One of the suggested experimental designs was to use the children with parental approval as the treatment group and the others as the control group. But it was observed that this would confound the study because parents who gave approval tended to be more educated (and possibly more hygienic). The actual design involved treatment and control groups randomly selected from the children with parental approval. The study was double-blinded, and placebos were given to the control group. The experimental study involved many thousands of children. Statistical analysis of the results showed that the vaccine was highly effective. Subsequently, the general use of the Salk vaccine practically wiped out the disease in the United States.

To eliminate or vastly reduce the effects of confounding variables, researchers often conduct *experiments* so that such variables can be *controlled.* In an **experimental study,** two groups are selected: a **treatment group** (in which individuals are given a treatment) and a **control group** (in which individuals are not given the treatment). The individuals in the experiment are called **subjects** (or **experimental units**). The goal is to measure the **response** of the subjects to the treatment—that is, whether or not the treatment has an effect. The next step is to make sure that the two groups are as similar as possible except for the treatment. If the two groups are alike except for the treatment, then any statistical difference in response between the groups can be confidently attributed to the treatment. Here are some typical experimental designs.

1. **Completely Randomized Design:** The effects of unknown variables that may confound the experiment can be reduced or eliminated by randomization, that is, by assigning individuals randomly to the treatment or control groups. Randomization ensures that the effects of any confounding variables are equally likely to occur in either group. So any difference between the two groups in the response to the treatment can be attributed to the treatment.
2. **Randomized Block Design:** Variables that are known (prior to the experiment) to affect the response can be controlled by **blocking**. The participants are arranged into **blocks** (or groups) consisting of subjects with similar characteristics, and then treatment and control groups are randomly selected within each block. For example, if sex is a known source of variability in response, then a male block and a female block are formed. Treatment and control subjects are then randomly selected from the male block and from the female block. This design specifically "controls" for a known confounding factor by ensuring equal participation of individuals from each block in the treatment and control groups.
3. **Matched Pair Design:** In this design, the subjects are matched in pairs based on variables that may affect the response to the treatment. For example, the subjects of the study may be matched in pairs based on age and sex (a young male with a young male, a senior female with a senior female, and so on). Then an individual from each pair is randomly assigned to the treatment group, and the other is assigned to the control group. In this example, the treatment and control groups have equal participation with respect to two possible confounding variables (age and sex).

A common confounding factor is the **placebo effect**, in which patients who *think* they are receiving a medication report an improvement (perceived or actual), even though the "treatment" they received was a **placebo**—a simulated or false treatment (sometimes called a "sugar pill"). To control the placebo effect, a researcher may use **single-blinding**, a method in which subjects don't know whether they are in the treatment or control group, or **double-blinding**, in which the researchers are also not privy to this information during the course of the experiment. **Replication**, the repetition of the experiment, can also reinforce the reliability of the results.

EXAMPLE 4 | Designing an Experiment

To test the effectiveness of a new flu vaccine a researcher conducts a study with 800 individuals: 500 women and 300 men. The subjects are to be assigned to treatment and control groups. Suggest different experimental designs for this study.

SOLUTION We suggest three possible designs. In each design, the subjects in the treatment group are given the vaccine, and those in the control group are given a placebo. To further unbias the procedures in the experiment, we suggest double-blinding.

Completely Randomized Design: We randomly assign 400 patients to the treatment group and 400 to the control group. (It is convenient but not necessary to have the same number of subjects in each group.) We can assign an ID number to each participant and then use a computer random-number generator to select the two groups.

Randomized Block Design: It is known that the sex of an individual frequently affects the response to medication (men and women often have different physiological responses to medication). To control for this factor, we use blocking: We put men and women in separate blocks. Subjects from each block are randomly assigned to the treatment or control group. So each group would contain 250 women and 150 men.

Matched Pair Design: It is also known that the age of an individual may affect the response to medication. So in this design, we place participants in pairs matched by age and sex. To do this, we may want to decide on age categories. Then each pair would consist of individuals of the same sex and the same age category. From each pair we pick one for the treatment group and one for the control group. (We can do this by flipping a coin.)

NOW TRY EXERCISE **17(b)** ■

▼ Sample Size and Margin of Error

One of the most important features of statistical thinking is the realization that statistics does not provide exact answers. Statisticians *estimate* properties of a population by examining a random sample from that population. Statistical conclusions are based on probability and are always accompanied by a confidence level. The 95% **confidence level** means that there is less than a 5% chance (or 0.05 probability) that the result obtained from the sample could be obtained by chance alone. In the popular press, poll results are accompanied by a *margin of error.* For example, a magazine article states:

> "A recent poll shows that Candidate X has 57% of the vote. The poll was conducted among a random sample of 600 registered voters. The margin of error of the poll is $\pm 3\%$."

This means that 57% of the voters in the *sample* prefer Candidate X. But how does this compare with the true proportion that prefer Candidate X in the entire *population*? Traditionally, poll results are stated at the 95% confidence level. So the statement in the magazine means that there is a 95% chance (or 0.95 probability) that the true proportion is between $57 - 3 = 54\%$ and $57 + 3 = 60\%$. That is, if we conduct this poll many times, we would expect the true proportion to be contained in the intervals we get about 95% of the time.

Each time the poll is conducted, a new value for p_0 and a new interval $(p_0 - d, p_0 + d)$ are obtained. At the 95% confidence level we would expect 95% of the intervals so constructed to contain the true proportion p. (Stated another way, we would expect only 5% of the intervals constructed in this way to fail to contain p.)

In general, the proportion p_0 of individuals with a particular characteristic in a sample is an *estimate* of the true proportion p of such individuals in the population. It is clear that the larger the sample size, the more accurate the estimate is likely to be. The **margin of error** of the estimate associated with the sample is a number d satisfying the following:

> There is a 95% probability that the interval $(p_0 - d, p_0 + d)$ contains the true proportion p.

Notice that the true proportion p is a fixed number that we do not know. The random sample allows us to construct an interval that is very likely (probability 0.95) to contain p. The interval $(p_0 - d, p_0 + d)$ is called the 95% **confidence interval**. The following formula allows us to find the margin of error associated with a given sample size.

MARGIN OF ERROR AND SAMPLE SIZE

At the 95% confidence level, the margin of error d and the sample size n are related by the formula*

$$d = \frac{1.96}{2\sqrt{n}}$$

*The formula is derived by using calculus and relies on properties of the binomial and normal distributions. The number 1.96 in the formula is a slightly more precise form of the second statement in the Empirical Rule (page 938): For a normal distribution 95% of the data is between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

EXAMPLE 5 | Margin of Error and Sample Size

A political poll is conducted to gauge the approval rating of Candidate X. A random sample of 600 voters is polled. The poll indicates that 324 of those surveyed approve of the candidate.

(a) Estimate the percentage of the voters in the population who prefer Candidate X.

(b) Find the margin of error of the poll (at the 95% confidence level).

(c) State the 95% confidence interval for the percentage of the population who prefer Candidate X.

SOLUTION

(a) The proportion of the sample who prefer Candidate X is

$$\frac{324}{600} = 0.54$$

So we estimate that 54% of the voter population prefer Candidate X.

(b) Since $n = 600$, we have

$$d = \frac{1.96}{2\sqrt{n}} = \frac{1.96}{2\sqrt{600}} \approx 0.04$$

So the margin of error of the poll is $\pm 4\%$.

(c) The 95% confidence interval for the proportion who prefer Candidate X is

$$(0.54 - 0.04, 0.54 + 0.04) = (0.50, 0.58)$$

So we can say with 95% confidence that between 50% and 58% of voters prefer Candidate X.

NOW TRY EXERCISE **21**

EXAMPLE 6 | Margin of Error and Sample Size

A political polling firm is hired to conduct a poll for Candidate X. Determine the required sample size if the candidate would like the poll to have the following margin of error.

(a) $\pm 3\%$ **(b)** $\pm 1\%$

SOLUTION

(a) We set $d = 0.03$ in the formula for margin of eror and solve for n:

$$0.03 = \frac{1.96}{2\sqrt{n}} \qquad \text{Set } d = 0.03 \text{ in formula}$$

$$n \approx 1067.1 \qquad \text{Solve for } n$$

So a random sample of 1067 or greater would provide a margin of error of $\pm 3\%$ or better.

(b) We set $d = 0.01$ in the formula for margin of error and solve for n:

From Example 6 we see that a very small margin of error requires a very large sample.

$$0.01 = \frac{1.96}{2\sqrt{n}} \qquad \text{Set } d = 0.03 \text{ in formula}$$

$$n \approx 9604 \qquad \text{Solve for } n$$

So a random sample of 9604 or greater would provide a margin of error of $\pm 1\%$ or better.

NOW TRY EXERCISE **25**

▼ Two-Variable Data and Correlation

So far in our study of statistics we've learned about **one-variable** data, in which only one property of each individual is measured. For example, we can measure the height of individuals. We now consider **two-variable** data, in which two properties of each individual are measured. For example, we can record the height *and* salary of each individual. In studying two-variable data, our goal is to discover whether there is a linear relationship between the two variables.

Two-variable data can be graphed in a coordinate plane, resulting in a **scatter plot**. In *Focus on Modeling,* page 130, we analyzed such data mathematically by finding the line that best fits the data, called the **regression line**. Associated with the regression line is a **correlation coefficient**, which is a mathematical measure of how well the data fit along the regression line, or how well the two variables are **correlated**. The correlation coefficient r satisfies

A positive correlation between height and salary means that taller people tend to have bigger salaries. A negative correlation would indicate that taller people tend to have smaller salaries.

$$-1 \le r \le 1$$

If r is close to zero, the variables have little correlation. The closer r is to 1 or -1, the closer the data points are to the regression line. (The slope of the regression line determines the sign of the correlation coefficient.)

Now let's think about these issues statistically. In statistics the data are a *sample* from a *population.* Our goal is to discover the relationship between the two variables in the population by examining the sample. So if we find that the sample shows a strong correlation between the two variables, we would intuitively assume that the same is true in the population. But there is another variable that is always involved in statistical conclusions: sample size. In statistics we think as follows. We first skeptically assume that these two variables are not correlated at all in the population and then ask:

What is the probability that the correlation in the sample is due to chance alone?

If this probability is small (less than 0.05), then we say that the correlation is **statistically significant** (at the 95% **confidence level**). (Assuming that the data are normally distributed, the value of this probability can be calculated precisely by using calculus.) For example, if the sample consists of only three individuals, even a strong correlation coefficient may not be significant. On the other hand, for a large sample, a small correlation coefficient may be significant. This is because if there is no correlation at all in the population, it's very unlikely that a large random sample would produce data that have a linear trend, whereas a small sample is more likely to produce correlated data by chance alone.

Table 1 relates the sample size with critical values of the correlation coefficient, at the 95% confidence level. For example, for a sample of size 10 we need a correlation coefficient greater than 0.632 to conclude that there is a statistically significant linear relationship between the two variables at the 95% confidence level. This means that if there is actually no correlation at all between the variables in the population, there is only a 0.05 probability of obtaining a sample of 10 with a correlation coefficient greater than 0.632.

TABLE 1

Critical values of the correlation coefficient r for sample size n (at the 95% confidence level)

n	r	n	r	n	r
3	0.997	11	0.602	35	0.334
4	0.950	12	0.576	40	0.312
5	0.878	13	0.553	50	0.279
6	0.811	14	0.532	60	0.254
7	0.754	15	0.514	70	0.235
8	0.707	20	0.444	80	0.220
9	0.666	25	0.396	90	0.207
10	0.632	30	0.361	100	0.197

TABLE 2

Study hours	Blood pressure
0	129
1	120
3	130
4	128
6	132

EXAMPLE 7 | Is the Correlation Statistically Significant?

A student wants to test the relationship between study time and blood pressure. Table 2 gives the data the student obtained.

(a) Find the correlation coefficient.

(b) Are the results of the study statistically significant? Explain how your conclusion is supported by the information in Table 1.

SOLUTION

Finding the regression line on a graphing calculator is explained on pages 130–135.

(a) We use a graphing calculator to find the regression line. The calculator also gives the correlation coefficient $r = 0.564$.

(b) The sample contains only five data points. According to Table 1 the correlation coefficient is not statistically significant (it is less than 0.878). This means that there is a greater than 0.05 probability that the observed correlation is due to chance alone.

NOW TRY EXERCISE **33**

EXAMPLE 8 | Is the Correlation Statistically Significant?

When laboratory rats are exposed to asbestos fibers, some of them develop lung tumors. Table 3 gives the data from several different experiments. (See Example 3, page 133.)

(a) Find the correlation coefficient.

(b) Are the results of the study statistically significant? Explain how your conclusion is supported by the information in Table 1.

TABLE 3

Asbestos exposure (fibers/mL)	Percent that develop tumors
50	2
400	6
500	5
900	10
1100	26
1600	42
1800	37
2000	28
3000	50

SOLUTION

(a) On page 134 we found the regression line for these data. The calculator also gives the correlation coefficient $r = 0.924$.

(b) The sample contains nine data points. According to Table 1 this correlation coefficient is statistically significant (because it is greater than 0.666). This means that there is less than a 0.05 probability that the observed correlation is due to chance alone.

NOW TRY EXERCISE **35**

Is Correlation the Same as Causation? The answer is no. For example, there is a very strong correlation between the number of grocery stores in a city and the number of robberies in that city. Can we conclude from this that grocery stores cause robberies? Of course not—population is a lurking variable in this situation. Larger cities have more of everything, including grocery stores and robberies. Nevertheless, correlation is an excellent statistical tool for discovering hidden relationships. For example, the discovery of a statistically significant correlation between smoking and lung cancer led scientists to search for the causal relationship between these variables. You can learn much more about this topic in the *Discovery Project* referenced on page 953.

14.7 EXERCISES

CONCEPTS

1. When we analyze a population statistically by performing a survey or making an observational study, it is important that we select a __________ sample, to avoid bias.

2. In selecting a sample for a statistical study, a researcher must be careful to avoid the following types of bias.

(a) If part of the population is excluded from the sampling process, the sample will suffer from __________ bias.

(b) If the wording of the questions in a survey are intended to elicit a specific response, the sample will suffer from __________ bias.

(c) If some members of the sample are inclined not to answer some of the questions in a survey, the sample will suffer from __________ bias.

(d) If the sample consists of individuals who have volunteered to take part in the survey, or study because of their personal interest, the sample will suffer from __________ bias.

3. Since a simple random sample is often difficult or costly to select, statisticians have developed other sampling methods.

(a) If we obtain a sample systematically from a list, we are conducting __________ sampling.

(b) If we first divide the population into groups and then randomly select a sample proportionally from each group, we are conducting __________ sampling.

(c) If we first divide the population into groups and then select a random sample of groups, we are conducting __________ sampling.

4. In an observational study (such as one that tests the effectiveness of an experimental drug), a researcher will usually divide the sample into two groups: One group receives the treatment being tested and is called the __________ group; the other group does not receive the treatment and is called the __________ group. This is done to ensure that the treatment is the actual cause of any positive results obtained and to eliminate the effect of unintended __________ variables.

5. When a poll is conducted, then at the 95% confidence level the margin of error d and the sample size n are related by the formula $d =$ __________.

6. When we analyze two-variable data by finding a regression line, then both the __________ coefficient r and the __________ size n are important in determining whether the correlation is statistically __________.

APPLICATIONS

7–12 ■ A sample is selected as described. Is the sample a random sample? If it isn't, identify the type of sampling bias.

7. Athletic Program A middle school wishes to determine how important an athletic program is to the parents of its students. It mails a survey to all 230 parents whose children are enrolled at the school and asks them whether they consider the program "important," "somewhat important," or "not important." They receive 37 responses, of which most consider the athletic program "important."

8. Cosmetic Product The Exfolique Cosmetics Company wishes to demonstrate how effective its products are in improving women's skin health. The company places surveyors at its cosmetics counters in department stores who question the first ten clients who show up at each counter about how satisfied they are with the effect of Exfolique products on their skin.

9. Call-In Poll A news-talk radio station for a certain large city conducts a call-in poll on the popularity of a mayoral candidate. The sample selected consists of listeners who call in to the radio station.

10. Study Hours To estimate the average number of hours freshmen study each week, a sample of freshmen from the university is selected as follows. The ID numbers of all freshmen are entered into a computer, and a random-number generator is used to randomly select 50 of those ID numbers. Students with the selected ID numbers are interviewed about their study habits.

11. Impaired Driving Survey A Pennsylvania county offers first-time drunk driving offenders the option of completing a six-week educational program in lieu of jail time after conviction. To measure the effectiveness of the program, they interview 200 graduates of the program and ask them about their impaired driving behavior (or lack of it) after completion of the program.

12. Food Waste Nanaimo Regional District officials have decided to add food waste to the materials that residents must recycle as part of their waste management program. The new system is criticized in the local press, so the district management mails surveys to all homeowners on its tax rolls explaining the many advantages of the program and asking whether the program should be continued. The responses to the survey are overwhelmingly supportive of the program.

13–16 ■ A sampling technique is described. Identify the type of sampling used.

13. Gypsy Moths After gypsy moths infest the trees in a large suburban park, officials decide to measure the degree of the infestation. Using a topographical map, park rangers divide the park into a grid of 1000 equally sized blocks. They then select 20 blocks at random; from each of these blocks they randomly select 15 trees and observe the degree of gypsy moth infestation in each of the selected trees.

14. Congressional Candidates An election in a swing congressional district pits two highly respected candidates against each other. A pollster obtains the list of registered voters for the district and interviews every 500th person on the list as to the person's preference to determine who is ahead in the election.

15. Customer Satisfaction A financial services company wishes to study customer satisfaction. To select a sample for the study, the market research branch of the company first classifies the customers into four groups: small, medium, large, and preferred investors. A random sample of customers is selected by randomly selecting from each group a number of customers that is proportional to the size of the class group.

16. Clean-Up Crew On a camping trip with 200 campers, some are selected for clean-up as follows. For each camper a die is rolled. The campers who get the numbers 1 and 2 are selected for the clean-up crew.

17–20 ■ An experiment is described. **(a)** Describe some possible confounding variable(s) in the experiment. **(b)** Suggest an experimental design that would eliminate the confounding variable(s). Identify the type of design you are suggesting.

17. Synthetic Oil Versus Conventional Oil A petrochemical researcher wants to compare the effectiveness of synthetic oil and conventional oil on the durability of automobile engines. His firm purchases ten Acura TLs and ten Skoda Octavias, all of them new 2011 models, for his experiment. The Acuras are lubricated with synthetic oil, and the Skodas are lubricated with conventional oil. After 100,000 km of travel with regular oil changes, the engines are dismantled, and their wear is compared.

18. Exercise and Cancer To determine whether exercise helps to prevent cancer, a medical professor selects 60 sedentary students from his classes and 35 people who exercise regularly at the gym in his uncle's retirement home. He then tracks their medical history for a five-year period, recording the number of cancer cases that arise.

19. Bottled Water and Health The Zxyzx bottled water company conducts a study on the health benefits of drinking their very expensive bottled water. The company selects 50 households that regularly use Zxyzx and 50 households that drink tap water. The study finds that the children in the Zxyzx households have fewer health problems.

20. Restless Arm Syndrome A new medication for restless arm syndrome is to be tested among 400 patients suffering with the ailment. The subjects of the study consist of 300 women and 100 men, ranging in age from 20 to 40 years of age. The medication is administered to the women, whereas the men get a placebo. Their responses are recorded and analyzed.

21–24 ■ To determine the approval rating of a candidate for governor, a random sample of registered voters is selected. The sample size n and the number x who plan to vote for the candidate are given.

(a) Estimate the percentage of the voters in the population of registered voters that prefer the candidate.

(b) Find the margin of error of the poll (at the 95% confidence level).

(c) State the 95% confidence interval for the percentage of the population who prefer the candidate.

21. $n = 400, \quad x = 180$ **22.** $n = 200, \quad x = 80$

23. $n = 1000, \quad x = 600$ **24.** $n = 10, \quad x = 4$

25–28 ■ A polling firm is hired to conduct a poll for Candidate X. The candidate wants the poll to have margin of error d. Find the required sample size n.

25. $d = 0.015$ **26.** $d = 0.025$

27. $d = 0.05$ **28.** $d = 0.005$

29. Tasting Phenylthiocarbamide (PTC) The ability to taste the chemical PTC is genetically determined. For those who can taste it, PTC has a bitter taste; for others it is tasteless. To estimate the proportion of people in a certain Scandinavian country who can taste PTC, a researcher selects a random sample of 1500 individuals. The researcher finds that 1230 of these can taste PTC. Estimate the percentage of people in that country who can taste PTC, and state the margin of error of your estimate.

30. Methylmercury in Fish Fish concentrate mercury in their bodies in the form of methylmercury, a biological form of mercury, which is highly toxic to humans. To estimate the proportion of fish in a particular lake that are contaminated with methylmercury, a researcher obtains a random sample of 600 fish from the lake. The researcher finds that 153 of these are contaminated. Estimate the percentage of fish in the entire lake that are contaminated, and state the margin of error of your estimate.

31. Car Accidents near Home To estimate the number of car accidents that occur "near home" (within 5 miles of home), a researcher obtains a random sample of 8000 accident reports. It is found that 4220 accidents occurred near home. Use these data to estimate the percentage of all accidents that occur near home and state the margin of error.

32. Tomato Seeds A seed company wishes to estimate the proportion of its hybridized tomato seeds that germinate. A random sample of 2200 of the tomato seeds is selected and planted; 2035 of them germinate. Estimate the percentage of the company's tomato seeds that will germinate, and state the margin of error of your estimate.

33–34 ■ Two-variable data are given. **(a)** Use a graphing calculator to find the correlation coefficient r. **(b)** Determine whether the correlation between the variables is statistically significant at the 95% significance level.

33. Video Games A student obtains the following two-variable data from three of his friends. The data give the number of hours per day that each student plays video games and their grade on the last precalculus test. The student claims that playing video games is strongly positively correlated to mathematics test scores.

Hours	Score
1	72
2	70
3	80

34. Too Much Studying A student claims that studying too much lowers your grade in calculus. The student offers the following data for his claim.

Hours of study	Test score
1	80
2	78
3	77
4	88

35–40 ■ Refer to the two-variable data in the indicated exercises on pages 135–139. **(a)** Use a graphing calculator to find the correlation coefficient *r*. **(b)** Determine whether the correlation between the variables is statistically significant at the 95% significance level.

35. Exercise 1

36. Exercise 4

37. Exercise 5

38. Exercise 6

39. Exercise 7

40. Exercise 8

DISCOVERY PROJECT **Correlation and Causation**

In this project we explore the relationship between correlation and causation. You can find the project at the book companion website: **www.stewartmath.com**

14.8 Introduction to Inferential Statistics

Testing a Claim About a Population Proportion Intuitively ▶ Testing a Claim About a Population Proportion Using Probability: The *P*-Value ▶ Inference About Two Proportions

The goal of **inferential statistics** is to get (or infer) information about an entire population by examining a random sample from that population. We can use a sample to test the validity of a hypothesis (or a claim) about the population mean, variance, standard deviation, and other properties of a population. In this introduction to statistical inference we use a random sample from a population to test a claim about a **population proportion**.

▼ Testing a Claim About a Population Proportion Intuitively

In statistics a **hypothesis** is a statement (or a claim) about a property of an entire population. In this section we study hypotheses about the true proportion p of individuals in the population with a particular characteristic: for example, the true proportion of registered voters who will vote Yes in an upcoming ballot or the true proportion of patients who react well to a new medication.

Suppose that individuals having a particular property are *assumed* to form a proportion p_0 of a population. How does this assumption compare to the *true* proportion p of these individuals? To answer this question, we begin by stating two opposing hypotheses.

The **null hypothesis**, denoted by H_0, states the "assumed state of affairs" and is expressed as an equality: $p = p_0$. The **alternative hypothesis** (also called the **research hypothesis**), denoted by H_1, is the proposed substitute to the null hypothesis and is expressed as an inequality: $p > p_0$, or $p < p_0$, or $p \neq p_0$.

To test a hypothesis, we examine a random sample from the population. The claim is either supported or refuted by data from that random sample.

If, under the assumption that the null hypothesis is true, the observed sample proportion (or a more extreme proportion, depending on H_1) is *very unlikely* to have occurred by chance alone, we **reject the null hypothesis**. Otherwise, the data does not provide us with enough evidence to reject the null hypothesis, so we **fail to reject the null hypothesis**. The next example illustrates how these ideas are used.

EXAMPLE 1 | Testing a Hypothesis Intuitively

It is assumed that voter sentiment on a California ballot measure is equally divided between Yes and No. An advocate for the measure claims that more than half of the voters support the measure. To test this claim, a *random* sample of 1000 voters is selected.

(a) Formulate a hypothesis test.

(b) What should we conclude about the claim if the sample has 820 Yes votes? 508 Yes votes? 535 Yes votes?

Photo by Hulton Archive/Getty Images

FLORENCE NIGHTINGALE (1820–1919) was a British nurse who tended to wounded soldiers during the Crimean war. She laid the foundation for the modern science of nursing by using data analysis to investigate the conditions of medical care in Britain. After the war she established the first professional school of nursing.

Florence exhibited a strong interest in mathematics at an early age. As a nurse she used her knowledge of mathematics to collect and exhibit data on sanitary conditions in hospitals. She was a pioneer in the use of graphical representations of data. She was one of the first to use pie charts and invented another method of presenting data graphically now known as Rose Diagrams.

Florence Nightingale was enamored by the powerful uses of statistics. She wrote: "Statistics is the most important science in the whole world, for upon it depends the practical application of every other science and of every art, the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience."

SOLUTION

(a) The null hypothesis is that the proportion of Yes is half: $p = 0.5$. The alternative hypothesis is that the proportion is greater than half: $p > 0.5$. We state these hypotheses as follows:

$$H_0: p = 0.5$$

$$H_1: p > 0.5$$

(b) If the null hypothesis is true, we would expect our random sample of 1000 voters to have about 500 Yes votes.

If the null hypothesis is true, it would be *very unlikely* to have 820 or more Yes votes in a random sample of 1000 voters. This sample provides overwhelming evidence against the null hypothesis. So in this case we reject the null hypothesis.

If the null hypothesis is true, it is *not unlikely* that we would have 508 Yes votes in our sample. So the sample does not provide enough evidence to reject the null hypothesis. So we fail to reject the null hypothesis.

If the null hypothesis is true, the 535 Yes votes in the sample is higher than expected. But it is difficult to decide on the basis of intuition alone whether this sample provides enough evidence to reject the null hypothesis. (We'll see how to answer this question in the next example.)

NOW TRY EXERCISE **3**

Testing a Claim About a Population Proportion Using Probability: The *P*-Value

Example 1 contains the main concept of testing a hypothesis. The only missing ingredient is the calculation of probabilities. In the solution to Example 1 we used the phrases "very unlikely" and "not unlikely." We need to replace these phrases with probabilities.

The ***P*-value** associated with the observed sample is the probability of obtaining a random sample with a proportion at least as extreme (depending on H_1) as the proportion in our random sample, given that H_0 is true. So a "very small" P-value tells us that it is "very unlikely" that the sample we got was obtained by chance alone, so we should reject the null hypothesis. The P-value at which we decide to reject the null hypothesis is called the **significance level** of the test and is denoted by α. Experience among statisticians indicates that the significance level $\alpha = 0.05$ (or 5%) is adequate for most purposes. In critical studies, such as the testing for the effectiveness of drugs, a much more stringent level of significance is used (usually $\alpha = 0.01$).

TESTING A HYPOTHESIS

To test a hypothesis, we use the following steps.

1. **Formulate the Hypotheses.** First state a null hypothesis and an alternative hypothesis.
2. **Obtain a Random Sample.** Obtain a sample that is selected arbitrarily and without bias.
3. **Calculate the *P*-Value.** Use a calculator to find the P-value associated with the sample obtained.
4. **Make a Conclusion.** A P-value smaller than the significance level of the test indicates that we should reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

The logic of hypothesis testing outlined here is the same as the logic used in legal proceedings. In court, the defendant is presumed to be innocent (this is the null hypothesis). Next, the facts of the case (the data) are presented in court. The jury must decide the case on reasonable doubt (corresponding to the P-value). Suppose the defendant is found fleeing the scene in the victim's car with the murder weapon by his side and the victim in the back seat. The jury will conclude that there is only a small chance (small P-value) of this happening if the hypothesis of innocence (the null hypothesis) is true, so the jury decides on guilt (rejects the null hypothesis). On the other hand, if the facts of the case show that the defendant was most likely in another town when the crime occurred and no facts were presented that he ever had access to the murder weapon, the jury may find reasonable doubt (large P-value) and fail to convict the defendant (fail to reject the null hypothesis).

The calculation of P-values is very difficult and requires advanced calculus. Fortunately, the formulas needed to calculate P-values are programmed into most graphing calculators. We now test the hypotheses in Example 1 by calculating P-values.

EXAMPLE 2 | Testing a Hypothesis Using the *P*-Value

It is assumed that voter sentiment on a California ballot measure is equally divided between Yes and No. An advocate for the measure claims that Yes voters are more than half of all voters. To test this claim, a *random* sample of 1000 voters is selected. The sample has 535 Yes voters.

(a) Formulate a hypothesis test.

(b) Calculate the P-value.

(c) Make a conclusion using the $\alpha = 0.05$ significance level.

(d) Explain how the P-value supports your conclusion.

SOLUTION

(a) Let p be the true proportion of Yes voters in the population. As in Example 1, the hypothesis test is formulated as follows:

$$H_0: p = 0.5$$

$$H_1: p > 0.5$$

(b) To find the P-value on a TI-83 calculator, follow the instructions on page 957. On the `1-PropZTest` menu, use $p_0 = 0.5$, $x = 535$, $n = 1000$, and Prop $> p_0$, and then calculate. The calculator output is shown on page 957. We see that, rounded to three decimals, the P-value is 0.013.

If the true proportion of Yes voters is 0.5, then in a random sample of 1000 voters we would expect *about* 500 Yes votes. But we should *not* expect exactly 500 votes. The question is "Is the deviation from 500 (or from what we expect) statistically significant?"

(c) Since the P-value is less than 0.05, we reject the null hypothesis. We conclude that on the basis of this sample, it is very likely that the proportion of Yes voters is greater than one half.

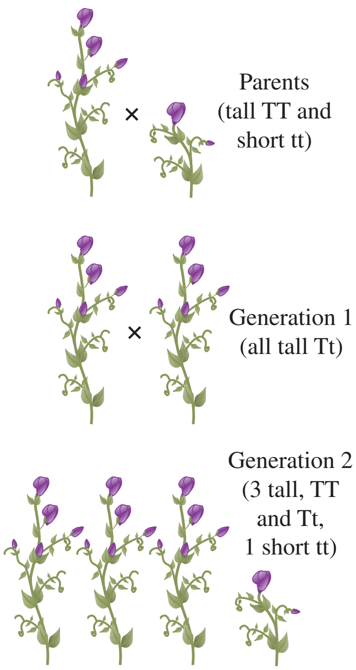
(d) The P-value tells us that there is only a very small chance (probability 0.013) that we would obtain 535 Yes votes in a random sample of 1000 voters, given that the null hypothesis is true (i.e., given that the true proportion of Yes voters is 0.5). That's why we reject the null hypothesis.

NOW TRY EXERCISE **5**

How is the *P*-value calculated? In Example 2 the P-value is the probability that a random sample of 1000 voters would contain 535 or more Yes voters, given that the null hypothesis is true. Since the null hypothesis states that $p = 0.5$, the P-value is the binomial probability of 535 or more "successes" (Yes votes).

$$P = C(1000, 535)(0.5)^{535}(0.5)^{465} + C(1000, 536)(0.5)^{536}(0.5)^{464} + \cdots + C(1000, 1000)(0.5)^{1000}(0.5)^{0}$$

This calculation involves huge numbers. Graphing calculators are programmed with advanced methods to approximate such probabilities. That's why we need the capability of a graphing calculator in Example 2(b).



EXAMPLE 3 | Hypothesis Test (Proportion)

In his pioneering genetic studies, Gregor Mendel experimented with hybridized (tall and short) pea plants. Mendel claimed that the cross of such hybridized plants results in tall plants 25% of the time. A biology student theorizes that the proportion of tall plants should be less than 25%. The student replicates Mendel's experiment and obtains 360 offspring of hybridized pea plants, of which 84 are tall. (See Exercise 53, page 911.)

(a) Formulate a hypothesis test.

(b) Calculate the *P*-value.

(c) Make a conclusion using the $\alpha = 0.05$ significance level.

(d) Explain how the *P*-value supports your conclusion.

SOLUTION

(a) The generally accepted claim is that the true proportion p of tall offspring is 25%, so the null hypothesis is $p = 0.25$. The student theorizes that $p < 0.25$.

$$H_0\colon p = 0.25$$

$$H_1\colon p < 0.25$$

(b) To find the *P*-value on a TI-83 calculator, follow the instructions on page 957. On the `1-PropZTest` menu, use $p_0 = 0.25$, $x = 84$, $n = 360$, and Prop $< p_0$, and then calculate. We find that the *P*-value is 0.394.

(c) The *P*-value is greater than 0.05, so we fail to reject the null hypothesis. We conclude that on the basis of this experiment, it is entirely possible that the true proportion is 0.25.

(d) The *P*-value tells us that there is a high probability (greater than 0.05) that the results of this experiment are due to chance alone, given that the null hypothesis is true. This means that the experiment does not provide enough evidence to refute Mendel's claim (the null hypothesis).

NOW TRY EXERCISE **7**

Is 23.3% the same as 25%?
In Example 3 the proportion of tall peas in the student's sample is 23.3%, but we did not reject the hypothesis that the true proportion is 25%. So is 23.3% the same as 25%? In mathematics the answer is a simple No. But in statistics the situation is quite different. In this example our answer is that we cannot say No on the basis of this sample at the 0.05 significance level. In other words, if the true proportion is 25%, we would expect to obtain the results of this sample more than 5% of the time (in fact, about 39.4% of the time).

It is important to understand that hypothesis tests do not determine truth; they determine only the probability of truth. For example, in the courtroom analogy described earlier, there may be another explanation for the defendant's predicament, and he may actually be innocent. The jury determines only that it is *unlikely* (small *P*-value) that the defendant is innocent given the facts of the case. For this reason we can have only a certain level of confidence in the jury decision. For any hypothesis test, the 5% significance level can also be expressed as a 95% **confidence level** in the conclusion of the test. This indicates that there is only a 1 in 20 chance (5% chance) that we falsely rejected the null hypothesis.

Inference About Two Proportions

Most statistical studies involve a comparison of two groups, usually called the **treatment group** and the **control group**. For example, when researchers are testing the effectiveness of an investigational medication, two groups of patients are selected. The individuals in the treatment group are given the medication; the individuals in the control group are not. The proportions of patients who recover in each group is compared. The goal of the study

STATISTICAL TESTS
On a Graphing Calculator

Graphing calculators are capable of calculating the *P*-values we use in this section. The following steps show how to obtain *P*-values on the TI-83 or TI-84 calculators.

Step 1 Choose the Test

Press the [STAT] key, choose **TESTS**, then select the appropriate test and press [ENTER].

```
EDIT   CALC   TESTS
1: Z-Test...
2: T-Test...
3: 2-Samp ZTest...
4: 2-Samp TTest...
5: 1-Prop ZTest...
6: 2-Prop ZTest...
7↓ZInterval...
```

```
EDIT   CALC   TESTS
1: Z-Test...
2: T-Test...
3: 2-Samp ZTest...
4: 2-Samp TTest...
5: 1-Prop ZTest...
6: 2-Prop ZTest...
7↓ZInterval...
```

Step 2 Enter the Required Information

Enter the information required for the test. The screens for the test of a proportion **1-propZTest** and **2-propZTest** are shown.

```
1-PropZTest
 p0:.5
 x:535
 n:1000
 Prop≠p0  <p0  >p0
 Calculate  Draw
```

```
2-PropZTest
 x1:122
 n1:135
 x2:52
 n2:68
 p1:≠p2  <p2  >p2
 Calculate  Draw
```

Step 3 Obtain Results

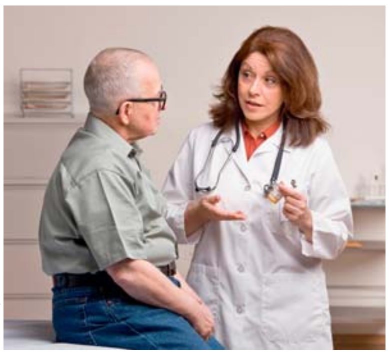
Choose **Calculate**, and press [ENTER]. The screen shows several results. We are interested in the *P*-value.

```
1-PropZTest
 prop>.5
 z=2.213594362
 p=.0134283016
 p̂=.535
 n=1000
```

```
2-PropZTest
 p1<p2
 z=-2.671183538
 p=.0037792605
 p̂1=.7647058824
 p̂2=.9037037037
↓p̂=.8571428571
```

is to determine whether the difference in the proportion that recovers is *statistically significant* (and not just the result of chance alone).

Let p_1 and p_2 be the true proportions of patients who recover in the control and treatment groups, respectively. The null hypothesis is that the medication (the treatment) has no effect: $p_1 = p_2$. The alternative hypothesis is that the treatment does have an effect: $p_1 < p_2$. The *P*-value is the probability that the difference between the proportions in the two samples is due to chance alone. So if the *P*-value is small (less than the significance level of the test), we reject the null hypothesis.


AVAVA /Shutterstock.com

EXAMPLE 4 | Two-Sample Hypothesis Test

In a study of an investigational medication the researcher selects a treatment group and a control group. Of the 68 patients in the control group, 52 recover; and of the 135 patients in the treatment group, 122 recover.

(a) Formulate a hypothesis test.

(b) Calculate the *P*-value.

(c) Make a conclusion using the $\alpha = 0.01$ significance level.

(d) Explain how the *P*-value supports your conclusion.

SOLUTION

(a) Let p_1 be the true proportion of patients who recover without the medication, and let p_2 be the true proportions of patients who recover with the medication. The null hypothesis is that the medication has no effect, that is, $p_1 = p_2$. The alternative hypothesis is that the medication is effective, that is, $p_1 < p_2$.

$$H_0: p_1 = p_2$$

$$H_1: p_1 < p_2$$

(b) To find the *P*-value on a TI-83 calculator, follow the instructions on page 957. On the `2-PropZTest` menu, use $x_1 = 52$, $n_1 = 68$, $x_2 = 122$, $n_2 = 135$, and $p_1 < p_2$, and then calculate. The calculator output is shown on page 957. We see that, rounded to four decimals, the *P*-value is 0.0038.

(c) Since the *P*-value is less than 0.01, we reject the null hypothesis. We conclude that on the basis of this study, the medication is very likely to be effective.

(d) The *P*-value tells us that it is very unlikely (probability 0.0038) that the results obtained were due to chance alone, given that the null hypothesis is true. That's why we reject the null hypothesis.

NOW TRY EXERCISE **9** ■

14.8 EXERCISES

CONCEPTS

1. A hypothesis is a statement or a claim about a ________ (population / sample).

2. To test a claim about a population proportion, we first state two hypotheses. The ________ hypothesis, denoted by H_0, is the assumed value of the proportion. The ________ hypothesis, denoted by H_1, is also called the research hypothesis.

3. A very large glass jar is filled with black and white beans. Your friend claims that the proportion of black beans is 0.3. By looking at the jar, you believe that the proportion of black beans is greater than 0.3. To test the claim that the proportion p of black beans is greater than 0.3, we mix the beans thoroughly and use a large cup to scoop out a random sample.

(a) State the null and alternative hypotheses.

H_0: ________

H_1: ________

(b) Suppose that the random sample we obtained consists of 200 beans, of which 180 are black. On the basis of intuition alone, which of the following conclusions would we make: reject H_0 or fail to reject H_0.

4. The students in Exercise 3 find that the proportion of black beans in the sample is 0.36. They use a calculator to find the *P*-value associated with this sample. This *P*-value is the ___________ of getting a random sample with the proportion of black beans greater than or equal to ________, assuming that the ___________ hypothesis is true. Use the $\alpha = 0.05$ significance level to answer the following.

(a) If the students obtain a *P*-value 0.01, they should ___________ the null hypothesis.

(b) If the students obtain a *P*-value of 0.2, they should ____________________ the null hypothesis.

SKILLS

5–8 ■ Perform a hypothesis test for the proportion of "successes" in a population as follows. Use the 0.05 significance level.

(a) State the null and alternative hypotheses.
(b) Calculate the *P*-value.
(c) Make a conclusion.
(d) Explain how the *P*-value supports your conclusion.

5. Test whether the proportion of "successes" in a population is greater than 0.5. A random sample of 100 has 59 successes.

6. Test whether the proportion of "successes" in a population is greater than 0.7. A random sample of 1000 has 716 successes.

7. Test whether the proportion of "successes" in a population is less than 0.6. A random sample of 200 has 114 successes.

8. Test whether the proportion of "successes" in a population is less than 0.6. A random sample of 200 has 107 successes.

9–12 ■ An experiment is performed to test the effectiveness of a medication. The treatment group has n_1 subjects, and x_1 of them recover; the control group has n_2 subjects, and x_2 recover. Let p_1 and p_2 be the true proportions of those who recover with and without the medication, respectively. Perform a two-sample hypothesis test as follows.

(a) State the null and alternative hypotheses.
(b) Calculate the *P*-value.
(c) Make a conclusion (use the $\alpha = 0.01$ significance level).
(d) Explain how the *P*-value supports your conclusion.

9. $n_1 = 60, \quad n_2 = 70, \quad x_1 = 20, \quad x_2 = 29$

10. $n_1 = 100, \quad n_2 = 100, \quad x_1 = 58, \quad x_2 = 49$

11. $n_1 = 1000, \quad n_2 = 800, \quad x_1 = 312, \quad x_2 = 207$

12. $n_1 = 500, \quad n_2 = 350, \quad x_1 = 220, \quad x_2 = 142$

APPLICATIONS

13–24 ■ For each situation perform a hypothesis as follows.

(a) State the null and alternative hypotheses.
(b) Calculate the *P*-value.
(c) Make a conclusion.

13. Political Poll In a poll of 1000 registered voters 548 indicate that they would vote for the Republican candidate for governor. Does this poll provide sufficient evidence that the Republican candidate is likely to win the election (i.e., get more than 50% of the vote)? Use the 0.05 significance level.

14. Political Poll In a poll of 1100 registered voters 560 indicate that they would vote for the Democratic candidate for the Senate. Does this poll provide sufficient evidence that the Democratic candidate is likely to win election to the Senate (i.e., get more than 50% of the vote)? Use the 0.05 significance level.

15. Choosing the Gender of a Baby A herbalist claims that his herbal drinks increase the proportion of male births to greater than 0.5. In a sample of 100 couples using the herbalist's system 62 have a male birth. Is the herbalist's claim supported by this sample? Use the 0.01 significance level.

16. Choosing the Gender of a Baby A herbalist claims that his herbal pills increase the proportion of female births to greater than 0.5. In a sample of 100 couples using the herbalist's system 55 have a female birth. Is the herbalist's claim supported by this sample? Use the 0.01 significance level.

17. Tasting Phenylthiocarbamide (PTC) The ability to taste the chemical PTC is genetically determined. For those who can taste it, PTC has a bitter taste; for others it is tasteless. It is known that 70% of the U.S. population can taste PTC. In a sample of 200 Native Americans it is found that 162 can taste this chemical. Test the hypothesis that the proportion of Native Americans who can taste PTC is greater than 0.70. Use the 0.01 significance level.

18. Cell Phones and Brain Cancer In a landmark 2006 study in Denmark on the health effects of prolonged exposure to cell phone radiation, a cohort of 400,095 cell phone users were considered. The proportion of the population who acquire brain cancer is 0.0015; the number of subjects in the study who acquired brain cancer is 580. Test the hypothesis that there is a higher incidence of brain cancer among cell phone users. Use the 0.01 significance level.

19. New Teaching Method A new method of teaching school mathematics claims to increase the proportion of students passing the entry-level mathematics (ELM) test. In a group of 416 students being taught with the new method, 280 pass the ELM test. Test the hypothesis that the proportion passing the exam is greater than the usual 0.70. Use the 0.05 significance level.

20. Methylmercury in Fish Fish concentrate mercury in their bodies in the form of methylmercury, a biological form of mercury, which is highly toxic to humans. Approximately 25% of the fish in lakes and streams in the United States have mercury levels above the safety levels determined by the U.S. Environmental Protection Agency. After a clean-up program to reduce the presence of mercury in the water, a random sample of 1500 fish was tested, and 343 of these had high mercury levels. Test the hypothesis that the clean-up program reduced the proportion of contaminated fish to less than 0.25. Use the 0.05 significance level.

21. Mission to Mars Polls are periodically conducted to gauge the public's support for funding space missions. In a poll of 600 Americans 316 say that it is important for the United States to be first to put a person on Mars. Test the claim that a majority (more than 50%) of Americans feel that it is important for the United States to be first to put a person on Mars. Use the 0.05 significance level.

22. Car Accidents near Home It is commonly assumed that 50% of car accidents occur "near home" (within 5 miles of home). In a randomly selected sample of 8000 accident reports, it is found that 4220 accidents occurred near home. Test the hypothesis that the proportion of accidents occurring within 5 miles of home is greater than 0.50. Use the 0.05 significance level.

23. Tomato Seeds A seed company claims that more than 95% of their hybridized tomato seeds germinate. To test the company's claim, a random sample of 2200 of the company's tomato seeds is selected and planted; 2080 of them germinate. Test the company's claim. Use the 0.05 significance level.

24. Dutch Elm Disease Dutch elm disease has been a persistent tree infestation across the northern United States. In a randomly selected sample of 2000 elm trees in a certain forest it was found that 650 have the disease. Test the hypothesis that more than 30% of the elm trees in the forest have the disease. Use the 0.05 significance level.

25–28 ■ Perform a two sample hypothesis test for each experiment as follows.

(a) Identify the two proportions, p_1 and p_2, that are to be compared.
(b) State the null and alternative hypotheses.
(c) Calculate the *P*-value
(d) Make a conclusion.

25. Vaccine Effectiveness To test the effectiveness of a flu vaccine, researchers set up an experiment. The vaccine is administered to the 150 subjects in the treatment group, and a placebo is given to the 70 subjects in the control group. In the treatment group 12 subjects get the flu, and in the control group 10 get the flu. Do the results indicate that the vaccine is effective in preventing the flu? Use the 0.01 significance level.

26. Helsinki Heart Study In a 1982–1987 study on the relationship between cholesterol levels and heart attacks, 4081 middle-aged men from Helsinki, Finland were the subjects. The 2051 men in the treatment group were given gemfibrozil, a cholesterol-lowering drug. The 2030 men in the control group were given a placebo. In the five-year period of the study, the incidence of heart attacks comprised 56 men in the treatment group and 84 men in the control group. Do the results indicate that gemfibrozil is effective in preventing heart attacks? Use the 0.01 significance level.

27. Gender and Nightmares In a 2009 study at the University of West England, 100 female students and 93 male students were interviewed about their dreams. When asked to describe their most recent dream, 34 of the women and 18 of the men reported having a nightmare. Is there a statistically significant difference in the proportion of men and women having nightmares? Use the 0.05 significance level.

28. Teaching Methods A study seeks to compare two mathematics teaching methods designed to prepare students for the college entry level mathematics exam (ELM). In the study, 380 students are taught by using method A, and 410 students are taught by using Method B. From the Method A group, 270 pass the ELM; from the Method B group, 290 pass. Is there a statistically significant difference in the proportion of students who pass the ELM in the two groups? Use the 0.05 significance level.

DISCOVERY ■ DISCUSSION ■ WRITING

29. The Significance Level Suppose you perform a hypothesis test at the 0.01 significance level and you *reject* the null hypothesis. What would your conclusion be if you used the 0.05 significance level?

30. The Significance Level Suppose you perform a hypothesis test at the 0.05 significance level and you *fail to reject* the null hypothesis. If you use the 0.01 significance level, would you still fail to reject the null hypothesis?

CHAPTER 14 | REVIEW

CONCEPT CHECK

1. What does the Fundamental Counting Principle say?

2. **(a)** What is a permutation of a set of distinct objects?
(b) How many permutations are there of *n* objects?
(c) How many permutations are there of *n* objects taken *r* at a time?

3. **(a)** What is a combination of *r* elements of aset?
(b) How many combinations are there of *n* elements taken *r* at a time?
(c) How many subsets does a set with *n* elements have?

4. In solving a problem involving picking *r* objects from *n* objects, how do you know whether to use permutations or combinations?

5. **(a)** What is meant by the sample space of an experiment?
(b) What is an event?
(c) Define the probability of an event *E* in a sample space *S*.
(d) What is the probability of the complement of *E*?

6. **(a)** What are mutually exclusive events?
(b) If *E* and *F* are mutually exclusive events, what is the probability of *E* or *F* occurring? What if *E* and *F* are not mutually exclusive?

7. **(a)** What is meant by the conditional probability of *E* given *F*? How is this probability calculated?
(b) What are independent events?
(c) If *E* and *F* are independent events, what is the probability of *E* and *F* occurring? What if *E* and *F* are not independent?

8. An experiment has two outcomes, *S* and *F*, where the probability of *S* is *p*. The experiment is performed *n* times.
(a) What type of probability is associated with this experiment?
(b) What is the probability that *S* occurs exactly *r* times?

9. Suppose that a game gives payouts $a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, \ldots, p_n$.
(a) What is the expected value of this game?
(b) What is the significance of the expected value?
(c) What is meant by a fair game? How do we determine whether the game is fair?

10. Suppose we have a set of one-variable data.
(a) What are "numerical data" and "categorical data"? Give examples of each type.
(b) What is meant by a 'measure of central tendency"? Describe three different measures of central tendency.
(c) What is meant by a "measure of spread"? Describe one such measure.
(d) What is a frequency table of data?
(e) What is a stemplot of data?
(f) What is a "five-number summary" of data?

11. Suppose we have a set of one-variable data.
(a) What is a histogram of the data? What information about the data does a histogram give us?
(b) What is a normal distribution?
(c) State the Empirical Rule for normally distributed data.

12. **(a)** Define "sample" and "population."
(b) To get information about a population from a sample, why is it important that the sample be randomly selected?
(c) Describe some of the methods of obtaining a random sample.
(d) Describe some types of sampling bias.
(e) What is a statistical experiment? Describe some possible experimental designs.

13. When using a sample to estimate a population proportion, what is "the margin of error"? How are sample size and margin of error related?

14. Suppose we have a set of two-variable data.
(a) Give some examples of two-variable data and possible relationships between the variables.
(b) How are the regression line, the correlation coefficient, the sample size, and the statistical significance of the correlation coefficient related?

15. **(a)** What is a hypothesis?
(b) In testing a hypothesis, what is the null hypothesis and what is the alternative hypothesis?
(c) What are the possible conclusions of a hypothesis test?

EXERCISES

1. A coin is tossed, a die is rolled, and a card is drawn from a deck. How many possible outcomes does this experiment have?

2. How many three-digit numbers can be formed by using the digits 1, 2, 3, 4, 5, and 6 if repetition of digits
(a) is allowed?
(b) is not allowed?

3. **(a)** How many different two-element subsets does the set {A, E, I, O, U} have?
(b) How many different two-letter "words" can be made by using the letters from the set in part (a)?

4. An airline company has overbooked a particular flight, and seven passengers must be "bumped" from the flight. If 120 passengers are booked on this flight, in how many ways can the airline choose the seven passengers to be bumped?

5. A quiz has ten true-false questions. In how many different ways can a student earn a score of exactly 70% on this quiz?

6. A test has ten true-false questions and five multiple-choice questions with four choices for each. In how many ways can this test be completed?

7. If you must answer only eight of ten questions on a test, how many ways do you have of choosing the questions you will omit?

8. An ice-cream store offers 15 flavors of ice cream. The specialty is a banana split with four scoops of ice cream. If each scoop must be a different flavor, how many different banana splits may be ordered?

9. A company uses a different three-letter security code for each of its employees. What is the maximum number of codes this security system can generate?

10. A group of students determines that they can stand in a row for their class picture in 120 different ways. How many students are in this class?

11. A coin is tossed ten times. In how many different ways can the result be three heads and seven tails?

12. The Yukon Territory in Canada uses a license-plate system for automobiles that consists of two letters followed by three numbers. Explain how we can know that fewer than 700,000 autos are licensed in the Yukon.

13. A group of friends have reserved a tennis court. They find that there are ten different ways in which two of them can play a singles game on this court. How many friends are in this group?

14. A pizza parlor advertises that they prepare 2048 different types of pizza. How many toppings does this parlor offer?

15. In Morse code, each letter is represented by a sequence of dots and dashes, with repetition allowed. How many letters can be represented by using Morse code if three or fewer symbols are used?

16. The genetic code is based on the four nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T). These are connected in long strings to form DNA molecules. For example, a sequence in the DNA may look like CAGTGGTACC . . . . The code uses "words," all the same length, that are composed of the nucleotides A, C, G, and T. It is known that at least 20 different words exist. What is the minimum word length necessary to generate 20 words?

17. Given 16 subjects from which to choose, in how many ways can a student select fields of study as follows?
(a) A major and a minor
(b) A major, a first minor, and a second minor
(c) A major and two minors

18. **(a)** How many three-digit numbers can be formed by using the digits 0, 1, . . . , 9? (Remember, a three-digit number cannot have 0 as the leftmost digit.)
(b) If a number is chosen randomly from the set {0, 1, 2, . . . , 1000}, what is the probability that the number chosen is a three-digit number?

19–20 ■ An **anagram** of a word is a permutation of the letters of that word. For example, anagrams of the word *triangle* include *griantle, integral,* and *tenalgir.*

19. How many anagrams of the word *TRIANGLE* are possible?

20. How many anagrams are possible from the word *RANDOM* ?

21. A committee of seven is to be chosen from a group of ten men and eight women. In how many ways can the committee be chosen using each of the following selection requirements?
(a) No restriction is placed on the number of men and women on the committee.
(b) The committee must have exactly four men and three women.
(c) Susie refuses to serve on the committee.
(d) At least five women must serve on the committee.
(e) At most two men can serve on the committee.
(f) The committee is to have a chairman, a vice chairman, a secretary, and four other members.

22. The U.S. Senate has two senators from each of the 50 states. In how many ways can a committee of five senators be chosen if no state is to have two members on the committee?

23. A jar contains ten red balls labeled 0, 1, 2, . . . , 9 and five white balls labeled 0, 1, 2, 3, 4. If a ball is drawn from the jar, find the probability of the given event.
(a) The ball is red.
(b) The ball is even-numbered.
(c) The ball is white and odd-numbered.
(d) The ball is red or odd-numbered.

24. If two balls are drawn from the jar in Exercise 23, find the probability of the given event.
(a) Both balls are red.
(b) One ball is white, and the other is red.
(c) At least one ball is red.
(d) Both balls are red and even-numbered.
(e) Both balls are white and odd-numbered.

25. A coin is tossed three times in a row, and the outcomes of each toss are observed.
(a) Find the sample space for this experiment.
(b) Find the probability of getting three heads.
(c) Find the probability of getting two or more heads.
(d) Find the probability of getting tails on the first toss.

26. A shelf has ten books: two mysteries, four romance novels, and four mathematics textbooks. If you select a book at random to take to the beach, what is the probability that it turns out to be a mathematics text?

27. A die is rolled, and a card is selected from a standard 52-card deck. What is the probability that both the die and the card show a six?

28. Find the probability that the indicated card is drawn at random from a 52-card deck.
(a) An ace
(b) An ace or a jack
(c) An ace or a spade
(d) A red ace

29. A card is drawn from a 52-card deck, a die is rolled, and a coin is tossed. Find the probability of each outcome.
(a) The ace of spades, a six, and heads
(b) A spade, a six, and heads
(c) A face card, a number greater than 3, and heads

30. Two dice are rolled. Find the probability of each outcome.
(a) The dice show the same number.
(b) The dice show different numbers.

31. Four cards are dealt from a standard 52-card deck. Find the probability that the cards are
(a) all kings
(b) all spades
(c) all the same color

32. In the "numbers game" lottery a player picks a three-digit number (from 000 to 999), and if the number is selected in the drawing, the player wins $500. If another number with the same digits (in any order) is drawn, the player wins $50. John plays the number 159.
(a) What is the probability that he will win $500?
(b) What is the probability that he will win $50?

33. In a TV game show, a contestant is given five cards with a different digit on each and is asked to arrange them to match the price of a brand-new car. If she gets the price right, she wins the car. What is the probability that she wins, assuming that she knows the first digit but must guess the remaining four?

34. A pizza parlor offers 12 different toppings, one of which is anchovies. If a a pizza is ordered at random (that is, any number of the toppings from 0 to all 12 may be ordered), what is the probability that anchovies is one of the toppings selected?

35. A drawer contains an unorganized collection of 50 socks; 20 are red and 30 are blue. Suppose the lights go out, so Kathy can't distinguish the color of the socks.
(a) What is the minimum number of socks Kathy must take out of the drawer to be sure of getting a matching pair?
(b) If two socks are taken at random from the drawer, what is the probability that they make a matching pair?

36. A volleyball team has nine players. In how many ways can a starting lineup be chosen if it consists of two forward players and three defense players?

37. Zip codes consist of five digits.
(a) How many different zip codes are possible?
(b) How many different zip codes can be read when the envelope is turned upside down? (An upside-down 9 is a 6; and 0, 1, and 8 are the same when read upside down.)
(c) What is the probability that a randomly chosen zip code can be read upside down?
(d) How many zip codes read the same upside down as right side up?

38. In the Zip+4 postal code system, zip codes consist of nine digits.
(a) How many different Zip+4 codes are possible?
(b) How many different Zip+4 codes are palindromes? (A palindrome is a number that reads the same from left to right as right to left.)
(c) What is the probability that a randomly chosen Zip+4 code is a palindrome?

39. Let $N = 3{,}600{,}000$. (Note that $N = 2^7 3^2 5^5$.)
(a) How many divisors does N have?
(b) How many even divisors does N have?
(c) How many divisors of N are multiples of 6?
(d) What is the probability that a randomly chosen divisor of N is even?

40. A fair die is rolled eight times. Find the probability of each event.
(a) A six occurs four times.
(b) An even number occurs two or more times.

41. Pacific Chinook salmon occur in two varieties: white-fleshed and red-fleshed. It is impossible to tell without cutting the fish open whether it is the white or red variety. About 30% of Chinooks have white flesh. An angler catches five Chinooks. Find the probability of each event.
(a) All are white.
(b) All are red.
(c) Exactly two are white.
(d) Three or more are red.

42. Two dice are rolled. John gets $5 if they show the same number; he pays $1 if they show different numbers. What is the expected value of this game?

43. Three dice are rolled. John gets $5 if they all show the same number; he pays $1 otherwise. What is the expected value of this game?

44. Mary will win $1,000,000 if she can name the 13 original states in the order in which they ratified the U.S. Constitution. Mary has no knowledge of this order, so she makes a guess. What is her expectation?

45–46 ■ A list of one-variable data is given. **(a)** Find the mean and median of the data. **(b)** Find the standard deviation.

45. 13 4 10 5 4 0

46. 10.3 6.5 7.2 4.6 9.1 8.4 5.0

47–48 ■ A list of one-variable data is given. **(a)** Find the five-number summary for the data set. **(b)** Draw a box plot for the data.

47.

9	7	6	8	7	6	2
5	6	4	9	8	6	7
6	7	9	9	10	8	7

48.

37	81	79	65	78	62
73	90	92	68	76	73
69	77	78	85	84	79

49–50 ■ A frequency table for a set of one-variable data is given. **(a)** Find the mean, median, and mode of the data. **(b)** Draw a histogram of the data, with bins of length 2 starting at 18.

49. Frequency Table

x	Frequency
20	1
21	3
22	6
23	8
24	7
25	3

50. Frequency Table

x	Frequency
20	1
21	0
22	4
23	5
24	6
25	7
26	8
27	10

51–52 ■ A stem-and-leaf plot of data is given. **(a)** Find the mean, median, and mode of the data. **(b)** Find the five-number summary for the data. **(c)** Draw a box plot for the data.

51.

Stem	Leaves
0	1 5
1	1 3 7
2	3 4 4 4 5 9
3	0 5 8 8
4	
5	4

1 | 3 means 13

52.

Stem	Leaves
2.0	0 1
2.1	
2.2	2 3 6
2.3	1 2 2 3 5 7
2.4	0 2

2.3 | 2 means 2.32

53–54 ■ A data set is approximately normally distributed with mean 41 and standard deviation 3.5. Use the Empirical Rule to find the proportion of data points that lie in the given interval.

53. Between 34 and 48

54. At most 44.5

55–56 ■ A data set is approximately normally distributed with mean 105 and standard deviation 22. Use a graphing calculator to find the proportion of data points that lie in the given interval.

55. Between 99 and 121

56. At least 79

57. Test Scores The stem-and-leaf plot shows the scores on a U.S. history exam.
(a) Find the mean and median score on the exam.
(b) Find the standard deviation for the scores.
(c) Find the five-number summary and draw a box plot for the scores.
(d) Draw a histogram of the data, with bins of length 5 starting at 40.
(e) A student in the class takes the exam late and receives a score of 83. Find the new mean and median score.

Stem	Leaves
4	5
5	
6	5 9
7	1 2 3 5 8
8	2 4 5 7 8 9
9	0 3 4 9

8 | 2 means 82

58. Gasoline Prices The lists on the following page show the price per gallon (in dollars) for gasoline at gas stations in East Springfield and West Springfield.
(a) Make a stem-and-leaf plot of the data. (3.1|2 means 3.12.)
(b) Find the mean and standard deviation for each gas station.
(c) Find the five-number summary and draw a box plot for each gas station. Compare the center and the spread in prices for the two stations.

(d) Draw histograms of the data, with bins of length 0.05, starting at 3.00.

East Springfield				West Springfield			
3.19	3.18	3.12	3.10	3.25	3.29	3.37	3.21
3.08	3.02	3.05	3.10	3.19	3.11	3.14	3.22
3.15	3.18	3.20	3.22	3.29	3.35	3.39	3.43
3.25	3.33			3.45	3.49		

59. Birch Heights The heights of birch trees in a Finnish national forest are approximately normally distributed with mean 11.3 meters and standard deviation 4.4 meters.

(a) Use the Empirical Rule to find the proportion of birch trees that have height between 6.9 and 15.7 meters and the proportion of birch trees that have height more than 20.1 meters.

(b) Use a graphing calculator to find the proportion of birch trees that have height less than 17 meters.

60. Body-Mass Index The body-mass index (BMI) is a measure for obesity. A person who has a BMI that is above 25 is considered overweight. The BMI for U.S. citizens is normally distributed with a mean of 22.2 and standard deviation of 3.9.

(a) Use the Empirical Rule to find the proportion of U.S. citizens who have a BMI less than18.3 and the proportion of U.S. citizens who have a BMI more than 26.1.

(b) Use a graphing calculator to find the proportion of U.S. citizens who have a BMI more than 25.

61–64 ■ A sample is selected as described. Is the sample a random sample? If it isn't, identify the type of sampling bias.

61. Satisfaction Survey To gauge customer satisfaction with their services, a car tire company places questionnaires in the waiting room. The completed questionnaires are collected and analyzed.

62. Department Store A group that opposes the building of a new department store in their neighborhood conducts a survey. The survey question states: "A large department store in our neighborhood will destroy local businesses, increase traffic congestion, and pollute the environment. Do you support the building of this store?"

63. Light Pollution Astronomy enthusiasts use the term *light pollution* to describe city lights that make it difficult to see stars at night. An astronomy club mails a survey to a sample of home owners of a community asking their opinion on removing all street lights. Very few surveys are returned, and most of these support the elimination of all street lights.

64. Texting and Driving To gauge drivers' attitudes toward texting while driving, a survey asking about the advantages of texting and driving is distributed to residents of a local retirement community.

65–66 ■ A study is described. State some possible confounding or lurking variable(s).

65. Caviar and the Common Cold To test the effect of eating caviar in preventing the common cold, a researcher selects a random sample of 50 individuals who regularly eat caviar and a random sample of 100 individuals who do not eat caviar. The researcher records the number of colds acquired in each group in a two-year period.

66. Religion and City Size A radio talk show host claims that people living in large cities are more religious than those who live in smaller towns. To test the claim, a random sample of 100 large cities (population over one million) and a random sample of 100 small towns (population under 100,000) are selected. The number of churches in each group is recorded.

67. Experimental Design To test the effectiveness of an investigational treatment for ADHD, a researcher selects a sample of 700 individuals (400 males and 300 females) diagnosed with the condition. Describe how treatment and control groups should be selected from the sample for the given experimental design. Sex (male or female) and BMI ("Normal" BMI ≤ 25 and "overweight" BMI > 25) may be factors affecting the results of the treatment.

(a) Completely randomized design.

(b) Block design (controlling for sex).

(c) Matched pair design (controlling for sex and BMI).

68. Pooraholics Anonymous Individuals with "pooraholic syndrome" tend to take jobs that are far below their potential or educational level. A psychologist forms a support group for such individuals. To test the effectiveness of her Pooraholics Anonymous support therapy, a sample of 200 individuals with the syndrome is selected (100 men and 100 women). It is thought that sex (male or female) and general happiness (happy or unhappy) are factors affecting the success of the therapy. Describe how treatment and control groups should be selected from the sample for the given experimental design.

(a) Completely randomized design.

(b) Block design (controlling for sex).

(c) Matched pair design (controlling for sex and happiness).

69. Popping Popcorn A popcorn company wants to estimate the proportion of the kernels of their brand ZIP popcorn that pop when heated. The company tests 500 randomly selected kernels and finds that 480 pop. Estimate the proportion of the kernels of brand ZIP popcorn that pop, and find the margin or error.

70. Popping Popcorn The popcorn company in Exercise 69 wants to estimate the proportion of brand POP popcorn that pops with a margin of error of 1%. What sample size should the company use?

71–72 ■ A claim is made. Use the appropriate hypothesis testing methods as follows: **(a)** State the null and alternative hypotheses. **(b)** Calculate the P-value. **(c)** Make a conclusion.

71. Dentists Recommend Toothpaste A toothpaste company advertises that nine out of ten dentists recommend their Brand X toothpaste. A consumer group claims that fewer than 90% of dentists recommend Brand X. To test their claim, a random sample of 500 dentists is selected. Of these, 405 recommend Brand X. Is the consumer group's claim supported by these data? Use the $\alpha = 0.05$ significance level.

72. Brand Loyalty for Toothpaste A customer is considered "loyal'" to a particular brand of toothpaste if the customer purchases the brand, then purchases the same brand again within three months. The makers of Toothpaste Y claim that its customers are more loyal than those who use Toothpaste X. To test this claim, two random samples of shoppers are selected: 60 who purchased Toothpaste X and 80 who purchased Toothpaste Y. Of the Toothpaste X group, 50 purchase the same brand again within three months, and of the Toothpaste Y group, 72 purchase the same brand again within three months. Is the loyalty claim of Toothpaste Y supported by these data? Use the $\alpha = 0.05$ significance level.

CHAPTER 14 TEST

1. Alice and Bill have four grandchildren, and they have three framed pictures of each grandchild. They wish to choose one picture of each grandchild to display on the piano in their living room, arranged from oldest to youngest. In how many ways can they do this?

2. A hospital cafeteria offers a fixed-price lunch consisting of a main course, a dessert, and a drink. If there are four main courses, three desserts, and six drinks to pick from, in how many ways can a customer select a meal consisting of one choice from each category?

3. An Internet service provider requires its customers to select a password consisting of four letters followed by three digits. Find how many such passwords are possible in each of the following cases:
   (a) Repetition of letters and digits is allowed.
   (b) Repetition of letters and digits is not allowed.

4. Over the past year, John has purchased 30 books.
   (a) In how many ways can he pick four of these books and arrange them, in order, on his nightstand bookshelf?
   (b) In how many ways can he choose four of these books to take with him on his vacation at the shore?

5. A commuter must travel from Ajax to Barrie and back every day. Four roads join the two cities. The commuter likes to vary the trip as much as possible, so she always leaves and returns by different roads. In how many different ways can she make the round-trip?

6. A pizza parlor offers four sizes of pizza and 14 different toppings. A customer may choose any number of toppings (or no topping at all). How many different pizzas does this parlor offer?

7. An *anagram* of a word is a rearrangement of the letters of the word.
   (a) How many anagrams of the word *LOVE* are possible?
   (b) How many different anagrams of the word *KISSES* are possible?

8. A board of directors consisting of eight members is to be chosen from a pool of 30 candidates. The board is to have a chairman, a treasurer, a secretary, and five other members. In how many ways can the board of directors be chosen?

9. One card is drawn from a deck. Find the probability of each event.
   (a) The card is red.
   (b) The card is a king.
   (c) The card is a red king.

10. A jar contains five red balls, numbered 1 to 5, and eight white balls, numbered 1 to 8. A ball is chosen at random from the jar. Find the probability of each event.
    (a) The ball is red.
    (b) The ball is even-numbered.
    (c) The ball is red or even-numbered.

11. Three people are chosen at random from a group of five men and ten women. What is the probability that all three are men?

12. Two dice are rolled. What is the probability of getting doubles?

13. In a group of four students, what is the probability that at least two have the same astrological sign?

14. An unbalanced coin is weighted so that the probability of heads is 0.55. The coin is tossed ten times.
    (a) What is the probability of getting exactly 6 heads?
    (b) What is the probability of getting less than 3 heads?

15. You are to draw one card from a deck. If it is an ace, you win \$10; if it is a face card, you win \$1; otherwise, you lose \$0.50. What is the expected value of this game?

16. A data set is given.

(a) Make a frequency table for the data.

(b) Find the mean, median, mode, and standard deviation of the data.

(c) Find the five-number summary of the data, and draw a box plot.

(d) Draw a histogram of the data, with bins of length 2 starting at 30.

38	36	37	35	39	38	31
34	34	37	39	37	38	38

17. The birth weights of infants born in India are normally distributed with a mean of 2900 grams and standard deviation of 450 grams.

(a) Use the Empirical Rule to find the probability that a randomly selected infant in India has a birth weight of at most 3350 grams.

(b) Use a graphing calculator to find the probability that a randomly selected infant in India has a birth weight of at least 2500 grams.

18. In a poll of 500 randomly selected residents of a small town, 340 support the building of a new community center. Estimate the proportion of residents who support the building project, and determine the margin of error.

19. The National Institute of Mental Health reports that 10% of people suffer from some form of depression. A coastal New Hampshire town claims that its residents are very happy, and consequently, fewer than 10% suffer from depression. In a random sample of 300 people from the town 21 suffer from depression. Test the town's claim as follows.

(a) State the null and alternative hypotheses.

(b) Calculate the P-value.

(c) Make a conclusion. (Use the $\alpha = 0.05$ significance level.)

20. In 2009 Bayer conducted a randomized double-blind clinical trial with 170 subjects of the drug moxifloxacin for the treatment of tuberculosis. Of the trial subjects, 85 were assigned to the treatment group and 85 to the control group. The drug moxifloxacin was administered to the treatment group, and ethambutol (another drug) was administered to the control group.

(a) From the description of this trial as "randomized," how were subjects assigned to the treatment and control groups?

(b) At the end of the trial 68 subjects in the treatment group and 54 in the control group tested negative for the presence of the tuberculosis bacterium. Use these data to test the hypothesis that moxifloxacin is more effective than ethambutol for the treatment of tuberculosis, and state the P-value. (Use the $\alpha = 0.01$ significance level.)

FOCUS ON MODELING

The Monte Carlo Method

A good way to familiarize ourselves with a fact is to experiment with it. For instance, to convince ourselves that the earth is a sphere (which was considered a major paradox at one time), we could go up in a space shuttle to see that it is so; to see whether a given equation is an identity, we might try some special cases to make sure there are no obvious counterexamples. In problems involving probability, we can perform an experiment many times and use the results to estimate the probability in question. In fact, we often model the experiment on a computer, thereby making it feasible to perform the experiment a large number of times. This technique is called the **Monte Carlo method**, named after the famous gambling casino in Monaco.

EXAMPLE 1 | The Contestant's Dilemma

In a TV game show, a contestant chooses one of three doors. Behind one of them is a valuable prize; the other two doors have nothing behind them. After the contestant has made her choice, the host opens one of the other two doors—one that he knows does not conceal a prize—and then gives her the opportunity to change her choice.

Should the contestant switch or stay, or does it matter? In other words, by switching doors, does she increase, decrease, or leave unchanged her probability of winning? At first, it may seem that switching doors doesn't make any difference. After all, two doors are left—one with the prize and one without—so it seems reasonable that the contestant has an equal chance of winning or losing. But if you play this game many times, you will find that by switching doors, you actually win about $\frac{2}{3}$ of the time.

Contestant: "I choose door number 2."

Contestant: "Oh no, what should I do?"

The authors modeled this game on a computer and found that in one million games the simulated contestant (who always switches) won 667,049 times—very close to $\frac{2}{3}$ of the time. Thus it seems that switching doors does make a difference: Switching increases the contestant's chances of winning. This experiment forces us to reexamine our reasoning. Here is why switching doors is the correct strategy:

1. When the contestant first made her choice, she had a $\frac{1}{3}$ chance of winning. If she doesn't switch, no matter what the host does, her probability of winning remains $\frac{1}{3}$.
2. If the contestant decides to switch, she will switch to the winning door if she had initially chosen a losing one or to a losing door if she had initially chosen the winning one. Since the probability of having initially selected a losing door is $\frac{2}{3}$, by switching the probability of winning then becomes $\frac{2}{3}$.

We conclude that the contestant should switch, because her probability of winning is $\frac{2}{3}$ if she switches and $\frac{1}{3}$ if she doesn't. Put simply, there is a much greater chance that she initially chose a losing door (since there are more of these), so she should switch. ■

An experiment can be modeled by using any computer language or programmable calculator that has a random-number generator. This is a command or function (usually called `Rnd` or `Rand`) that returns a randomly chosen number x with $0 \le x < 1$. In the next example we see how to use this to model a simple experiment.

EXAMPLE 2 | Monte Carlo Model of a Coin Toss

When a balanced coin is tossed, each outcome—"heads" or "tails"—has probability $\frac{1}{2}$. This doesn't mean that if we toss a coin several times, we will necessarily get exactly half heads and half tails. We would expect, however, the proportion of heads and of tails to get closer and closer to $\frac{1}{2}$ as the number of tosses increases. To test this hypothesis, we could toss a coin a very large number of times and keep track of the results. But this is a very tedious process, so we will use the Monte Carlo method to model this process.

To model a coin toss with a calculator or computer, we use the random-number generator to get a random number x such that $0 \le x < 1$. Because the number is chosen randomly, the probability that it lies in the first half of this interval $\left(0 \le x < \frac{1}{2}\right)$ is the same

```
PROGRAM:HEADTAIL
:0→J:0→K
:For(N,1,100)
:rand→X
:int(2X)→Y
:J+(1-Y)→J
:K+Y→K
:END
:Disp"HEADS=",J
:Disp"TAILS=",K
```

as the probability that it lies in the second half $\left(\frac{1}{2} \le x < 1\right)$. Thus we could model the outcome "heads" by the event that $0 \le x < \frac{1}{2}$ and the outcome "tails" by the event that $\frac{1}{2} \le x < 1$.

An easier way to keep track of heads and tails is to note that if $0 \le x < 1$, then $0 \le 2x < 2$, and so $[\![2x]\!]$, the integer part of $2x$, is either 0 or 1, each with probability $\frac{1}{2}$. (On most programmable calculators, the function `Int` gives the integer part of a number.) Thus we could model "heads" with the outcome "0" and "tails" with the outcome "1" when we take the integer part of $2x$. The program in the margin models 100 tosses of a coin on the TI-83 calculator. The graph in Figure 1 shows what proportion p of the tosses have come up "heads" after n tosses. As you can see, this proportion settles down near 0.5 as the number n of tosses increases—just as we hypothesized.

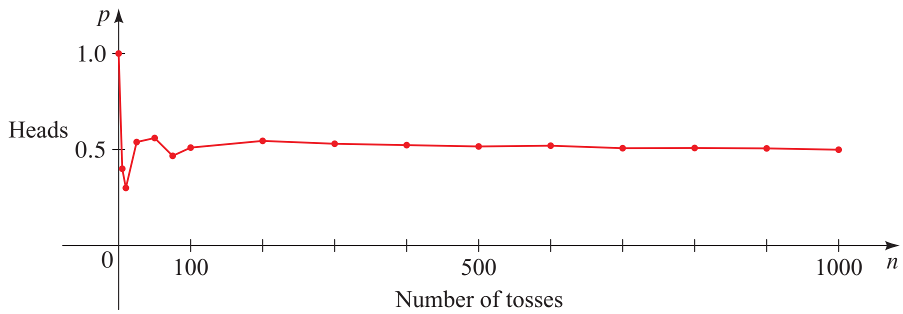


FIGURE 1 Relative frequency of "heads"

In general, if a process has n equally likely outcomes, then we can model the process using a random-number generator as follows: If our program or calculator produces the random number x, with $0 \le x < 1$, then the integer part of nx will be a random choice from the n integers $0, 1, 2, \ldots, n - 1$. Thus we can use the outcomes $0, 1, 2, \ldots, n - 1$ as models for the outcomes of the actual experiment.

PROBLEMS

1. **Winning Strategy** In a game show like the one described in Example 1, a prize is concealed behind one of ten doors. After the contestant chooses a door, the host opens eight losing doors and then gives the contestant the opportunity to switch to the other unopened door.
   - **(a)** Play this game with a friend 30 or more times, using the strategy of switching doors each time. Count the number of times you win, and estimate the probability of winning with this strategy.
   - **(b)** Calculate the probability of winning with the "switching" strategy using reasoning similar to that in Example 1. Compare with your result from part (a).
2. **Family Planning** A couple intend to have two children. What is the probability that they will have one child of each sex? The French mathematician D'Alembert analyzed this problem (incorrectly) by reasoning that three outcomes are possible: two boys, or two girls, or one child of each sex. He concluded that the probability of having one of each sex is $\frac{1}{3}$, mistakenly assuming that the three outcomes are equally likely.
   - **(a)** Model this problem with a pair of coins (using "heads" for boys and "tails" for girls), or write a program to model the problem. Perform the experiment 40 or more times, counting the number of boy-girl combinations. Estimate the probability of having one child of each sex.
   - **(b)** Calculate the correct probability of having one child of each sex, and compare this with your result from part (a).
3. **Dividing a Jackpot** A game between two players consists of tossing a coin. Player A gets a point if the coin shows heads, and player B gets a point if it shows tails. The first player to get six points wins an $8000 jackpot. As it happens, the police raid the place when player A has five points and B has three points. After everyone has calmed down, how

should the jackpot be divided between the two players? In other words, what is the probability of A winning (and that of B winning) if the game were to continue?

The French mathematicians Pascal and Fermat corresponded about this problem, and both came to the same correct conclusion (though by very different reasonings). Their friend Roberval disagreed with both of them. He argued that player A has probability $\frac{3}{4}$ of winning, because the game can end in the four ways *H*, *TH*, *TTH*, *TTT*, and in three of these, A wins. Roberval's reasoning was wrong.

(a) Continue the game from the point at which it was interrupted, using either a coin or a modeling program. Perform this experiment 80 or more times, and estimate the probability that player A wins.

(b) Calculate the probability that player A wins. Compare with your estimate from part (a).

4. Long or Short World Series? In the World Series the top teams in the National League and the American League play a best-of-seven series; that is, they play until one team has won four games. (No tie is allowed, so this results in a maximum of seven games.) Suppose the teams are evenly matched, so the probability that either team wins a given game is $\frac{1}{2}$.

(a) Use a coin or a modeling program to model a World Series, in which "heads" represents a win by Team A and "tails" represents a win by Team B. Perform this experiment at least 80 times, keeping track of how many games are needed to decide each series. Estimate the probability that an evenly matched series will end in four games. Do the same for five, six, and seven games.

(b) What is the probability that the series will end in four games? Five games? Six games? Seven games? Compare with your estimates from part (a).

(c) Find the expected value for the number of games until the series ends. [*Hint:* This will be $P(\text{four games}) \times 4 + P(\text{five}) \times 5 + P(\text{six}) \times 6 + P(\text{seven}) \times 7$.]

5. Estimating π In this problem we use the Monte Carlo method to estimate the value of π. The circle in the figure has radius 1, so its area is π, and the square has area 4. If we choose a point at random from the square, the probability that it lies inside the circle will be

$$\frac{\text{area of circle}}{\text{area of square}} = \frac{\pi}{4}$$

The Monte Carlo method involves choosing many points inside the square. Then we have

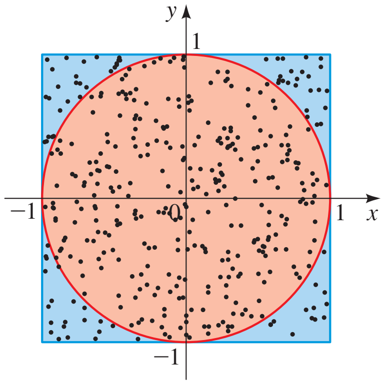
$$\frac{\text{number of hits inside circle}}{\text{number of hits inside square}} \approx \frac{\pi}{4}$$

Thus 4 times this ratio will give us an approximation for π.

To implement this method, we use a random-number generator to obtain the coordinates (x, y) of a random point in the square and then check to see whether it lies inside the circle (that is, we check if $x^2 + y^2 < 1$). Note that we need to use only points in the first quadrant, since the ratio of areas is the same in each quadrant. The program in the margin shows a way of doing this on the TI-83 calculator for 1000 randomly selected points.

```
PROGRAM:PI
:0→P
:For(N,1,1000)
:rand→X:rand→Y
:P+((X²+Y²)<1)→P
:End
:Disp "PI IS
APPROX",4*P/N
```

Carry out this Monte Carlo simulation for as many points as you can. How do your results compare with the actual value of π? Do you think this is a reasonable way to get a good approximation for π?

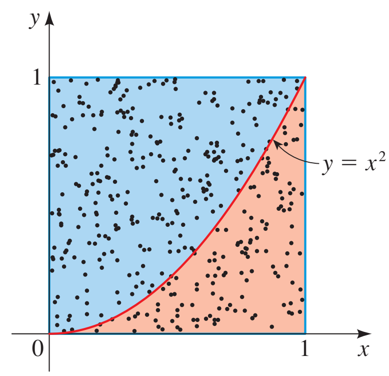


The "contestant's dilemma" problem discussed on page 967 is an example of how subtle probability can be. This problem was posed in a nationally syndicated column in *Parade* magazine in 1990. The correct solution was presented in the column, but it generated considerable controversy, with thousands of letters arguing that the solution was wrong. This shows how problems in probability can be quite tricky. Without a lot of experience in probabilistic thinking, it's easy to make a mistake. Even great mathematicians such as D'Alembert and Roberval (see Problems 2 and 3) made mistakes in probability. Professor David Burton writes in his book *The History of Mathematics,* "Probability theory abounds in paradoxes that wrench the common sense and trip the unwary."

6. Areas of Curved Regions The Monte Carlo method can be used to estimate the area under the graph of a function. The figure below shows the region under the graph of $f(x) = x^2$, above the x-axis, between $x = 0$ and $x = 1$. If we choose a point in the square at random, the probability that it lies under the graph of $f(x) = x^2$ is the area under the graph divided by the area of the square. So if we randomly select a large number of points in the square, we have

$$\frac{\text{number of hits under graph}}{\text{number of hits in square}} \approx \frac{\text{area under graph}}{\text{area of square}}$$

Modify the program from Problem 5 to carry out this Monte Carlo simulation and approximate the required area.



7. Random Numbers Choose two numbers at random from the interval $[0, 1)$. What is the probability that the sum of the two numbers is less than 1?

(a) Use a Monte Carlo model to estimate the probability.

(b) Calculate the exact value of the probability. [*Hint:* Call the numbers x and y. Choosing these numbers is the same as choosing an ordered pair (x, y) in the unit square $\{(x, y) \mid 0 \le x < 1, 0 \le y < 1\}$. What proportion of the points in this square corresponds to $x + y$ being less than 1?]