OBJECTIVES

In this chapter, you will learn:

1. how to collect data using unbiased random samples.
2. how to compute measures of central tendency.
3. how to compute measures of spread and
4. how to summarize a set of data numerically and visually.

9.1 Populations, Samples, and Data

INITIAL PROBLEM

A university mathematics department is going to conduct a study on "Improving Problem-Solving Skills." The researchers will ask for student volunteers from a precalculus class, select a group of five students, and teach them some problem-solving techniques. Twenty-five of Professor Spark's students have indicated they would like to participate in the study. How can the professor select 5 students from the 25 volunteers in a fair way, so that no one can claim the professor showed favoritism?

A solution of this Initial Problem is on page 573.

Whether we are in a classroom or simply going about our daily and professional lives, we need information. Sometimes the information we need is not readily available, and we have to seek it out before we can proceed.

Suppose that we want to find the average height of 50-year-old men in the United States. Of course, we cannot ask every 50-year-old man in the United States how tall he is. One way to simplify the task is to look at a smaller group of men that is representative of the entire group.

Tidbit

The term *statistics* was first applied to collections of data relating to matters important to a government, such as population, tax assessment, etc. An important early example is the Doomsday Book, the record of William the Conqueror's survey of England in the latter part of the 11th century.

POPULATIONS AND SAMPLES

One of the most common uses of statistics is gathering and analyzing information about specific groups of people or objects. For example, an insurance company may need to know the average height and weight of 50-year-old males, political advisors may need to know the percentage of people who support the President's foreign policies, or a manufacturer may need to know the percentage of defective parts produced during a manufacturing process. When analyzing information about a group, the entire set that we are studying is called the **population**. The population may consist of people, as it does for the insurance company interested in the heights and weights of 50-year-old men. However, the population can also consist of inanimate objects, as it does for the manufacturer interested in the percentage of defective parts. In the case of the manufacturer, the population is the set of all parts produced by the manufacturer. The population may even consist of events. For example, a population might be defined to be all the transactions occurring at a bank branch during a particular year or all the hurricanes during the 20th century. Whatever the population consists of, its members are called **elements**.

EXAMPLE 9.1 Suppose you wish to determine voter opinion regarding a ballot measure to fund the proposed new library. To do so, you survey potential voters among the pedestrians on Main Street during the lunch hour. What is the population in this survey?

SOLUTION The group you are interested in is the set of all people who are going to vote on a ballot measure in the upcoming election. Thus, the population consists of *all* those people who intend to vote on the ballot measure, no matter where they are and what they are doing at the time you conduct your survey. Figure 9.1 uses a diagram to represent the population schematically.

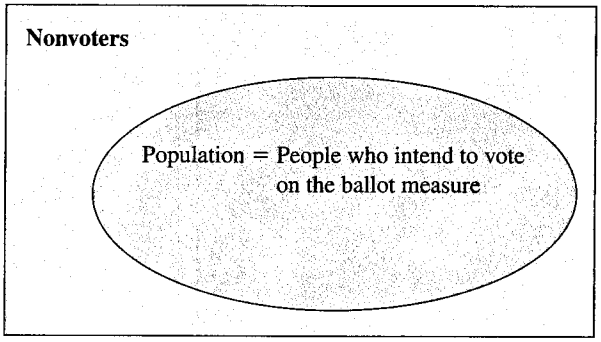


Figure 9.1 ■

Any characteristic of the individuals in our population is called a **variable**. In Example 9.1, the variable we are interested in is the potential voter's opinion "for" or "against" the library proposal, and when we interview such a potential voter, we say that we **measure** that variable. In Example 9.1, we were interested in only one variable, but in many cases, we will measure several variables.

A **census** involves measuring a variable for every individual in the population. For a small population, it may be possible to measure the variables for every member, but in general, it is too time-consuming or expensive to survey every member of a population. For example, an insurance company probably doesn't have the time or money needed to weigh and measure *every* 50-year-old male.

Instead of dealing with the entire population under study, we will usually select a portion, or subset, from the population and analyze that instead. A subset of the population is called a **sample**. Let's revisit the ballot measure in Example 9.1.

EXAMPLE 9.2 What is the sample in the survey discussed in Example 9.1? What variable is being measured?

SOLUTION A sample was defined to be a subset of the population. From Example 9.1, we know that the population consists of all those people who intend to vote on the ballot measure in the upcoming library election. Thus, the sample consists of those persons interviewed on the street who say they will be voting on the ballot measure. If a

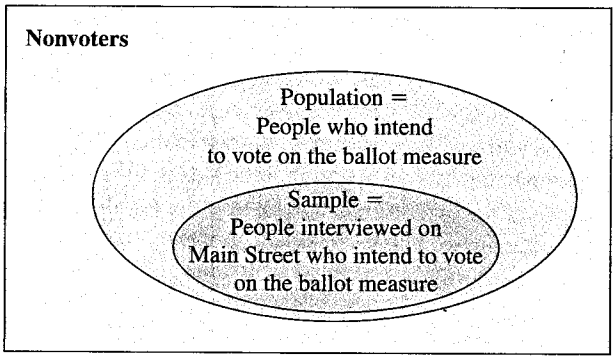


Figure 9.2

Tidbit

The U.S. Constitution requires a new count of the population every 10 years. The Census Bureau collects the population data that are used to determine voting districts, apportion seats in the House of Representatives, and allocate funding for federal programs. The census has traditionally been performed by an actual head count (enumeration). However, the method of conducting the census has become a divisive issue. The Clinton administration proposed using statistical sampling to improve the accuracy and lower the cost of the 2000 census. House Republicans filed two lawsuits, claiming that sampling was not only unconstitutional, but would actually be less accurate. In the end, the U.S. Supreme Court ruled against the use of sampling in conducting the census.

person you interview is not going to vote on the ballot measure, then that person is not in the sample, even if he or she has an opinion on the library issue. Figure 9.2 illustrates that the sample must be a subset of the population. The variable to be measured is the voter's intent to vote "yes" or "no" on the ballot measure. ■

DATA AND BIAS

The measurement information recorded from a sample is data. There are several types of data. If the measurements are naturally numerical, then we are measuring a **quantitative variable** and we obtain **quantitative data**. For example, the heights and weights measured for 50-year-old men would be quantitative data. Voter opinions "for" or "against" the library proposal are not numerical. Data that cannot be measured on a natural numerical scale are called **qualitative data**, and the variable is said to be a **qualitative variable**. Of course, numerical codes can be used to represent qualitative data. For example, we might record a 1 to represent a voter in favor of a proposal and a 0 to represent a voter opposed to a proposal.

We can further classify qualitative data. **Ordinal data** are qualitative data for which there is a natural ordering. An example of ordinal data would be the rankings of pizzas on a scale of "excellent," "good," "fair," and "poor." A numerical code for ordinal data should reflect the natural ordering. For the pizza rankings, we might use a code of 4 for excellent, 3 for good, 2 for fair, and 1 for poor. **Nominal data** are qualitative data for which there is no natural ordering. An example of nominal data would be eye color. For nominal data, the number values in a numerical code would serve as identification only. Figure 9.3 illustrates the types of data that can be obtained when measuring a variable.

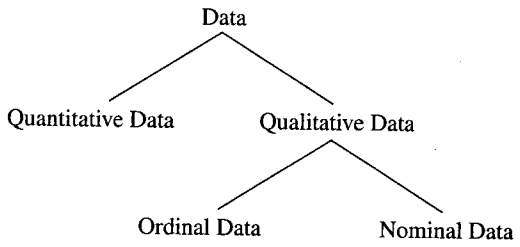


Figure 9.3

EXAMPLE 9.3 Suppose you wish to determine voter opinion regarding a ballot measure to fund the proposed new library, but you are also interested in profiling the voters who favor and those who oppose the library. You survey potential voters among the pedestrians on Main Street during the lunch hour to determine their political affiliation and age, as well as their opinion on the library measure. Classify the variables as quantitative or qualitative.

SOLUTION Political affiliation and opinion on the library measure are qualitative variables. Age is a quantitative variable. ■

If a sample has characteristics that are typical of the entire population, then it is said to be a **representative sample**. For example, in finding the average height of 50-year-old males, if a sample is representative, then the average height will be approximately the same as that of the population of 50-year-old males.

One of the most important uses of statistics is **statistical inference**, in which an estimate or prediction is made for the entire population based on data collected from a sample. For example, if the average height in a sample of 50-year-old men is 70 inches, we might claim that the average height of all 50-year-old men in the United States is about 70 inches. A sample that is not representative of the population can easily lead to an erroneous conclusion. This was exactly what happened with the *Literary Digest*'s prediction of the outcome of the 1936 election between Landon and Roosevelt, as discussed at the beginning of this chapter. A **bias** is a flaw in the sampling procedure that makes it more likely that the sample will not be representative of the population.

As an example, suppose a local late-night news program had a call-in telephone poll on a gun control issue and charged callers 50 cents to participate in the survey. Such a telephone poll has many sources of bias. One important source of bias is that it takes effort and some expense to participate. This means that people who feel strongly enough about gun control to make a call and are also willing to part with 50 cents are more likely to participate. Thus, the people who call the program probably do not fairly represent the opinions of all the local citizens on the issue of gun control. Other sources of bias in this survey include there being nothing to prevent people just visiting or passing through from participating and there being nothing to prevent people from voting more than once.

Another form of bias that can also affect the results of a survey is the wording of questions. For example, if callers on the gun control issue are asked to answer the question "Should government be allowed to limit a citizen's right to defend his or her home?" the survey can be expected to be biased against gun control.

EXAMPLE 9.4 Suppose you wish to determine voter opinion regarding the elimination of the capital gains tax (a profit made on an investment is called a "capital gain"). To do this, you survey potential voters on a street corner near Wall Street in New York City. Identify a source of bias in this poll.

SOLUTION One source of bias in choosing this sample is that many people involved in trading stocks work on Wall Street; thus, a disproportionate number of persons in the sample are likely to be employed in stock-related jobs. The incomes of these people could be enhanced by the elimination of the capital gains tax, so they are likely to favor eliminating that tax. ■

Even if a population consists of objects, the sampling procedure may be biased, as we see in the next example.

EXAMPLE 9.5 Suppose that an automobile manufacturer wants to test the reliability of one lot of 1000 alternators produced at a certain factory. Technicians will test the first 30 alternators from the lot for defects. Describe the population, the sample, the variable to be measured, and any potential sources of bias.

SOLUTION The population is the entire lot of 1000 alternators produced at the factory. The sample is the set of the *first* 30 alternators from the lot. The variable is the status of the alternator, which is either "defective" or "nondefective." Choosing to use the first 30 alternators can introduce bias in the sampling. It is possible that these 30 alternators are made with special care or that defects are less likely at the start of a run because the workers are fresh. On the other hand, it is possible that the number of defects in this group are more likely if "bugs" are worked out of the system at start-up. ■

A sample may be biased in many ways. The most easily biased samples are those obtained using surveys. Table 9.1 lists the most frequently occurring types of bias in surveys.

Table 9.1

COMMON SOURCES OF BIAS IN SURVEYS
Faulty Sampling: The sample selected is not representative.
Faulty Questions: Questions are worded to influence the answers.
Faulty Interviewing: Interviewers fail to survey the entire sample, misread questions, and/or misinterpret respondents' answers.
Lack of Understanding or Knowledge: The person being interviewed does not understand what is being asked or needs more information to answer the question.
False Answers: The person being interviewed intentionally gives incorrect information.

SIMPLE RANDOM SAMPLES

The *Literary Digest*'s incorrect prediction of the 1936 presidential election dramatically demonstrated that obtaining more data will not correct the errors introduced by biased sampling. On the other hand, if the sampling is unbiased, then, as we will see in Chapter 11, obtaining more data will make the sample more likely to be representative. One kind of sample that is unbiased is the simple random sample. Given a population and a desired sample size (that is, the number of persons or things desired in the sample), a **simple random sample** is any sample that is chosen in such a way that all samples of the same size are equally likely to be chosen.

Tidbit

A selection method equivalent to "drawing names from a hat" was used in the first Vietnam-era draft lottery of 1970. Numbers 1–366, corresponding to dates of birth from January 1 through December 31, were placed in capsules that were put in a drum and mixed. Draft numbers were then assigned according to the order in which numbers were drawn from the drum. The first number selected was 258, meaning young men born on the 258th day of the year (September 14) were the first ones called up. Although this method was believed to be fair to all candidates for the draft, statisticians who analyzed the results claimed that the sampling was biased. Because of poor mixing of the numbers in the drum, young men born earlier in the year were less likely to be called up than those born later in the year.

For example, suppose you have three tickets to a sold-out concert, and you are trying to decide which two of your four good friends to take with you without showing any favoritism. You need to choose two persons from Alanis, Brandy, Carol, and Deborah. You can choose a simple random sample of size two by writing their names on four slips of paper, mixing them, and selecting two slips without looking. Because any pair of names is equally likely to be chosen, this sample is a simple random sample.

Another way to choose a simple random sample is to use a random-number generator or a table of random numbers. A **random-number generator** is a computer or calculator program designed to produce numbers that are as random as possible; that is, the numbers have no apparent pattern. A **random-number table** is a table produced with a random-number generator. Part of a random-number table is shown in Figure 9.4. Note that the numbers in the far left column identify the rows in the table; they are not part of the body of the table.

In order to produce a simple random sample using a random-number table, we pick an arbitrary place on the table to begin and then move across or down the table in some systematic way. For example, suppose we wish to choose a simple random sample of size 5 from a group of 10 people. First, we give each of the 10 people in the population a one-digit label for identification: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Then we pick a place on the table to begin. For simplicity, let's begin at the top of the first column of the table in Figure 9.4, in the row labeled 101. The first random number in that row is 03918. Because we want one-digit numbers, we will use only the first digit in each row of this column. We move down the column and take the first five digits, ignoring any duplication. The first digits in this column are 0, 1, 4, 6, 4, 9, and so on. After crossing out the duplicate 4, we obtain our simple random sample, in this case, 0, 1, 4, 6, and 9. This selection is illustrated in Figure 9.5.

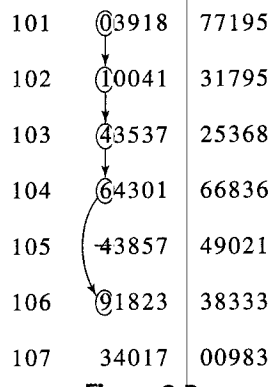


Figure 9.5

101	03918	77195	47772	21870	87122	99445
102	10041	31795	63857	64569	34893	20429
103	43537	25368	95237	17707	34280	04755
104	64301	66836	12201	60638	85624	33306
105	43857	49021	49026	93608	51382	49238
106	91823	38333	37006	78545	23827	39103
107	34017	00983	48659	39445	90910	29087
108	49105	95041	94232	50784	59181	44253
109	72479	24246	35932	33358	34853	77573
110	84281	57601	78425	36246	79348	41681
111	61589	93355	41310	17068	65700	54464
112	25318	28496	80120	31632	06746	90642
113	40113	91130	74270	27914	80511	70243
114	58420	96471	28464	72438	37667	16233
115	18075	32457	50011	42175	41029	07733
116	52754	43382	02151	46182	40557	94157
117	05255	73603	15957	99738	62835	62959
118	76032	69846	63316	48201	11580	45699
119	97050	48883	17828	98601	74821	06605
120	29030	55519	63362	55720	15296	78787
121	45609	12114	36541	53609	09322	28694
122	07608	55455	49299	90355	35334	29000
123	94901	06633	04618	82809	76952	21697
124	50581	84325	17532	57302	81752	25570
125	22265	14648	32967	10792	81713	68326
126	59294	06043	86457	78791	44380	62238
127	45473	93910	79160	19436	00813	75916
128	40239	02596	12487	99703	08901	49759
129	30241	44100	59953	83094	05261	46901
130	43837	77175	96514	61955	75287	24839
131	25050	80925	64073	70415	39896	69297
132	01445	23629	74556	24642	01672	92860
133	85236	77764	06026	33455	17737	08377
134	05946	75867	30147	53490	50415	24093
135	61189	32931	99257	50892	66516	45434
136	91267	07544	22194	04212	20015	15407
137	17039	95693	69650	40076	57722	38787
138	58541	34646	17657	30584	94546	09286
139	85563	13994	46354	93939	12491	41648
140	48576	89126	32012	39665	43906	76405
141	00543	87408	87066	74781	13065	35705
142	27954	32772	58815	88341	28322	05945
143	89156	74789	42290	03617	10054	13262
144	62334	04229	42057	10099	35791	10708
145	76172	20142	30526	88296	61844	89118

Figure 9.4

Tidbit

Random-number generators are built into CD players to randomly shuffle selections on a CD or a group of CDs. Many computer games also use random-number generators to incorporate an element of unpredictability in the game.

EXAMPLE 9.6 Choose a simple random sample of size 5 from the following 12 semifinalists in a contest: Astoria, Beatrix, Charles, Delila, Elsie, Frank, Gaston, Heidi, Ian, Jose, Kirsten, and Lex.

SOLUTION First, assign numerical labels to the contestants. The labels must be two-digit numbers because more than 10 people are in the population. We might designate the semifinalists as follows.

00 = Astoria	01 = Beatrix	02 = Charles
03 = Delila	04 = Elsie	05 = Frank
06 = Gaston	07 = Heidi	08 = Ian
09 = Jose	10 = Kirsten	11 = Lex

Although we will generally be systematic in assigning labels to the elements in a population, the only requirement is that all elements have different labels with the same number of digits. For example, we could have used the set of two-digit numbers 01, 05, 06, 10, 11, 16, 17, 23, 24, 25, 30, 41 to represent the 12 people.

Next, we decide where to start and how to move through the random-number table. We could start anywhere, but let's start at the top of the third column of random numbers in Figure 9.4 with the number 47772 and read down the column. Because we want two-digit numbers, we may look at only the last two digits in each row of the column. (This choice is also arbitrary. We could have picked any two digits in the number.) Going down the column, we read 72, 57, 37, 01, 26, 06, 59, and so on.

Because of the labels we assigned to the people, we are interested only in numbers from 00 through 11. Eliminating all numbers larger than 11 in this column yields the numbers 01, 06, 10, and 11. We have four contestants so far, but since we need a sample of size 5, we need another number. If we look at the last two digits in the fourth column, we find 70, 69, and 07, so 07 is the fifth number in the range from 00 through 11. Three of our selections are illustrated in Figure 9.6. Numbers that are unacceptable because they are out of the desired range or because they are duplicates are crossed out in the table.

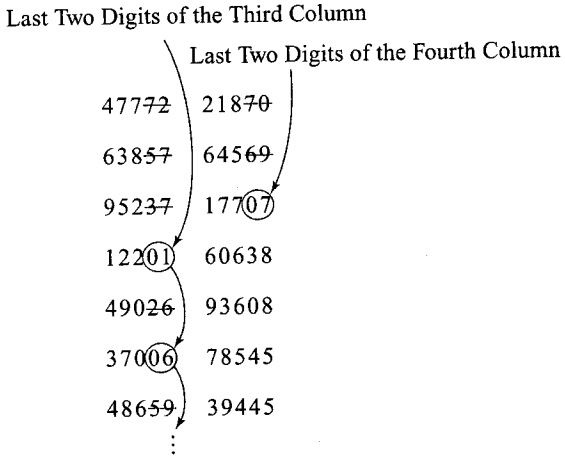


Figure 9.6

The numbers representing our simple random sample are 01, 06, 10, 11, and 07. Looking up the names of the corresponding contestants gives us our random sample of Beatrix, Gaston, Kirsten, Lex, and Heidi. ■

When using a random-number table, we customarily begin at a random position in the table and then select the numbers in a systematic way. For clarity in the examples, however, we usually begin the process in a convenient place, such as the top of a column or

the beginning of a row. We have used columns because they are easier to read, but we could have scanned across rows instead. The main requirement is that we systematically look at new random numbers, never looking at numbers in the table more than once.

EXAMPLE 9.7 Choose a simple random sample of size 8 from the states of the United States.

SOLUTION We will first assign numerical labels to the states. We could list the states in alphabetical order and assign labels that way or we could use some other list of the 50 states. Let's use the following list of states ranked by area, from largest to smallest (Table 9.2).

Table 9.2

State	Area Ranking	Area (square miles, including water)	State	Area Ranking	Area (square miles, including water)
Alaska	1	656,425	Iowa	26	56,276
Texas	2	268,601	New York	27	54,475
California	3	163,707	North Carolina	28	53,821
Montana	4	147,046	Arkansas	29	53,182
New Mexico	5	121,593	Alabama	30	52,423
Arizona	6	114,006	Louisiana	31	51,843
Nevada	7	110,567	Mississippi	32	48,434
Colorado	8	104,100	Pennsylvania	33	46,058
Oregon	9	98,386	Ohio	34	44,828
Wyoming	10	97,818	Virginia	35	42,769
Michigan	11	96,810	Tennessee	36	42,146
Minnesota	12	86,943	Kentucky	37	40,411
Utah	13	84,904	Indiana	38	36,420
Idaho	14	83,574	Maine	39	35,387
Kansas	15	82,282	South Carolina	40	32,007
Nebraska	16	77,358	West Virginia	41	24,231
South Dakota	17	77,121	Maryland	42	12,407
Washington	18	71,303	Hawaii	43	10,932
North Dakota	19	70,704	Massachusetts	44	10,555
Oklahoma	20	69,903	Vermont	45	9615
Missouri	21	69,709	New Hampshire	46	9351
Florida	22	65,758	New Jersey	47	8722
Wisconsin	23	65,503	Connecticut	48	5544
Georgia	24	59,441	Delaware	49	2489
Illinois	25	57,918	Rhode Island	50	1545

Alaska is first, so it is assigned 01 and Rhode Island is last, hence it is number 50.

Next we must decide on how to use the random-number table to choose our random sample. Let's start at the top row, left column, of the table in Figure 9.4 and proceed from

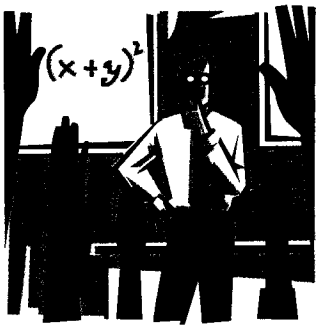
left to right using the last two digits from each entry. This procedure gives us the two-digit numbers 18, 95, 72, 70, 22, 45, 41, 95, 57, 69, 93, 29, 37, 68, 37, 07, 80, 55, 01, and so on. Since, we want eight numbers from 01 to 50, and we do not use repetitions, this leaves the following numbers: 18, 22, 45, 41, 29,37, 07, and 01 (check this).

We have eight different numbers in the desired range, so we look up the states to which they correspond. Our simple random sample consists of the following eight states:

18–Washington	22–Florida	45–Vermont	41–West Virginia
29–Arkansas	37–Kentucky	07–Nevada	01–Alaska

■

SOLUTION OF THE INITIAL PROBLEM



A university mathematics department is going to conduct a study on "Improving Problem-Solving Skills." The researchers will ask for student volunteers from a precalculus class, select a group of five students, and teach them some problem-solving techniques. Twenty-five of Professor Spark's students have indicated they would like to participate in the study. How can the professor select 5 students from the 25 volunteers in a fair way, so that no one can claim the professor showed favoritism?

SOLUTION Choose a simple random sample using a table of random numbers. Assign the 25 students the numbers 00, 01, . . . , 24 in order. Looking at the first two digits of each number in the last column of Figure 9.4 and going down that column, we obtain the numbers 99, 20, 04, 33, 49, 39, 29, 44, 77, 41, 54, 90, 70, 16, 07, 94, 62, 45, 06, 78, ... The first five numbers in this list that are 24 or less are 20, 04, 16, 07, and 06. The students that were assigned those numbers will be able to participate in the study.

PROBLEM SET 9.1

Problems 1 through 8

Identify the population being studied, the sample that is actually observed, and the variable.

1. A light bulb company says its bulbs last 2000 hours. To test this claim, an independent laboratory purchases a package of 8 bulbs, which are kept lit until they burn out. Five of the bulbs burn out before 2000 hours.
2. Divers recover a chest of 1000 gold coins from a sunken Spanish galleon found off the coast of Panama. The archaeologists working on the salvage project take 20 coins from the top of the chest and test them to see if they are pure gold.
3. The registrar's office at a university is interested in the percentage of full-time students who commute to school on a regular basis. One hundred full-time students are randomly selected and briefly interviewed. Of these students, 75 commute on a regular basis.
4. The mathematics department at a university is concerned about the amount of time mathematics students regularly set aside for studying. The department distributes a questionnaire in three mathematics classes having 82 students.
5. In Hewlett-Packard's 2003 annual survey, 4100 Hewlett-Packard customers were asked how they felt about their relationship with the company. Almost two-thirds of the customers believed they had a good relationship with Hewlett-Packard. *Source*: www.interex.org.
6. A newly developed biosensor called Cyranose is an electronic nose that is able to sniff out lung cancer. The device works by picking up the scent of compounds exhaled in the breath of patients with lung cancer. Doctors in Cleveland, Ohio, tested Cyranose on 59 people in the Cleveland area. Some of the people tested were patients with lung cancer, some had other lung cancer disorders, and some were healthy. The biosensor detected lung cancer successfully in the 14 patients who had lung cancer. *Source*: www.worldhealth.net.

7. Of the 7140 registered voters in a certain city, 3460 are Democrats, 3250 are Republicans, and 430 are Independents. A preelection canvassing of adults in a given neighborhood reveals the following numbers of registered voters: 185 Democrats, 210 Republicans, and 25 Independents.

8. There are 12,545 students attending a city's six high schools. The school district conducts a survey to determine students' access to the Internet. Of the 550 students contacted, 385 have a computer at home and 325 have Internet service at home.

9. For each immigrant entering the United States, the government creates a record that includes the country of origin, an identification number, and profession. Which of these variables are quantitative and which are qualitative?

10. The birth record of a baby includes the date and time of birth, the weight, the name of the baby, and the gender. Which of these variables are quantitative and which are qualitative?

11. An ecologist surveys trees in one acre of a forest, recording the location of each tree, the variety of tree (such as pine, oak, or Douglas fir), the approximate age, the approximate height, and the health of the tree (critical, poor, good, excellent). Which of these variables are quantitative and which are qualitative? Which of the qualitative variables are ordinal and which are nominal?

12. An investor sells a bond, and the total dollar amount of the transaction is recorded, as are the name of the bond and its rating. Which of these variables are quantitative and which are qualitative? Which of the qualitative variables are ordinal and which are nominal?

Problems 13 through 16

Identify and discuss any sources of bias in the sampling method.

13. A Minnesota-based toothpaste company claims that 90% of dentists prefer the formula in its toothpaste to any other. To substantiate this claim, the company conducts a study. Managers send questionnaires to 100 dentists in the Minneapolis–St. Paul area asking if they prefer the company's toothpaste formula to others.

14. A magazine devoted to exercise, vitamins, and healthy living is interested in the habits of older adults related to exercise and nutritional supplements. The current issue includes an article on the subject and a questionnaire for readers to fill out and mail in.

15. A soft drink company produces a lemon-lime drink that it says people prefer by a margin of two to one over its main competitor, a cola. To prove this claim, the company sets up a booth in a large shopping mall, where customers are allowed to try both drinks. The customers are filmed for possible inclusion in a television commercial and are asked which drink they prefer.

16. A sociologist working for a large school system is interested in demographic information about the families with children in the schools served by the system. Two hundred students are randomly selected from the school system's database and a questionnaire is sent to the home address of the parents or guardian.

Problems 17 through 20

Identify the population being studied and the sample actually observed. Discuss any sources of bias in the sampling procedure.

17. A biologist wants to estimate the number of fish in a lake. She catches 250 fish, tags them, and releases them back into the lake. Later, she catches 500 fish and finds that 18 of them are tagged.

18. A college professor is up for promotion. Teaching performance, as judged through student evaluations, is a significant factor in the decision. The professor has to choose one of his classes to complete student evaluations. The day of the evaluations, he passes out questionnaires and then remains in the room to answer any questions students might have about how to fill out the form.

19. A drug company wishes to claim that 9 of 10 doctors recommend the active ingredients in its product. It commissions a study of 20 doctors. If at least 18 doctors say they recommend the active ingredients in the product, the company will feel justified in making this claim. If not, the company will commission another study.

20. Two college students are running for student body president. Candidate Johnson believes that the student body resources should be used to enhance the social atmosphere of the college and that the first priority should be dances, concerts, and other social events. Candidate Jackson believes that sports should be the first priority and wants to use student body resources to subsidize student sporting events and enlarge the recreation facility. The student newspaper conducts a poll. An interviewer goes to a coffeehouse near the college one evening and asks students which candidate they prefer. Another interviewer goes to the gym and asks students which candidate they prefer.

21. Suppose city council members would like to survey a representative sample of city residents to determine whether they favor adding fluoride, a known tooth decay preventive, to the city's water system. The sampling could be done in any of the following three ways:

(i) All adults entering the city's public library on a Saturday could be questioned.

(ii) Ten city blocks could be randomly selected and every adult resident on each block questioned.

(iii) Two telephone numbers could be published in the newspaper. People who favor adding fluoride to the water system would call one telephone number while those who are against adding fluoride would call the other number.

a. For each of the three sampling methods, discuss possible sources of bias and then indicate which of the three methods would be likely to yield the most representative sample.

b. Describe a sampling method that might yield a more representative sample than any of the three described in this problem.

22. A school board would like to survey a representative sample of parents of children in the district to determine if parents would be willing to pay a book fee for needed textbook upgrades. The sampling could be done in any of the following three ways:

(i) Parents attending a Parent–Teacher Association meeting could be questioned.

(ii) A questionnaire could be left at the office of every school in the district so that parents visiting the school could fill out the survey.

(iii) At each school in the district, 20 parents could be questioned as they arrive to pick up their children after school.

a. For each of the three sampling methods, discuss possible sources of bias and then indicate which of the three methods would be likely to yield the most representative sample.

b. Describe a sampling method that might yield a more representative sample than any of the three described in this problem.

23. An instructor will randomly select 5 students from a class of 36. Each student is represented by a two-digit number from 00 to 35. Use the table in Figure 9.4 to select a sample of students by taking the last two digits of each random number. Begin with row 115 and proceed down the third column.

24. An automobile distributor received 80 new cars for a particular sales region. Ten cars will be randomly selected for detailed inspections before the shipment is finally accepted. The cars are numbered 00 to 79. Use the table in Figure 9.4 to select the sample of cars by taking the second and third digits of each random number. Begin with row 110 and proceed down the fourth column.

25. A university's science department has 250 graduate students. The dean will randomly select 10 of the graduate students and interview them about financial aid, program requirements, and other matters. The students are numbered 000 to 249. Use the table in Figure 9.4 to select the students by taking the first three digits of each random number. Begin with row 110 and proceed down the second column.

26. A pediatric dental group treats a combined total of 1146 patients. The patients are numbered 0000 to 1145. An independent auditor will conduct a thorough review of patient care and billing procedures on a random sample of 15 patients. Use the table in Figure 9.4 to select the sample of patients by taking the last four digits in each random number. Begin with row 130 and proceed down the first column.

27. Professional baseball has 14 American League player representatives, one for each of the American League teams. The 2004 player representatives are listed in the following table. These players serve as representatives in labor negotiations. Suppose union leaders randomly select a special committee of 5 players from the 14 player representatives. Explain how to generate the sample using a

American League	
Player	Team
Scott Schoeneweis	Anaheim Angels
Jason Johnson	Baltimore Orioles
Johnny Damon	Boston Red Sox
Jeff Liefer	Chicago White Sox
Charles Nagy	Cleveland Indians
Damon Easley	Detroit Tigers
Jason Grimsley	Kansas City Royals
Denny Hocking	Minnesota Twins
Mike Stanton	New York Yankees
Barry Zito	Oakland Athletics
Paul Abbott	Seattle Mariners
John Flaherty	Tampa Bay Devil Rays
Jeff Zimmerman	Texas Rangers
Vernon Wells	Toronto Blue Jays

Source: www.mlb.com.

random-number table, carefully describing the specific steps in your sampling procedure. Carry out your plan, and list the five players in your sample.

28. Professional baseball has 16 National League player representatives, one for each of the National League teams. The 2004 player representatives are listed in the following table. These players serve as representatives in labor negotiations. Suppose union leaders randomly select a special committee of 6 players from the 16 player representatives. Explain how to use a random-number table to generate the sample, carefully describing the specific steps in your sampling procedure. Carry out your plan, and list the six players in your sample.

National League	
Player	Team
Craig Counsell	Arizona Diamondbacks
Mike Remlinger	Atlanta Braves
Joe Girardi	Chicago Cubs
Aaron Boone	Cincinnati Reds
Todd Zeile	Colorado Rockies
Charles Johnson	Florida Marlins
Gregg Zaun	Houston Astros
Paul Lo Duca	Los Angeles Dodgers
Ray King	Milwaukee Brewers
Michael Barrett	Montreal Expos
Al Leiter	New York Mets
Doug Glanville	Philadelphia Phillies
Kevin Young	Pittsburgh Pirates
Steve Kline	St. Louis Cardinals
Kevin Jarvis	San Diego Padres
Russ Ortiz	San Francisco Giants

Source: www.mlb.com.

29. Based on figures from the 2000 census, the top 30 U.S. cities by population are listed in the following table. Suppose coordinators of a federally funded math program will select a sample of 5 of these cities to pilot the program in all elementary schools citywide.

a. Select a random sample of size 5 by beginning in column 2, row 117 of the table in Figure 9.4, and reading down the column selecting the second and third digits of each random number. List the cities in the sample.

b. Select a random sample of size 5 by beginning in column 5, row 140, and reading down the column selecting the last two digits of each random number. List the cities in the sample.

	City, State	Population in 2000
00	New York, NY	8,008,278
01	Los Angeles, CA	3,694,820
02	Chicago, IL	2,896,016
03	Houston, TX	1,953,631
04	Philadelphia, PA	1,517,550
05	Phoenix, AZ	1,321,045
06	San Diego, CA	1,223,400
07	Dallas, TX	1,188,580
08	San Antonio, TX	1,144,646
09	Detroit, MI	951,270
10	San Jose, CA	894,943
11	Indianapolis, IN	791,926
12	San Francisco, CA	776,733
13	Jacksonville, FL	735,617
14	Columbus, OH	711,470
15	Austin, TX	656,562
16	Baltimore, MD	651,154
17	Memphis, TN	650,100
18	Milwaukee, WI	596,974
19	Boston, MA	589,141
20	Washington, DC	572,059
21	Nashville-Davidson, TN	569,891
22	El Paso, TX	563,662
23	Seattle, WA	563,374
24	Denver, CO	554,636
25	Charlotte, NC	540,828
26	Fort Worth, TX	534,694
27	Portland, OR	529,121
28	Oklahoma City, OK	506,132
29	Tucson, AZ	486,699

c. Select a random sample of size 5 by beginning in column 1, row 102, and reading down the column selecting the third and fourth digits of each random number. List the cities in the sample.

d. Six of the 30 cities in the list are West Coast cities: Los Angeles, CA; San Diego, CA; San Jose, CA; San Francisco, CA; Seattle, WA; and Portland, OR. What fraction of cities in the list are West Coast cities? What fraction of cities in each sample in parts (a), (b), and (c) are West Coast cities?

e. Should one suspect bias in the sampling procedure if more than one West Coast city is selected? Explain.

30. Consider the 30 most populous cities in the United States as listed in the previous problem.

a. Select a random sample of size 10 by beginning in column 1, row 106 of the table in Figure 9.4, and reading down the column selecting the first two digits of each random number. List the cities in the sample.

b. Select a random sample of size 10 by beginning in column 4, row 127, and reading down the column selecting the second and third digits of each random number. List the cities in the sample.

c. Select a random sample of size 10 by beginning in column 3, row 111, and reading down the column selecting the first two digits of each random number. List the cities in the sample.

d. Nine of the 30 cities in the list have populations over 1 million. What fraction of cities in the list have populations over 1 million? What fraction of cities in each sample in parts (a), (b), and (c) have populations over 1 million?

e. Should one suspect bias if more than three cities with populations over 1 million are selected? Explain.

Extended Problems

31. The Nielsen Media Group selects households in the United States for its television-rating service. Results of the rating service determine which programs will be broadcast on television, which programs will be canceled, which programs will air during prime time, and so on. Research the Nielson Media Group. On the Internet, visit www.nielsenmedia.com. How many families are included in the rating service? Does the research group accept volunteers? Explain why or why not. How is the sample of families selected? Write a report summarizing your findings.

32. Many Internet news sites conduct opinion polls. For example, on the CNN site at www.cnn.com, readers may participate in "Quick Vote," a poll on issues currently in the news. Visit a news site with an opinion poll and participate in the current poll. Results of the "Quick Vote" are displayed with a disclaimer stating that it is "not a scientific survey." For the opinion poll in which you participated, describe the population, the sample, and the results of the poll. Are the results of the poll representative of Internet users in general? Explain.

33. Citizens of the United States have a civic duty to respond to a jury summons and, if chosen, to serve as a juror. How are citizens selected for jury duty? From what population is a sample of jurors taken? Is the jury a random sample of the population? The population of potential jurors may vary from state to state. Jurors are sometimes excused from jury duty. What excuses are routinely accepted? How does excusing some potential jurors from jury duty affect the selection process and the ability to determine whether the jury is representative of the population? Research juror-selection procedures in the county in which you live. Write a report to summarize your findings.

34. In order to select a random sample, we used a table of randomly generated numbers. If the numbers in the table in Figure 9.4 are indeed random, then we might expect each of the 10 digits 0 through 9 to occur approximately one-tenth of the time. Consider sets of numbers, such as census population data, tax-return data, baseball statistics, accounting balance sheets, or street addresses. It is commonly assumed that if numbers are sampled from a set of

data, then the digits 0 through 9 will all occur equally often in the first position of the number. In 1963, Dr. Frank Benford, a physicist at General Electric Company, discovered that first digits do not occur equally often in sets of data. In fact, Bedford found that 30% of the time, the first digit is a 1. He found that about 18% of the time, the first digit is a 2. This phenomenon is known as **Benford's law**.

a. Research Benford's law. How did Benford discover this pattern about the frequency of various digits? With what probabilities do other digits occur as first digits? How has Benford's law been applied to tax evasion and medical research fraud? For information on the Internet, search keywords "Benford's law." Summarize your findings in a report.

b. Obtain a large set of data such as the New York Stock Exchange page in your local newspaper. Randomly sample the stock prices and determine how many prices in the sample begin with the digit 1. How many begin with the digit 2? Do your results satisfy Benford's law? Explain.

9.2 Survey Sampling Methods

INITIAL PROBLEM

As a researcher at a consulting company, you must arrange interviews of at least 800 people nationwide to obtain marketing information for a manufacturer of DVD players. Assume that a different interviewer will be needed for each U.S. county, and each interviewer hired will cost $50 plus $10 for each person he interviews. Your budget is $15,000.

Before you can start arranging interviews, you must decide how to choose the people to be interviewed. You are considering two possible sampling methods and the associated costs, as follows.

1. a simple random sample of all adults in the United States
2. a simple random sample of adults in randomly selected U.S. counties

Which of the two methods should you choose, and why?
A solution for this Initial Problem is on page 585.

In the previous section, we discussed bias in sampling. One way to avoid such bias is to use random sampling. In the previous section, we considered the simple random sample, which is elementary in theory but can be expensive and time-consuming in practice. Statisticians developed the body of knowledge called **sample survey design** to provide alternatives to simple random sampling with the goal of collecting more information at less cost. In this section, we will present some terminology and methods used in survey sampling.

INDEPENDENT SAMPLING

A simple random sample gives a sample of a *fixed size* from a population. For example, suppose we wanted to use a simple random sample to select 50% of the customers coming into a store. We might record every customer's name and telephone number on a slip of paper and at the end of the day randomly select half the slips of paper. Thus, if 200 customers (the population) came into the store, we would randomly select 100 of the 200 slips of paper with name and address information (the sample). This procedure presents serious difficulties. For one thing, some customers may not want

to provide their names and phone numbers. Second, once the sample is chosen, those persons selected must be contacted and interviewed. It would have been a lot easier to do the interviews while the customers were in the store.

A more efficient way to sample 50% of the customers coming into the store would be to flip a coin for each customer and interview the customers for whom the coin came up heads. If 200 customers enter the store during the day, it would be unlikely that the coin would come up heads *exactly* 100 times, but there is a 50% chance that any individual customer would be interviewed. The coin-flipping procedure for choosing the sample is an example of independent sampling. In **independent sampling**, each member of the population has the same *fixed chance* of being selected for the sample regardless of whether other members of the population were selected, but the size of the sample cannot be fixed ahead of time. Because in this instance each customer has a 50% chance of being selected, we call this particular sample a **50% independent sample.**

EXAMPLE 9.8 Find a 50% independent sample of the 12 semifinalists—Astoria, Beatrix, Charles, Delila, Elsie, Frank, Gaston, Heidi, Ian, Jose, Kirsten, and Lex—from Example 9.6.

SOLUTION Instead of flipping a coin to pick the 50% independent sample, we will use a random-number table. Because the random-number table has 10 different digits, there is a 50% chance that one of the five digits 0, 1, 2, 3, or 4 will occur at any spot in the table. Likewise, there is a 50% chance that one of the five digits 5, 6, 7, 8, or 9 will appear. Thus, we let the digits 0 through 4 represent "select this contestant" and let the other digits represent "do not select this contestant." Then it is just as likely that a contestant will be selected as not. (Our choice of digits 0–4 to represent "select" is arbitrary.)

We arbitrarily choose column 6 of the table in Figure 9.4, and, starting at the top, look at the first 12 digits because we have 12 semifinalists. The first 12 digits of the column are 99445 20429 04. Each digit determines whether the contestant in that position will be chosen. For example, the first digit is a 9, so we will not choose the first contestant, Astoria. The second digit is also a 9, so we will not choose the second contestant, Beatrix. The third digit in our sequence of random numbers is a 4, so we will choose the third contestant, Charles. Continuing in this manner, we see that the contestants in our 50% independent sample are Charles, Delila, Frank, Gaston, Heidi, Ian, Kirsten, and Lex. ■

Note that the sample selected in Example 9.8 contained 8 out of 12 or $66\frac{2}{3}\%$ of the population, but nonetheless it is considered a 50% independent sample because each semifinalist had a 50% chance of being selected. In general, a 50% independent sample may contain more than 50%, less than 50%, or exactly 50% of the population.

Tidbit

The Internal Revenue Service picks an independent sample of tax returns for audit. Customs agents also frequently choose an independent sample of people entering the country to check for smuggled contraband.

EXAMPLE 9.9 Suppose a factory produces 100 automobiles in 1 day. Use the random-number table to choose a 10% independent sample of the automobiles produced that day.

SOLUTION Let's number the automobiles produced in a single day as 1, 2, . . . 100. To choose a 10% independent sample, we must find a way to give each car a 10% chance of being selected using the random-number table. Each of the 10 digits 0, 1, 2 . . . 9 has the same chance of appearing in a random-number table. Thus, the digit 0 occurs 10% of the time, so we will let the digit 0 represent "select this car."

We must now decide which portion of the table to use. Let's consider only the numbers in the five-digit-wide first column and ignore the rest of the table. We will read down the column, starting in row 101, going from the left to the right, row by row. The first digit is 0, so we choose automobile 1 as part of the sample. The next digit is a 3, so we do not choose automobile 2 as part of the sample. Continuing in the same way, we see

that of the first 100 digits in column 1, the digit 0 occurs at positions 1, 7, 8, 19, 33, 39, 62, 70, 73, 81, 88, 93, 95, 98, 100. Thus, we select the automobiles with those labels. These choices are illustrated in Figure 9.7. Shaded squares represent the cars that are selected.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 9.7 ■

In Example 9.9, the sample chosen consists of 15 cars, or 15% of the population. Nonetheless, our selection is a 10% independent sample because any particular car had a 10% chance of being selected.

SYSTEMATIC SAMPLING

In **systematic sampling**, we decide ahead of time what proportion of the population we wish to sample. For example, suppose we wish to select 1 of every 10 elements of the population. We call this a 1-in-10 systematic sample. We choose one of the numbers 1, 2, 3, . . . 10 at random. Suppose 7 is the number we pick at random. Working from a list of all elements of the population, we select the 7th element from every group of 10 elements, that is, we select the 7th element, the 17th element, the 27th element, and so on.

In general, a **1-in-*k* systematic sample** is selected in the following way. First we randomly choose one of the numbers from 1 to k and designate that randomly selected number as r. In the preceding example, k was 10 and r was 7. Then our 1-in-k sample consists of the following elements: rth, $(r + k)$th, $(r + 2k)$th, etc.

EXAMPLE 9.10 Suppose a factory produces 100 automobiles in 1 day. Use systematic sampling to select a 1-in-10 systematic sample of the automobiles produced in 1 day for quality-control testing.

SOLUTION Let's label the automobiles produced in a single day as 1, 2, . . . 100 and use the random-number table to choose our 1-in-10 systematic sample. The random number we pick from 1 to 10 will serve as our value of r. Suppose we use the last digit in the first row of the table in Figure 9.4, which happens to be 5. So, we have $r = 5$. (Note: Had the last digit in the row been a "0", we would have let $r = 10$.)

Now that we know $r = 5$ and $k = 10$, we know we will choose cars labeled 5 (because $r = 5$), 15 (because $r + k = 5 + 10 = 15$), 25 (because $r + 2k = 5 + 20 = 25$), and so on. Our systematic sample will consist of cars 5, 15, 25, 35, 45, 55, 65, 75, 85, and 95. These choices are illustrated in Figure 9.8. Notice how this representation differs from the visual representation of the 10% independent sample that was shown in Figure 9.7.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 9.8

In Example 9.10, we selected *exactly* 1 of every 10, or 10%, of the 100 automobiles. By contrast, in Example 9.9, we picked a total of 15 automobiles, but the likelihood of picking any given car was 10%.

A systematic sample is easier to choose than an independent sample. However, the regularity in the selection of the elements in a systematic sample can be a source of bias. For example, a 1-in-7 systematic sample of a daily phenomenon may be biased by always occurring on the same day of the week.

QUOTA SAMPLING

For a geographically dispersed population such as registered voters in the United States, both simple random sampling and independent sampling are difficult and expensive to carry out. On the other hand, pollsters, politicians, lobbyists, market analysts, and others are extremely interested in determining voters' opinions on a wide range of issues. The goal of a public opinion pollster is to satisfy the desire for an accurate assessment of the voters' opinions rapidly and at reasonable cost. One method that achieves this practical goal is quota sampling.

U.S. Census data provide a profile of the population with respect to a number of variables. For example, according to the *Statistical Abstract of the United States*, 12.7% of the population of the United States in 2002 was African American. A representative sample of the population of the United States, therefore, should be 12.7% African American. We often want to consider characteristics such as race and gender in order to select a sample that closely resembles the population from which it was drawn. **Quota sampling** forces the sample to be representative for known important variables by requiring that quotas are filled for respondents in various categories. If we select a representative sample of the U.S. population, 12.7% of the selected respondents should be African American. Similar quotas would also be set for other variables thought to be important, such as gender, age, occupation, and so on. George Gallup introduced quota sampling in the 1930s and used it to predict successfully the winner of the presidential elections of 1936, 1940, and 1944.

If you wished to use quota sampling to gauge student opinion on some university issue, you might interview people passing by a busy location on campus. Assuming that a student's major may be an important variable, you would want to ensure that you interviewed students with various majors in proportion to the number of students with those majors. Of course, the student's gender, age, or some other unanticipated characteristic might be more important than the student's major. Here we see one of the difficulties with quota sampling: There is no sure way to know ahead of time which variables are sufficiently important to require quotas.

Gallup used quota sampling to predict that Thomas Dewey would win the 1948 presidential election with 50% of the vote, compared with 44% of the vote for incumbent president Harry Truman (other candidates accounting for the remaining votes). Instead, President Truman was reelected with 50% of the vote compared to Dewey's 45%. This stunning failure led to a reassessment of scientific polling methods.

STRATIFIED SAMPLING

Recall that a population consists of the entire group (such as people, objects, or events) under study. It is likely that the population will not be homogeneous, especially if it consists of people. For instance, men may differ from women, high-school graduates may differ from college graduates, and Democrats may differ from Republicans in how they view a particular issue. Stratified sampling is another method of sampling that takes into account the variations in the groups that make up a population, while avoiding the bias that can occur in quota sampling. In **stratified sampling**, the population is subdivided into two or more nonoverlapping subsets, each of which is called a **stratum** (see Figure 9.9). Ideally, the strata should be chosen so that they are more homogeneous than the entire population. For example, a pollster might divide the population of registered voters into three nonintersecting groups on the basis of age: stratum 1 might be voters aged 18 to 30, stratum 2 voters 31–50, and stratum 3 voters over 50.

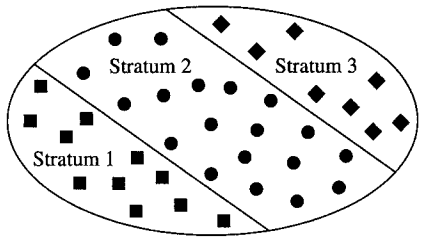


Figure 9.9
Population

Tidbit

In agriculture, farmers are increasingly using hand-held global positioning system (GPS) receivers. The GPS receiver can accurately locate the farmers' position in a field to within about 7 feet. The system can be used to calculate the size of a field, plot an outline, create a grid for sampling, select a sample, and identify which sampling unit the farmer is in. Farmers use the system to mark crop varieties, sample for pest infestations or plant disease, and note problem areas.

A **stratified random sample** is obtained by selecting a simple random sample from each stratum. For example, the pollster who divided voters into three age categories (strata) would choose a stratified random sample of voters by selecting a simple random sample from each age group. A stratified random sample can be less costly because the homogeneity of the strata allows a smaller sample to be used.

EXAMPLE 9.11 Select a stratified random sample of 10 men and 10 women from a population of 200. Suppose there are equal numbers of men and women in the population. Use the first two digits of the second and third columns in the table in Figure 9.4, beginning in row 101, for selecting men and women, respectively.

SOLUTION In this example, there are two strata: men and women. We must select a simple random sample from each of these groups. We number the 100 men 01, 02, . . . 99, 00. Reading down the first two digits of the second column of the table in Figure 9.4 to select a simple random sample of the men, we get the following ten 2-digit numbers: 77, 31, 25, 66, 49, 38, 00, 95, 24, and 57. These are the men selected for the male portion of our sample.

Similarly, we number the 100 women 01, 02, . . . 99, 00. Reading down the first two digits of the third column of the table in Figure 9.4 to select a simple random sample of the women, we get the following ten 2-digit numbers: 47, 63, 95, 12, 49, 37, 48, 94, 35, and 78. Combining these two groups gives us our stratified random sample of 10 men and 10 women. A visual representation of our stratified random sample is shown in Figure 9.10, where the shaded squares indicate individuals who were selected.

Men

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Women

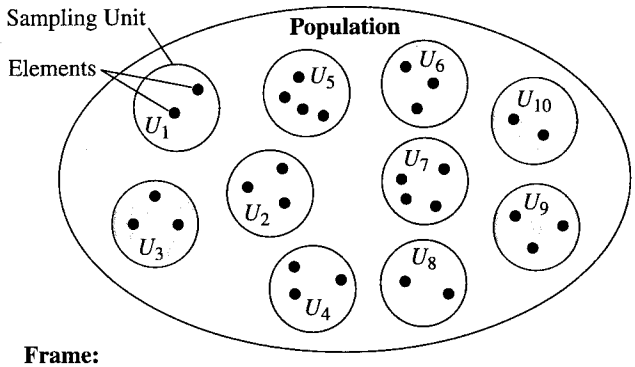
1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Figure 9.10

CLUSTER SAMPLING

It is often convenient to group elements of the population together, especially for a geographically dispersed population. For example, it may be easier to survey households than individual voters. We could choose a sample of households and then interview each member of that household. The first step in this process is dividing the entire population into households. Such nonoverlapping subsets of the population are called **sampling units** or **clusters**. Sampling units may vary in size, and, in principle, may consist of a single element. For instance, one household may consist of 10 persons, while another household may have only one person.

A **frame** is a complete list of the sampling units and a **sample** is a collection of sampling units selected from the frame. For instance, the frame might be all the households in a particular city, and then the sample would be the households selected to be interviewed. A population, its elements, the sampling units, the frame, and a sample are illustrated in Figure 9.11. The shaded circles represent the sample. Notice that different sampling units contain different numbers of elements.



Frame:
$U_1, U_2, U_3, U_4, U_5, U_6, U_7, U_8, U_9, U_{10}$

Sample:
U_3, U_9, U_{10}

Figure 9.11

In **cluster sampling**, a simple random sample determines the sampling units to be included in the sample. Typically, a cluster will be geographically small, and cluster sampling will be used when the cost of obtaining measurements increases with increasing distance.

If you wish to use cluster sampling to gauge student opinion on some university issue, you might use student residences as a device for simplifying the interviewing process. For example, if all the students live in dormitories, then each floor of a dorm could serve as a sampling unit. Then a frame would be a list of all floors in all dormitories on campus. To carry out the cluster sampling, a simple random sample of sampling units (floors) would be chosen from that list of dormitory floors. Interviewers would poll all the residents of the selected floors.

EXAMPLE 9.12 Select a cluster sample of 12 individuals from a population of 96 people who all live in four-person suites. Use the first two digits of the fourth column of the table in Figure 9.4 as a source of random numbers, starting in row 101.

SOLUTION The sampling units will be the 24 four-person suites, which we will number 01 through 24. We need a simple random sample of three of these suites. Reading down the first two digits of the fourth column of the table in Figure 9.4, the first three two-digit numbers we find are 21, 64, and 17. We do not find another two-digit number in the range 01 to 24 other than 17 until we get to the 25th row, where we find 10. Thus, the selected suites are numbers 10, 17, and 21. These suites comprise our sample. Figure 9.12 shows a graphical representation of the sample, in which the suites are numbered, and shading indicates selected suites and individuals.

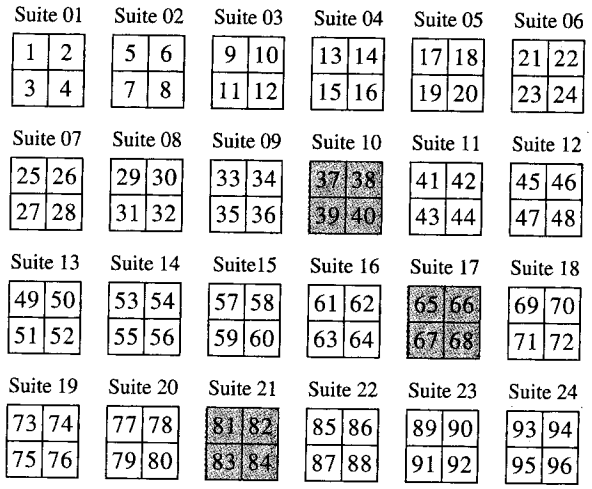


Figure 9.12

At first glance, it may appear that the method of cluster sampling is very much like stratified random sampling. The processes of cluster sampling and stratified random sampling are similar in that each divides the population into subsets: sampling units for cluster sampling and strata for stratified random sampling. The distinction between the sampling units of cluster sampling and the strata of stratified sampling is that every element in every selected sampling unit will be measured, but only a selected sample from each stratum will be measured. When sampling units are defined, they should be small enough that every element in a sampling unit can be measured, but the individuals in a sampling unit need not have anything more in common than that they live near each other. On the other hand, strata should be defined so that the individuals in them are similar in some important way, such as gender or ethnicity. Typically, cluster sampling involves many sampling units, but each sampling unit will contain only a few individuals. In contrast, stratified random sampling involves only a few strata, but each stratum will contain many individuals.

SUMMARY

Collecting a simple random sample from a given population can be prohibitively difficult or expensive. The sampling methods discussed in this section can overcome these problems. Independent sampling and systematic sampling are used when the population elements arrive in sequence so that the cost of gathering the sample can be significantly reduced by using one of these sampling methods. If the population can be subdivided into a small number of categories of similar individuals, we might choose to use the method of stratified sampling. When a list of all population elements is unavailable or it is possible to reduce the cost of sampling by reducing travel costs incurred in collecting the data, we might use cluster sampling.

Table 9.3 summarizes the kinds of samples we have discussed in this chapter.

Table 9.3

Type of Sample	Description	Fixed Sample Size?
Simple random sample	Draw a sample of a given size from the entire population using a predetermined random method—similar to "drawing names from a hat."	Yes
Independent sample	Select a sample so that each member of the population has the same predetermined chance of being selected.	No
1-in-k systematic sample	Order the population into groups of size k and then select the member in the same position of each group.	Yes, if population size is known
Quota sample	Establish quotas so the sample models the population on one or more important characteristic.	Yes
Stratified random sample	Divide population into strata based on some characteristic and take a simple random sample of each stratum.	Yes
Cluster sample	Divide population into clusters, select a sample of clusters, and measure all members of those clusters.	Depends on makeup of sampling units

SOLUTION OF THE INITIAL PROBLEM

As a researcher at a consulting company, you must arrange interviews of at least 800 people nationwide to obtain marketing information for a manufacturer of DVD players. Assume that a different interviewer will be needed for each U.S. county, and each interviewer hired will cost \$50 plus \$10 for each person he interviews. Your budget is \$15,000.

Before you can start arranging interviews, you must decide how to choose the people to be interviewed. You are considering two possible sampling methods and the associated costs, as follows.

1. a simple random sample of all adults in the United States
2. a simple random sample of adults in randomly selected U.S. counties

Which of the two methods should you choose, and why?

SOLUTION A simple random sample is unbiased, so option (1) might seem to be the best choice. Unfortunately, as we shall see, your budget is not large enough for this kind of sample. The United States has 3130 counties; most people selected to be interviewed are likely to live in different counties. Suppose that these 800 people live in only 400 different counties. Then the cost of hiring the interviewers would be (400)(\$50) = \$20,000. You will also need to pay \$10 for each interview conducted, that is, (800)(\$10) = \$8000. The total cost of this simple random sample then would be \$20,000 + \$8000 = \$28,000, which is more than your budget of \$15,000.

Now we consider option (2), which we will see is more economical. Suppose you were to first use a simple random sample to choose 100 counties and that next you choose a simple random sample of 8 people in each county. The cost of hiring the interviewers would be (100)($50) = $5000 for the 100 different counties. Again, you will need to pay $10 for each interview conducted, that is, (800)($10) = $8000. Thus, the total cost under this plan would be $5000 + $8000 = $13,000, which is $2000 under your budgeted amount.

Option (2) is an example of **multistage sampling**, which is a sampling method that uses successive applications of the sampling methods we have discussed.

PROBLEM SET 9.2

1. For each of the following samples, indicate which sampling technique was used.
   - **a.** A newspaper randomly selected 80 urban and 80 rural residents and interviewed them about the governor's new tax proposal.
   - **b.** A scientist surveyed every seventh person entering a fast-food restaurant about his or her sleeping habits.
   - **c.** A farmer divided a map of field corn into nonoverlapping regions. He randomly selected six of the regions and examined all the corn plants in each region for pest infestation.
   - **d.** Forty percent of women who gave birth in a certain hospital had cesarean sections. An independent analyst surveyed 300 women who recently gave birth to assess their level of satisfaction with the care they received. Of the 300 women in the sample, the analyst randomly selected 120 women from the group who had cesarean sections, and 180 women from the group who did not.
2. For each of the following samples, indicate which sampling technique was used.
   - **a.** A quality-control inspector selected every 20th DVD player as it came off an assembly line.
   - **b.** A consulting firm randomly selected 90 patients of all patients who were treated at a certain hospital in the past year and interviewed them about their satisfaction with patient care.
   - **c.** A city planner divided a city into parcels measuring 1 city block by 1 city block. Fifty parcels were randomly selected and everyone in each parcel was interviewed about a recent flood.
   - **d.** An airport security guard rolled a die for every person who passed through a security check point. If the die landed with a 1 showing, the guard selected that person for a more detailed security screening.
3. An opinion pollster will take a sample of people entering a shopping mall. For each person passing through the door, the pollster will flip a coin. If the coin shows heads, then the person will be selected for the sample. Suppose the coin is flipped 200 times and shows heads 75 times.
   - **a.** What percent of the first 200 people entering the mall are included in the sample?
   - **b.** Is this a 50% independent sample? Explain.
4. A border guard at a certain checkpoint will take a sample of cars passing from the United States into Canada. For each car that stops at the checkpoint, the guard will flip a coin. If the coin shows tails, then the car will be selected for the sample. Suppose the coin is flipped 65 times and shows tails 48 times.
   - **a.** What percent of the first 65 cars stopping at the checkpoint are included in the sample?
   - **b.** Is this a 50% independent sample? Explain.
5. A jury-duty coordinator will send notices to a sample of the 2345 registered voters of a small town. In order to have a sufficiently large pool of potential jurors, the coordinator has to send notices to 20% of the registered voters. Explain why a 20% independent sample might not be a good choice of method for this jury-duty selection process.
6. A pollster will conduct an opinion poll to assess the approval rating for the governor of Utah. The pollster will select at least 1000 of the 1 million registered voters in Utah to participate in the poll. Explain why a 0.1% independent sample might not be a good choice of method for this opinion poll selection process.

7. In Example 9.9, we used the first column of digits to find a 10% independent sample from a set of 100 automobiles. Repeat the process illustrated in the example using 0 to mean "select this car," but this time find a 10% independent sample using the third column of digits from the table in Figure 9.4, beginning at the top of the column.

8. Find a 20% independent sample of the letters of the alphabet (A = 1, B = 2, . . . Z = 26) using the third column of digits from the table in Figure 9.4, beginning at the top of the column and reading row by row down the column. Use the digits 1 and 2 to mean "select this letter."

Problems 9 through 14

Governors for each of the 50 United States in 2004 are given in the following table. Label the governors alphabetically by state, as listed in the table, so that the Alabama governor Robert Riley is 1 and continue until the Wyoming governor is 50.

9. Find a 20% independent sample from the governors of the United States in 2004. Use the digits 0 and 1 to mean "select this governor," and proceed down the second column of the table in Figure 9.4, beginning in row 101 and going from left to right.

10. Find a 20% independent sample from the governors of the United States in 2004. Use the digits 8 and 9 to mean "select this governor," and proceed down the sixth column of the table in Figure 9.4 beginning in row 101 and going from left to right.

U.S. GOVERNORS, 2004	
Alabama: Robert Riley	Montana: Judy Martz
Alaska: Frank Murkowski	Nebraska: Mike Johanns
Arizona: Janet Napolitano	Nevada: Kenny Guinn
Arkansas: Mike Huckabee	New Hampshire: Craig Benson
California: Arnold Schwarzenegger	New Jersey: James McGreevey
Colorado: Bill Owens	New Mexico: Bill Richardson
Connecticut: John Rowland	New York: George Pataki
Delaware: Ruth Ann Minner	North Carolina: Michael Easley
Florida: Jeb Bush	North Dakota: John Hoeven
Georgia: Sonny Perdue	Ohio: Bob Taft
Hawaii: Linda Lingle	Oklahoma: Brad Henry
Idaho: Dirk Kempthorne	Oregon: Ted Kulongoski
Illinois: Rod Blagojevich	Pennsylvania: Edward Rendell
Indiana: Joseph Kernan	Rhode Island: Don Carcieri
Iowa: Thomas Vilsack	South Carolina: Mark Sanford
Kansas: Kathleen Sebelius	South Dakota: Mike Rounds
Kentucky: Ernie Fletcher	Tennessee: Phil Bredesen
Louisiana: Kathleen Blanco	Texas: Rick Perry
Maine: John Baldacci	Utah: Olene Walker
Maryland: Robert Ehrlich	Vermont: James H. Douglas
Massachusetts: Mitt Romney	Virginia: Mark Warner
Michigan: Jennifer Granholm	Washington: Gary Locke
Minnesota: Tim Pawlenty	West Virginia: Bob Wise
Mississippi: Haley Barbour	Wisconsin: Jim Doyle
Missouri: Bob Holden	Wyoming: Dave Freudenthal

Source: National Governors Association.

11. **a.** Find a 30% independent sample from the governors of the United States in 2004. Use the digits 1, 2, and 3 to mean "select this governor." Use the second column of the table in Figure 9.4. Read down the column beginning in row 130, and go from left to right, row by row.

b. What percentage of governors was actually included in your sample from part (a)?

12. **a.** Find a 50% independent sample from the governors of the United States in 2004. Use the digits 0, 1, 2, 3, and 4 to mean "select this governor." Use the third column of the table in Figure 9.4. Read down the column beginning in row 130, and go from left to right, row by row.

b. What percentage of governors was actually included in your sample from part (a)?

13. Suppose you took an independent sample with the resulting sample yielding only the governors from Illinois and Missouri. To generate the independent sample, you used the following entries from a random-number table.

04212	40076	30584	93939	39665
74781	88341	03617	10099	88296

Describe this independent sample in terms of percentages by filling in the blank below.

This sample was a ___% independent sample.

14. Suppose you took an independent sample and the result was that the sample consisted of the governors from Arizona, Connecticut, New Jersey, New Mexico, and Virginia. To generate the independent sample, you used the following entries from a random-number table.

77175	80925	23629	77764	75867
32931	07544	95693	34646	13994

Describe this independent sample in terms of percentages by filling in the blank below.

This sample was a ___% independent sample.

15. Suppose you use a 1-in-15 systematic sampling to pick a sample from 180 printers coming off an assembly line and that you pick the number 9 at random to start the systematic sample.

a. What is the value of k in this sampling process and what is its significance?

b. What is the value of r in this sampling process and what is its significance?

c. If the printers are numbered 1 to 180, which printers are included in the sample?

16. Suppose you use a 1-in-20 systematic sampling to pick a sample from 250 blocks of cheese on a conveyer belt and that you pick the number 11 at random to start the systematic sample.

a. What is the value of k in this sampling process and what is its significance?

b. What is the value of r in this sampling process and what is its significance?

c. If the blocks of cheese are numbered 1 to 250, which ones are included in the sample?

17. Use 1-in-10 systematic sampling to pick a sample from the 50 governors shown in the table after problem 10. Use the fourth digit from the first column in row 128 of the table in Figure 9.4 to start the systematic sample.

18. Use 1-in-10 systematic sampling to pick a sample from the 50 governors shown in the table after problem 10. Use the fifth digit from the sixth column in row 136 of the table in Figure 9.4 to start the systematic sample.

19. Pick a sample of letters of the alphabet using 1-in-5 systematic sampling. Use the table in Figure 9.4, beginning in column 3 and row 102. Select the first digit from 1 through 5 in the table to start the systematic sampling. Label the letters of the alphabet, using the natural order of the alphabet, with A = 1, and list the letters that are included in the sample.

20. Pick a sample of letters of the alphabet using 1-in-3 systematic sampling. Use the table in Figure 9.4 and begin in column 6 and row 115. Select the first digit from 1 through 3 in the table to start the systematic sampling. Label the letters of the alphabet in alphabetical order with A = 1, and list the letters that are included in the sample.

21. Suppose a theater owner hands a questionnaire to the 9th, 19th, 29th, 39th, 49th, and 59th adults who leave the theater.

a. What kind of survey did the theater owner conduct? Be specific.

b. Can you determine how many adults were in the theater? Explain.

22. Suppose a catalog company records the 8th, 23rd, 38th, . . . , and 208th calls to its customer service department in 1 day.

a. What kind of survey did the company conduct? Be specific.

b. Can you determine how many calls the company made on that day? Explain.

23. In 2002, approximately 141,661,000 males and 146,708,000 females were living in the United States, according the U.S. Census Bureau. If you plan to conduct a quota sample of size 800 such that the percentage of males and females in the sample is the same as the percentage in the general population, how many males and how many females should you include in the sample?

24. In 2002, approximately 288,369,000 people were living in the United States. The following table contains population information listed by race. If you plan to conduct a quota sample of size 5000 such that the percentages of each race in the sample is the same as the percentages in the general population, then how many people of each race should you include in the sample?

Race	Population
African American	36,746,000
American Indian and Alaska Native	2,752,000
Asian	11,559,000
Caucasian	232,647,000
Native Hawaiian and Pacific Islander	484,000
Two or more races	4,181,000

Source: U.S. Census Bureau

25. In 2002, according to the U.S. Census Bureau, approximately 18.8% of the population lived in the Northeast, 22.6% lived in the Midwest, 35.8% lived in the South, and 22.8% lived in the West. Suppose an opinion poll uses a quota sample in which the percentages of people in the sample living in each region of the country are the same as the percentages in the general population. How might the results of the survey be affected if all the people surveyed from the South lived in Florida?

26. Suppose an opinion poll uses a quota sample in which the percentages of males and females in the sample are the same as the percentages in the general population. How might the results of the survey be affected if all the males included in the survey are interviewed by phone between the hours of 8 A.M. and 3 P.M.?

27. Suppose that a class consists of 80 women and 80 men. You wish to survey the class to determine which movies to show in a foreign film series. Since men and women may have different tastes, you decide to take a stratified sample of 10 men and 10 women from the class.

a. Identify the strata in this sample.

b. Number the men from 01 to 80 and the women from 01 to 80. Use the table in Figure 9.4 to choose the sample. Use the second and third digits of column 2 for men and the second and third digits of column 3 for the women. Begin in row 113 in each case, and read down the column.

28. Suppose there are 240 freshmen, 220 sophomores, 232 juniors, and 184 seniors in a small college. You plan to take a stratified random sample of 4 focus groups of size 6 students from each class.

a. Identify the strata in this sample.

b. Number the students in each class with three digits, beginning with 001. Use the table in Figure 9.4 to select the samples. Begin using the first three digits of column 2 in row 107 and read down the column to select the sample of freshmen. After you've selected the last freshman, begin on the next row to pick the sample of sophomores, and continue in this manner.

29. A small college has an enrollment of 2000. Of these, 950 are freshmen and sophomores, 800 are juniors and seniors, and 250 are graduate students. The administration takes a stratified random sample of size 40 to ask their opinion about a proposed "technology fee" for upgrading computer facilities.

a. Identify the strata in this sample and comment on the likelihood that members of each stratum will have opinions that are more homogeneous than the general population.

b. The administration wants the proportion of each stratum in the sample to be the same as in the population. How many students should be selected from each stratum? Explain your reasoning.

c. Number the freshmen and sophomores from 001 to 950, and the other groups similarly. Use your answer in (b) and the table in Figure 9.4 to select the samples. Use column 1 for the freshmen and sophomores, column 3 for the juniors and seniors, and column 5 for the graduate students. Begin with the first three digits in row 105 in each case and read down the column.

30. An obstetrician has 156 expectant patients. Of the obstetrician's patients, 75 are expecting their first child, 54 their second, and 27 their third. The doctor would like to take a stratified random sample of 25 of her patients to ask their opinion about a new type of pain relief drug available to women in labor.

a. Identify the strata in this sample and comment on the likelihood that members of each stratum will have opinions that are more homogeneous than the general population.

b. The doctor wants the proportion of each stratum in the sample to be the same as in the population. How many patients should be selected from each stratum? Explain your reasoning.

c. Number the patients who are expecting their first child from 01 to 75, and the other patients similarly. Use your answer in (b) and the table in Figure 9.4 to select the samples. Use column 2 for the patients expecting their first child, column 3 for the patients expecting their second child, and column 4 for the patients expecting their third child. Begin with the last two digits in row 117 in each case and read down the column.

31. The Albany College of Pharmacy has a student body of approximately 700. Suppose the campus dormitory houses three students in each of the 80 rooms. A student will conduct a survey to determine dormitory residents' opinions about a campus issue. The student will interview a total of 60 residents from the dormitory. Use cluster sampling to select the sample.

a. Identify the sampling units and determine how many sampling units will be selected.

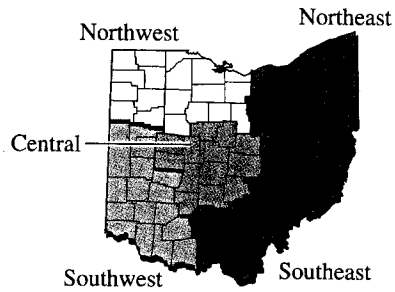
b. Number the rooms 01 to 80. Use the second and third digits of column 2 in the table in Figure 9.4, beginning in row 115 going down the column. Which rooms are selected in the sample?

32. As part of a research project, you will investigate how many chocolate chips are in Moonbeam Chocolate Chip Cookies. The nearby convenience store has 30 packages of these cookies, and each package contains 12 cookies. You will examine a total of 72 cookies. Use cluster sampling to select the sample.

a. Identify the sampling units and determine how many sampling units will be selected.

b. Number the packages 01 to 30. Use the fourth and fifth digits of column 4 in the table in Figure 9.4, beginning in row 128 going down the column. Which packages are selected in the sample?

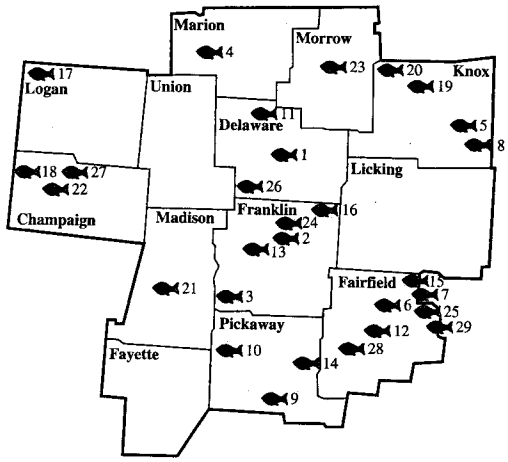
33. In the state of Ohio, people aged 16 to 66 must have a fishing license in order to fish in any public water. Suppose the Ohio Division of Wildlife wants to investigate the use of fishing licenses. The following map shows how Ohio is broken into five regions: Northwest, Northeast, Southwest, Southeast, and Central. Each Ohio region is broken up into counties.



Source: www.dnr.state.oh.us/wildlife/fishing/lakemaps/lmaps.htm.

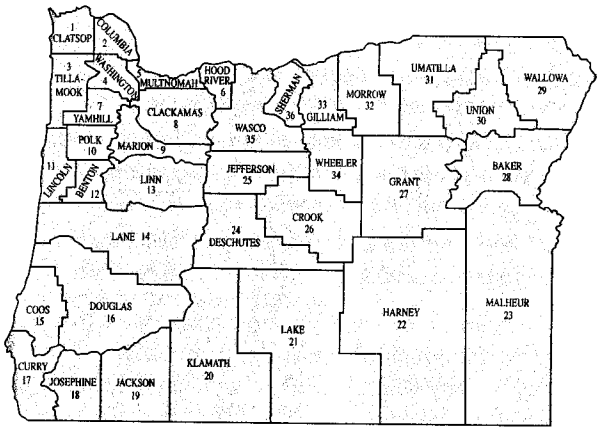
a. Describe how to use cluster sampling to study the use of public water fishing licenses if regions are used as clusters.

b. The following map shows all 13 counties and 29 public waters in the central region of Ohio. The numbered fish in the map indicate the locations of public waters in each county. Give at least two reasons why cluster sampling, with counties as clusters, might be a poor sampling choice for this study.



c. Select a simple random sample of 10 public waters in central Ohio. Use the second and third digits of the second column beginning in row 125 of the table in Figure 9.4 and go down the column.

34. Suppose the Department of Agriculture in the state of Oregon wants to monitor pesticide use for field crops by farms in the state. The following map numbers the 36 counties of Oregon. Over 40,000 farms are in Oregon, with some farms in each of the 36 counties.

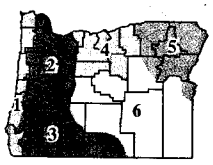


Source: U.S. Census Bureau, County Maps

a. Explain why the Department of Agriculture might want to use cluster sampling rather than taking a simple random sample of farms.

b. Use the first and second digits of column 4 in the table in Figure 9.4, beginning in row 122, going down the column to select a cluster sample of 10 counties.

c. The following map shows Oregon's growing regions; the list describes the main farming industry for each region. Explain how stratified random sampling could be used to investigate pesticide use.



Region 1: Coastal Dairy Farms

Region 2: Very diverse 170 different crops

Region 3: Fruit Trees, Potatoes, Livestock

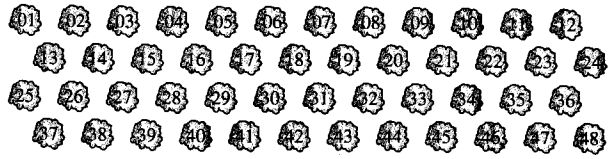
Region 4: Fruit Trees

Region 5: Wheat, Potatoes, Alfalfa, Carrots, Watermelon, Onion, Field Corn

Region 6: Livestock, Onions, Potatoes, Sugarbeets

Source: Oregon Department of Agriculture

35. A farmer maps her fruit tree orchard and numbers the trees as shown in the following diagram. To monitor for disease, she will take a sample of trees and visually inspect each tree in the sample. It will take 30 minutes per tree to examine for disease, at a cost of $35 per hour.



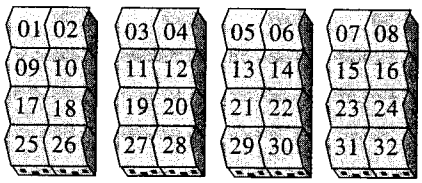
a. Select a 40% independent sample. Use the table in Figure 9.4. Begin in column 1 of row 109 and read across the row. Let the digits 1, 2, 3, and 4 indicate that the tree will be selected for the sample. List the trees that you selected and shade the selected trees in the diagram. What percentage of trees did you did you select? How much will it cost to inspect the trees in the sample?

b. Select a simple random sample of 16 trees. Use the table in Figure 9.4. Beginning in column 4 of row 135, use the first two digits in the row and read down the column. List the trees that you selected and shade the selected trees in the diagram. What percentage of trees did you select? How much will it cost to inspect the trees in the sample?

c. Select a 1-in-3 systematic sample. Use the fourth digit in column 3 of row 142 of the table in Figure 9.4 to begin the sample. List the trees that you selected and shade the selected trees in the diagram. What percentage of trees did you select? How much will it cost to inspect the trees in the sample?

d. Compare the selected trees in parts (a), (b), and (c). Which sampling method do you think is the most appropriate to use in this case? Explain.

36. The owner of a complex of 32 townhouses, which are numbered from 01 through 32 as shown in the following diagram, is willing to pay for 16 termite inspections. The eight townhouses in each block are connected.



Block 1 Block 2 Block 3 Block 4

a. Select a simple random sample of 16 of the townhouses. Use the second and third digits in column 5 of row 116 in the table in Figure 9.4, and read down the column. List the townhouses that you selected and shade the selected townhouses in the diagram.

b. If you assume that each block of eight townhouses is homogeneous, then the townhouses form four strata. Select a stratified random sample by randomly sampling 4 townhouses in each stratum. Use the table in Figure 9.4. For block 1, use the last two digits of column 1 in row 104 and read down the column. For block 2, use the last two digits of column 3 in row 101. For block 3, use the first two digits of column 2 in row 103. For block 4, use the first two digits of column 3 in row 102. List the townhouses that you selected and shade the selected townhouses in the diagram.

c. Select a cluster sample using blocks as clusters. Use the table in Figure 9.4 to select the sample. Begin in column 2 of row 110 and read across the row. List the townhouses that you selected and shade the selected townhouses in the diagram.

d. Compare the selected townhouses in parts (a), (b), and (c). Which sampling method do you think is the most appropriate to use in this case? Explain.

37. Suppose the Centers for Disease Control and Prevention (CDC) want to conduct a survey to determine health habits of children in a certain state. For each scenario below, identify the sampling technique described and discuss the pros and cons of using that sampling technique for the survey.

a. The CDC randomly selects several cities in the state and interviews all children in every school of each selected city.

b. The CDC lists all the schools in the state, randomly selects a certain number of schools from the list, and interviews every child in each of the selected schools.

c. The CDC lists every school-age child in the state, randomly selects a certain number of children from the list, and interviews each child selected.

d. The CDC divides the state into urban and rural areas, randomly selects a fixed number of schools from both areas, and interviews all the children in each selected school.

38. Suppose a fruit tree grower would like to assess the level of pest infestation in his crop of fruit trees. For each part below, identify the sampling technique described and discuss the pros and cons of using that sampling technique for the study.

a. The grower randomly selects a tree to examine and then moves row by row through the orchard, selecting every 15th tree for inspection.

b. For each tree, the grower flips a coin. If the coin lands heads, the grower inspects the tree.

c. The grower numbers each row of trees, randomly selects a certain number of rows, and inspects every tree in those rows.

d. Each tree is numbered, and a certain number of trees are randomly selected for inspection.

39. Advisors to the President are considering a new tax break for adults who have children. Before taking action, they would like to determine public opinion about the tax break, so they will take a sample of 350 adults. Describe a sampling technique that would be appropriate to use in this situation, and give reasons for your selection.

40. A snack factory makes potato chips, tortilla chips, and pretzels and packages them in lunch-size bags. Daily factory production is 35% potato chips, 40% tortilla chips, and 25% pretzels. A sample of size 100 will be selected to estimate how many bags are underweight. Describe a sampling technique that would be appropriate to use in this situation, and give reasons for your selection.

41. Suppose a new warning label is being considered for placement on packages of cigarettes. The Surgeon General calls for a survey to be conducted to determine the effectiveness of the warning label. A sample of size 500 will be taken. Describe a sampling technique that would be appropriate to use in this situation, and give reasons for your selection.

42. A university would like to survey recent graduates to determine average salaries. In the past year, 1700 students graduated with bachelor's degrees, 650 with master's degrees, and 45 with doctor's degrees. The university wants to sample 200 graduates. Describe a sampling technique that would be appropriate to use in this situation, and give reasons for your selection.

Extended Problems

43. On August 24, 1992, Hurricane Andrew struck the south Florida coast. At the time, it was the most expensive natural disaster in history. The Epidemiology Program at the Florida Department of Health and Rehabilitative Services in Miami conducted modified cluster samples in South Florida to obtain a population-based needs assessment after the hurricane. Research the methods used after Hurricane Andrew. How were the samples taken? What information was gained? What conclusions were drawn about the sampling methods used? For information on the Internet, search keywords "Hurricane Andrew modified cluster sampling." Write a report summarizing your findings.

44. Consider the job of an archaeologist who must decide how best to excavate a site that is a 100-meter-by-100-meter piece of farmland. Suppose, too, that several small artifacts turned up recently when a farmer plowed the land and that the artifacts seemed to be randomly distributed. It would be cost-prohibitive, disruptive, and a waste of time to excavate the entire site if no artifacts were found in many areas. Therefore, the archaeologist must use sampling to determine which areas of the site to excavate.

- **a.** Suppose that the archaeologist has a grant to excavate only 40 areas, each measuring 5 meters by 5 meters. Discuss the advantages and disadvantages of systematic random sampling. Explain how the 40 areas might be selected, and use your method to choose the sample. Sketch a diagram of the whole site and its forty 5-meter-by-5-meter areas that would be included in your sample.
- **b.** Suppose that the archaeologist has the same grant described in part (a). Discuss the advantages and disadvantages of simple random sampling. Explain how the 40 areas might be selected, and use your method to choose the sample. Sketch a diagram of the whole site and the forty 5-meter-by-5-meter regions that would be included in your sample.
- **c.** Research archaeological excavation methods. Write a report that summarizes the conditions under which systematic random sampling is used and those under which simple random sampling is used to determine excavation areas. On the Internet use search keywords "archaeological excavation patterns and sampling."

45. Farmers need to know which pests will harm a crop, and they must take action at the appropriate time. They routinely use sampling techniques to monitor pest infestations. However, each pest has a unique life cycle, feeding habits, invasion patterns, and activity levels. Understanding the specific pests that could harm a particular crop helps to determine the type of sampling that should be done when monitoring for infestations.

- **a.** Turf grasses and field crops can be damaged by chinch bugs. Research the chinch bug. How does it cause damage? At what point in its life cycle does it cause the most damage? Does it have any unique invasion patterns? Based on your research, select a sampling technique that a farmer could use to monitor for a chinch bug infestation and justify your choice. Summarize your findings and recommendations in a short report.
- **b.** Alfalfa, fruits, and vegetables can all be damaged by lygus bugs. Research the lygus bug. How does it cause damage? At what point in its life cycle does it cause the most damage? Does it have any unique invasion patterns? Based on your research, select a sampling technique that a farmer could use to monitor for a lygus bug infestation and justify your choice. Summarize your findings and recommendations in a short report.

46. Sometimes the purpose of sampling is not to determine a certain characteristic of individuals in a population such as their opinion on an issue, but rather to determine the total number of individuals in the population. For a situation in which it is impossible to count the entire population, a **capture–recapture** technique may be used. This method, also known as a **Lincoln index**, was developed for use in wildlife biology to monitor populations of birds, animals, fish, and insects. It is carried out by first taking a sample of size s_1 of a particular type of animal and marking the selected animals. The marked animals are then released and allowed to fully mix with the rest of the population. At a later time, a second sample of size s_2 is taken, and the number m of marked animals in this sample is noted. The total number of animals in the population can be estimated by assuming that the number of marked animals is proportional to the total number of animals in each case. In other words,

$$\frac{\text{animals marked by researcher}}{\text{total population of animals}} = \frac{\text{marked animals in sample}}{\text{total number of animals in sample}},$$ or

$$\frac{s_1}{N} = \frac{m}{s_2}.$$

Cross multiplying, we get

$$Nm = s_1 s_2.$$

Solving for N, we get

$$N = \frac{s_1 s_2}{m}.$$

For example, suppose a biologist captures 75 rats, tags them, and releases them back into the population. In this case, $s_1 = 75$. After a suitable amount of time, the biologist captures 120 rats and finds 15 have tags. Therefore, $s_2 = 120$ and $m = 15$. The population total is then estimated to be

$$N \approx \frac{(75)(120)}{15} = \frac{9000}{15} = 600 \text{ rats}.$$

a. In order to get an accurate estimate of the true population total using the Lincoln index, we make several assumptions:

- animals may not enter or leave the area,
- animals are equally likely to be captured in any sample,
- capture and marking techniques do not affect recapture, and
- markers will not be lost.

Discuss how reasonable you think each of these assumptions is as you consider two different populations: fish in a pond and deer in a forest.

b. Conduct a simulation of the capture–recapture technique on a population for which the total is known. Obtain a jar of beans and count them to find the total number N in the bean population. Mix the beans, take a random sample of size s_1, and mark each bean in your sample. Put the marked beans back in the jar and mix thoroughly. Take a second random sample of size s_2 and count the number m of marked beans in the sample. Use these values of the variables and the formula $N \approx \frac{s_1 s_2}{m}$ to estimate the total number of beans in the population. How close was your estimate to the true number N of beans in the population? Repeat the experiment five times. For each experiment, calculate $\frac{\text{Actual Total} - \text{Estimated Total}}{\text{Actual Total}} \times 100\%$ and explain what information these values provide.

c. Consider the experiment in part (b) and the assumptions listed in part (a). Conduct the simulation again, but this time after the marked beans have been placed back in the jar, add several new unmarked beans to the jar or else randomly remove several beans from the population. Then carry out the remainder of the experiment. Adding beans to the jar or subtracting beans from the jar simulates animals entering or leaving the area. How does your estimate of a total population change if you cannot assume that animals may not enter or leave the area? Write a summary about how adding or removing beans affected the simulation of the capture–recapture technique.

d. Research the capture–recapture technique and list 3 populations for which this method is used to estimate population totals. For population listed, discuss whether the assumptions from part (a) are reasonable. For information on the Internet, use search keywords "capture–recapture technique."

9.3 Measures of Central Tendency and Variability

INITIAL PROBLEM

Suppose you have a choice of two stockbrokers. Each will build a portfolio of stocks for you from his or her recommended lists. Over the past year, the percentage gains of the first stockbroker's recommendations were 21%, −3%, 16%, 27%, 9%, 11%, 13%, 6%, and 17%. The second stockbroker's recommendations had percentage gains of 11%, 13%, 16%, 8%, 5%, 14%, 15%, 17%, and 18%. Your goal is to minimize your risk while maintaining a steady rate of growth. Which stockbroker should you choose?

A solution of this Initial Problem is on page 612.

In Chapter 8, we considered various ways of displaying data visually. In this section, we look at ways to use a few numerical values to describe the data, especially the center of the data, and to describe the variability or spread of the data.

MEASURES OF CENTRAL TENDENCY

Statistics that tell us about the location of the values in a data set are called **measures of location**. The most important measures of location, called **measures of central tendency** or **measures of center**, give us information about where the center of the data lies. The most important measures of central tendency are the mean, the median, and the mode.

The Mean

Tidbit

Averages are useful for comparison. Consider average rainfall, for example. The wettest place in the world by far is Mawsynram, in Meghalaya State, India, with an average rainfall of 467.5 inches per year. The driest place is on the Pacific coast of Chile between Arica and Antofagasta, with an average rainfall of less than 0.004 inch per year.

Dictionaries give many definitions of the word "average" in describing data, but in this textbook (and in any scientific or technical context), "average" means the mean, which we define below.

Definition

THE MEAN

If the N numbers in a data set are denoted by $x_1, x_2, \ldots, x_N$, the **mean** of the data set is

$$\frac{x_1 + x_2 + \cdots + x_N}{N}.$$

For example, if our data set is 1, 2, 3, 4, 5, 6, then the mean of the data set is

$$\frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5.$$

Tidbit

The U.S. Bureau of Labor and Statistics calculates the Consumer Price Index (CPI) monthly. This index provides a measure of the mean change in prices paid by urban consumers for a market basket of goods and services. The CPI currently uses prices in the years 1982 to 1984 as a reference base.

Mathematicians sometimes use other kinds of means such as the geometric mean or the harmonic mean, so to prevent confusion, the mean we have defined is referred to as the **arithmetic mean**.

EXAMPLE 9.13 Find the mean of each of the following data sets.

a. 1, 1, 2, 2, 3 **b.** 1, 1, 2, 2, 11 **c.** 1, 1, 2, 2, 47

SOLUTION

a. The mean is $\frac{1 + 1 + 2 + 2 + 3}{5} = \frac{9}{5} = 1\frac{4}{5}$.

b. The mean is $\frac{1 + 1 + 2 + 2 + 11}{5} = \frac{17}{5} = 3\frac{2}{5}$.

c. The mean is $\frac{1 + 1 + 2 + 2 + 47}{5} = \frac{53}{5} = 10\frac{3}{5}$. ■

Notice that the data sets in Example 9.13 are identical except for the largest value in each set, yet their means are different. Thus, even a single data point can have a significant effect on the mean if it is much larger than (or much smaller than) the rest of the data points. Why might this matter? The following example suggests one possible reason.

EXAMPLE 9.14 A recent college graduate is investigating employment possibilities with a small company that has five employees. Literature about the company states that the mean salary at the company is $48,000. How might this statement be misleading?

SOLUTION The mean salary for this company is the sum of all five salaries divided by five. However, this number does not reveal what the salaries are for each employee. For example, the five employees might each earn $48,000 per year, in which case the mean salary would be given by

$$\frac{48{,}000 + 48{,}000 + 48{,}000 + 48{,}000 + 48{,}000}{5} = \frac{240{,}000}{5} = 48{,}000.$$

On the other hand, it could also be the case that the owner of the company makes $120,000 a year, and each of the four other employees earns $30,000 per year. In this case the mean salary would be given by

$$\frac{30{,}000 + 30{,}000 + 30{,}000 + 30{,}000 + 120{,}000}{5} = \frac{240{,}000}{5} = 48{,}000.$$

Notice that the means are the same in each case, but the two scenarios present very different prospects for the job seeker. ■

Sample Mean versus Population Mean

The set of data points used to calculate a mean might consist of values taken from the entire population or might consist of a smaller set of values taken from a sample of the population. Measurements from a sample are often used to make inferences about the whole population; often, the mean of a sample is used to estimate the mean of the population. For example, to find the mean height of 50-year-old men in the United States, we would not first find the sum of the heights of *all* 50-year-old males in the United States. Instead, we would probably take a sample of 50-year-old males and add their heights together. Then we would calculate the mean height of our sample by dividing by the number of men in the sample. If the sample is representative of the population, then the mean of the sample is a good approximation of the mean height of *all* 50-year-old males in the United States.

The mean of a sample is often used to approximate the mean of a population, so different notations are used for these two types of means.

Notation

SAMPLE MEAN AND POPULATION MEAN

The mean of a sample is denoted by $\bar{x}$ (read "x-bar"). The mean of a population is denoted by μ (this Greek letter is pronounced "mew").

The Median

A second measure of central tendency is the median. Informally, the median is known as the middle number of a data set.

Definition

THE MEDIAN

To find the **median**, arrange the data points of a data set in order from smallest to largest.

1. If the number of data points is *odd*, the data point in the middle of the list is the median of the data set.
2. If the number of data points is *even*, the mean of the two data points in the middle is the median of the data set.

Consider the data set 4, 5, 5, 6, 6. The data are arranged in order, and the middle value is 5. Thus, the median is 5. To find the median of the data set 6, 5, 10, 6, 5, 4, we must first arrange the data points in order. Arranging the data points from smallest to largest, we have 4, 5, 5, 6, 6, 10. Because the number of data points is even, there is no middle value. The median is the mean of the *two* middle data points 5 and 6; that is, the median is $\frac{5+6}{2} = \frac{11}{2} = 5.5$.

EXAMPLE 9.15 Find the median and the mean of each of the following data sets. Note that the data points in each set are arranged in order.

a. 0, 2, 4 **b.** 0, 2, 4, 10 **c.** 0, 2, 4, 10, 1000

SOLUTION

a. The median of the data set is the middle data point, or 2.

The mean is $\frac{0+2+4}{3} = \frac{6}{3} = 2$.

b. Since there is an even number of data points, the median is the mean of the two data points nearest the middle. The median is $\frac{2+4}{2} = 3$.

The mean of the data set is $\frac{0+2+4+10}{4} = \frac{16}{4} = 4$.

c. The median of this odd number of data points is the middle value, or 4.

The mean is $\frac{0+2+4+10+1000}{5} = \frac{1016}{5} = 203.2$. ■

Tidbit

In March 2004, the Internal Revenue Service reported 2.2 million delinquent 2003 tax accounts, which, if uncollected, would cost the United States a total of $14.1 billion in individual income taxes and $2.3 billion in corporate taxes. Due to insufficient staffing, the IRS announced it was unable to take the steps necessary to collect these accounts. The median size of a delinquent 2003 tax account was $14,000, and the largest delinquent account was $50 million. Had the mean been used rather than the median, the "typical" delinquent 2003 tax account reported by the IRS would have been much larger than $14,000.

Recall that one very large or one very small data point can change the mean dramatically. Part (c) of Example 9.15 again shows how one large number in a data set may have a large effect on the mean, but has very little effect on the median. Next let's take another look at the salary example discussed earlier.

EXAMPLE 9.16 Determine the median salary at the five-employee company, given the following groups of salaries.

a. $48,000, $48,000, $48,000, $48,000, $48,000

b. $30,000, $30,000, $30,000, $30,000, $120,000

SOLUTION In each case, the salaries are listed in order, and there is an odd number of salaries. So the median will be the middle number.

a. In this case the median is $48,000, which is exactly the same value as the mean we calculated earlier.

b. In this case the median is $30,000, much smaller than the mean of $48,000 that we calculated for this same set of data. ■

In this situation, the median gives a more representative picture of salaries at the small company. If the job seeker knew the median salary of employees was $30,000, she would know that at least half of the employees at the company make salaries less than or equal to $30,000, and at least half of the employees make salaries greater than or equal to $30,000. Because it is not affected by one very large or very small salary, the median is often used instead of the mean to indicate typical incomes, typical home prices, etc.

In the previous two examples, we found the values of the mean and the median for several sets of data. The relationship between the mean and the median can sometimes be predicted by examining a visual representation of a data set. In Chapter 8, we saw that bar graphs and histograms could illustrate sets of data. The rough shape of a graph of a data set often reveals something about the relationship between the mean and the median.

The bar graph shown in Figure 9.13 depicts the number of hours students spent studying for a physics exam during the week prior to the test.

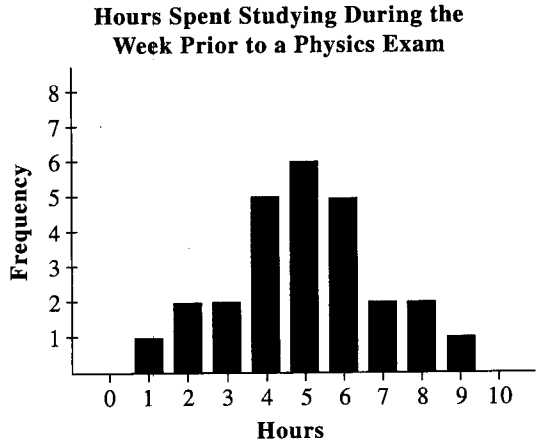


Figure 9.13

Notice that the graph is symmetric. It peaks in the middle and tapers off evenly at each end. The data distribution is said to be **symmetric** because the bar graph is symmetric.

We can use the given bar graph to determine the mean and the median for the data set. Using the frequencies represented by the heights of the bars, we know that the number of students for whom data is provided is $1 + 2 + 2 + 5 + 6 + 5 + 2 + 2 + 1 = 26$. Since the number of data points is even, the median will be the mean of the two middle numbers. Again, using the frequencies, we could make a list of the number of hours students spent studying:

$$1, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8, 9.$$

The median will be the mean of the two middle numbers, or $\frac{5 + 5}{2} = 5$. The mean of the data is

$$\frac{1 + 2 + 2 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 5 + 5 + 5 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 6 + 7 + 7 + 8 + 8 + 9}{26} = \frac{130}{26} = 5.$$

We could simplify the calculation of the mean by using the frequencies, as follows:

$$\frac{1(1) + 2(2) + 2(3) + 5(4) + 6(5) + 5(6) + 2(7) + 2(8) + 1(9)}{26} = \frac{130}{26} = 5.$$

Notice that in this case, the mean is the same as the median. In the case of a symmetric data distribution, the mean is equal to the median.

In contrast, Figure 9.14 shows an asymmetric bar graph.

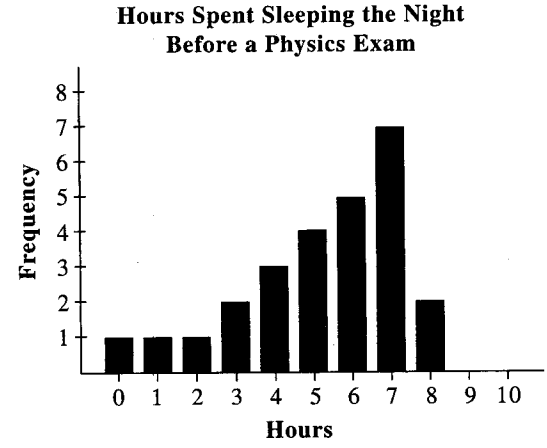


Figure 9.14

Notice that the data are more heavily concentrated on the right end of the graph, indicating that more people slept for 6 to 8 hours than slept for 1 to 3 hours the night before the exam.

Again, using the frequencies obtained from the bar graph, we can find the mean and median of this set of data. The number of students is $1 + 1 + 1 + 2 + 3 + 4 + 5 + 7 + 2 = 26$. Because there are 26 students, the median is the mean of the 13th and 14th data points. To find these two values, we could list the data points as we did for the other graph, or we could add frequencies from the left side of the graph until we get to 13 and 14. The 13th data point is 6 and the 14th data point is also 6. Thus, the median is $\frac{6 + 6}{2} = 6$. The mean of the data is

$$\frac{1(0) + 1(1) + 1(2) + 2(3) + 3(4) + 4(5) + 5(6) + 7(7) + 2(8)}{26} = \frac{136}{26} \approx 5.23$$

Notice that in this case the mean is less than the median. The small data points (such as 0 hours of sleep and 1 hour of sleep) located far from the rest of the data affected the mean. The graph of this distribution has shorter bars on the left side and taller bars on the right side, so we say the graph is skewed left. When the graph of an asymmetric distribution has taller bars on the left and shorter ones on the right, we say the graph is skewed right. These terms can also be defined using the mean and median of a distribution.

Definition

SKEWED DISTRIBUTIONS

A distribution is **skewed left** if the mean is less than the median.
A distribution is **skewed right** if the mean is greater than the median.

Two asymmetric bar graphs are shown in Figure 9.15. Notice that the distribution that is skewed left, with the taller bars on the right, has a mean of 5.68, which is less than the median value of 6. The distribution that is skewed right, with the taller bars on the left, has a mean that is greater than the median.

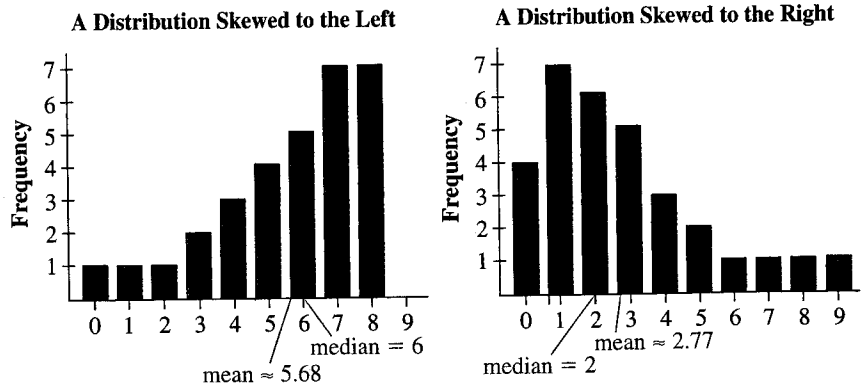


Figure 9.15

If a data distribution is skewed left, at least one small value must significantly decrease the mean. If a data distribution is skewed right, then at least one large data point must inflate the mean.

The Mode

The mode, another measure of central tendency in a data set, is the most commonly occurring value in a data set.

Definition

THE MODE

In a data set, the number that occurs most frequently is called the **mode**. A data set can have more than one mode if more than one number occurs most frequently. If every number in a data set appears equally often, then we say the distribution has no mode.

Table 9.4 lists some demographic data on the 50 states. If you look carefully through the list of percentages for Non-farm Manufacturing Employment, you will notice that 12.8% occurs three times, while no other percentage occurs more than twice in that column. Therefore, the mode for the percentage of Non-farm Manufacturing Employment in the United States in 2002 was 12.8%. More states had 12.8% of the workers employed in non-farm manufacturing jobs than any other percentage.

If you look through the list of percentages for Resident Population under 18, you will notice that 23.2%, 23.8%, 24.9%, and 26.5% all occur three times, while no other percentage occurs more than twice. There are four modes for this data. For the list of Per Capita Personal Income, no values are repeated, and thus there is no mode for this set of data.

This discussion about Table 9.4 illustrates two characteristics of the mode. The mode is relatively easy to determine because it requires no arithmetic, and it is most useful for a large set of data that can take on only a relatively small set of possible values. It is not unusual for a data set to have two modes. When a data set has more than two modes, however, the usefulness of the mode may be limited.

EXAMPLE 9.17 Find the mode(s) of the following Economics 101 test scores that were examined in Section 8.1.

26, 32, 54, 62, 67, 70, 71, 71, 74, 76, 80, 81, 84, 87, 87, 87, 89, 93, 95, 96

SOLUTION The score of 87 occurs three times and the score of 71 occurs twice. All the other scores occur only once. Thus, the mode for this data set is 87. ■

The dot plot in Figure 9.16 illustrates that the test score that occurs three times in Example 9.17 is the mode.

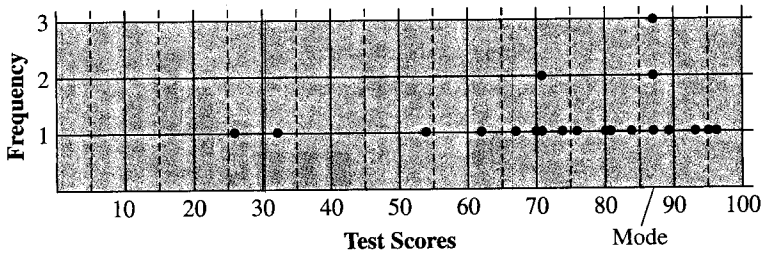


Figure 9.16

It is also easy to determine the mode when looking at a bar graph or histogram. Just look for the tallest bar. For example, examine the bar graph (Figure 9.14) that represents the number of hours students slept the night before the physics exam. Based on the frequencies, or heights of the bars, we can see that the mode is 7 hours of sleep. More students got 7 hours of sleep than any other number.

Table 9.4

DATA ON THE STATES

	2002 (July) Population Total	2002 Per Capita Personal Income (In 1996 Dollars)	2002 Resident Population Under 18 (Percent)	2002 Resident Population 65 and Older (Percent)	2002 Non-farm Manufacturing Employment (Percent)	2002 Unemployment Rate (Percent)
United States	287,941,220	27,857	25.3	12.3	11.7	5.8
Alabama	4,481,078	22,624	24.7	13.1	16.3	5.9
Alaska	640,841	28,947	29.9	6.1	3.8	7.7
Arizona	5,439,091	23,573	27.1	12.9	8.1	6.2
Arkansas	2,707,509	21,169	25.0	13.9	18.7	5.4
California	34,988,261	29,707	26.9	10.6	11.3	6.7
Colorado	4,498,077	29,959	25.5	9.6	7.6	5.7
Connecticut	3,459,006	38,450	25.2	13.6	12.8	4.3
Delaware	806,105	29,512	23.5	13.1	8.9	4.2
Florida	16,681,144	26,646	23.2	17.1	5.7	5.5
Georgia	8,539,735	25,949	26.5	9.5	12.1	5.1
Hawaii	1,234,514	27,011	23.7	13.4	2.7	4.2
Idaho	1,343,194	22,560	27.6	11.3	11.4	5.8
Illinois	12,585,204	30,075	25.8	11.9	12.8	6.5
Indiana	6,158,327	25,425	25.9	12.3	20.4	5.1
Iowa	2,934,776	25,461	23.8	14.7	15.7	4.0
Kansas	2,712,896	26,237	25.6	13.1	13.7	5.1
Kentucky	4,089,985	23,030	22.8	12.4	15.4	5.6
Louisiana	4,477,042	22,910	26.5	11.6	8.5	6.1
Maine	1,297,750	24,979	21.6	14.4	11.2	4.4
Maryland	5,441,531	32,680	25.3	11.3	6.4	4.4
Massachusetts	6,412,554	35,333	22.8	13.4	10.7	5.3
Michigan	10,042,495	27,276	25.6	12.3	17.0	6.2
Minnesota	5,025,081	30,675	24.9	12.0	13.4	4.4
Mississippi	2,867,635	20,142	26.5	12.1	16.7	6.8
Missouri	5,679,770	26,052	24.6	13.3	12.0	5.5
Montana	910,670	22,526	23.8	13.5	5.0	4.6
Nebraska	1,726,437	26,804	25.4	13.4	11.7	3.6
Nevada	2,168,304	27,172	26.3	11.1	4.1	5.5
New Hampshire	1,275,607	30,912	24.2	12.0	13.8	4.7
New Jersey	8,577,250	35,521	24.8	13.1	9.2	5.8
New Mexico	1,855,143	21,555	27.0	11.9	5.0	5.4
New York	19,151,066	32,451	24.1	12.9	7.7	6.1
North Carolina	8,311,899	24,949	24.9	12.0	16.7	6.7
North Dakota	633,799	24,293	23.2	14.8	7.2	4.0
Ohio	11,410,396	26,474	25.2	13.3	16.3	5.7
Oklahoma	3,488,201	23,026	25.0	13.2	10.3	4.5
Oregon	3,523,281	25,867	24.3	12.6	12.8	7.5
Pennsylvania	12,328,459	28,565	23.2	15.5	13.5	5.7
Rhode Island	1,068,897	28,198	22.4	14.2	13.0	5.1
South Carolina	4,105,848	22,868	23.8	12.3	16.1	6.0
South Dakota	760,452	24,214	25.7	14.2	10.2	3.1
Tennessee	5,792,297	24,913	24.2	12.4	16.0	5.1
Texas	21,723,220	25,705	28.0	9.9	10.1	6.3
Utah	2,319,743	21,883	30.8	8.6	10.6	6.1
Vermont	616,500	26,620	22.7	12.9	13.5	3.7
Virginia	7,273,572	29,641	24.4	11.2	9.2	4.1
Washington	6,067,146	29,420	24.9	11.2	10.8	7.3
West Virginia	1,805,230	21,327	21.6	15.3	9.4	6.1
Wisconsin	5,440,367	26,941	24.6	13.0	19.0	5.5
Wyoming	499,192	27,530	24.5	11.9	3.8	4.2

Tidbit

One unusual college grades on the basis of 5 points for an A, 4 points for a B, 3 points for a C+, 2 points for a C− (there is no plain C and no other + or − grades), 1 point for a D, and no points for an F. The same college also gives one credit for every course regardless of the number of lecture or lab hours required.

The Weighted Mean

For some data sets, different data points have different levels of importance, and these differences are often taken into account when calculating an average. An example probably familiar to you is grade point average (GPA). In most colleges and universities, each course has a given number of credits. The numerical value of the grade you receive in the course is weighted by the number of credits for the course when your GPA is calculated. The usual assignment of point values to letter grades is 4 for an A, 3 for a B, 2 for a C, 1 for D, and 0 for an F although many colleges and universities extend this system by using + and − grades as well.

EXAMPLE 9.18 Suppose that your grades for one semester consisted of an A in a five-credit course, a B in a four-credit course, and a C in two different three-credit courses. What would your GPA be for that semester?

SOLUTION As you may know, to calculate your GPA, we first multiply the numerical value of each grade by the number of credits for that course and then add these products:

$$4(5) + 3(4) + 2(3) + 2(3).$$

This step takes into account the fact that the five-credit course is worth more than the three-credit courses. Next we divide this sum by the total number of credits to get your GPA, as shown.

$$\text{GPA} = \frac{4(5) + 3(4) + 2(3) + 2(3)}{5 + 4 + 3 + 3} \approx 2.93$$

Thus, your GPA for the semester would be approximately 2.93. ■

This kind of mean, in which some data points are worth more than other data points, is called a weighted mean. In the preceding example, each numerical grade is a data point, and the number of credits is the **weight** associated with that grade. The box below summarizes the method for calculating a weighted mean.

Definition

THE WEIGHTED MEAN

If the numbers in a data set are $x_1, x_2, \ldots, x_N$ and these numbers have weights of $w_1, w_2, \ldots, w_N$, respectively, then the **weighted mean** of the data is

$$\frac{w_1x_1 + w_2x_2 + \cdots + w_Nx_N}{w_1 + w_2 + \cdots + w_N}.$$

Grade point averages are not the only applications of weighted means. Let's look at another example in which a weighted mean is appropriate.

EXAMPLE 9.19 Table 9.5 shows the approximate 2002 per-capita income and population for eight northwestern European countries: Belgium, France, Germany, Ireland, Luxembourg, the Netherlands, Switzerland, and the United Kingdom. Use the data in the table to determine the per-capita income for the entire group of nations.

Table 9.5

Nation	Per-Capita Income (in thousands of U.S. dollars)	Population (in millions)
Belgium	23.8	10.27
France	23.1	59.77
Germany	23.3	83.25
Ireland	22.6	3.88
Luxembourg	39.2	0.45
Netherlands	24.3	16.07
Switzerland	37.9	7.30
United Kingdom	24.8	59.78

Source: www.nationmaster.com.

SOLUTION The product of the per-capita income of a country and the population of the country gives the total income of the country. Thus, the overall per-capita income of the eight countries can be computed as a weighted mean of the per-capita incomes given in the table, using the populations of the countries as the weights. We obtain a weighted mean of

$$\frac{23.8(10.27) + 23.1(59.77) + 23.3(83.25) + 22.6(3.88) + 39.2(0.45) + 24.3(16.07) + 37.9(7.30) + 24.8(59.78)}{10.27 + 59.77 + 83.25 + 3.88 + 0.45 + 16.07 + 7.30 + 59.78}$$

$$= \frac{5819.881}{240.77} \approx 24.2.$$

The numerator in the preceding computation is the total income of the eight countries, and the denominator (the sum of the weights) is the total population of the eight countries. The weighted mean of 24.2 thousands, or \$24,200, is the approximate 2002 per-capita income in northwestern Europe. (By comparison, the 2002 per-capita income in the United States was about \$34,900.) ■

MEASURES OF VARIABILITY

The measures of central tendency that we have discussed so far describe only part of the behavior of a data set. We also need to know how the data set varies from its center. Statistics that tell us about how the data vary from the center are called **measures of variability** or **measures of spread**.

The Range

The measure of variability that is most easily calculated is the range, which is simply the difference between the largest number and the smallest number in the data set.

Definition

THE RANGE

If the numbers in a data set $x_1, x_2, \ldots, x_N$ are arranged in increasing order from smallest to largest, then the **range** of the data set is $x_N - x_1$.

EXAMPLE 9.20 Compute the mean and the range for each of the following data sets.

a. 3, 4, 5, 6, 7, 8 **b.** 0, 2, 5, 7, 8, 11

SOLUTION

a. The mean for the first set of data is $\frac{3 + 4 + 5 + 6 + 7 + 8}{6} = 5.5$

The range for the first set of data is $8 - 3 = 5$.

Note that the range is a single number, 5. We do not say the range is 3 to 8.

b. The mean for the second set of data is $\frac{0 + 2 + 5 + 7 + 8 + 11}{6} = 5.5$

The range for the second set of data is $11 - 0 = 11$. ■

Notice that the two sets of data have exactly the same mean. The fact that the range of the second set of data is larger than the range of the first data set means that the second data set is more spread out than the first. The values in the first set of data are more closely clustered around the mean. However, the range does not always give a full picture of the variability in a data set, as the next example shows.

EXAMPLE 9.21 Compute the range for each of the following data sets.

a. 0, 8, 9, 6, 1, 4, 6, 0, 1, 5, 3, 0, 9, 8, 0, 5, 6, 9, 5, 0

b. 0, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 7, 9

SOLUTION For both data sets, the largest number is 9 and the smallest number is 0, so the range in each case is $9 - 0 = 9$. ■

Although the two data sets in Example 9.21 both have the same range, the bar graphs of the two data sets as pictured in Figure 9.17 show that the data set in (b) is more closely concentrated near the center than is the data set in (a). The graphs show that although the range is an easily computed measure of variability, it may not be a good indicator of the spread of the data. The range is not as sensitive as the measures of variability we will consider in the remainder of this section.

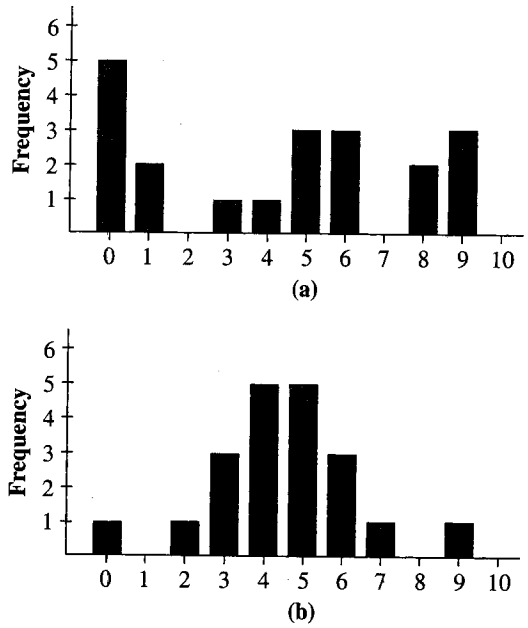


Figure 9.17

Quartiles

A second way to measure the variability of a data set is to use quartiles (you may have encountered quartiles in the reporting of your test scores). Suppose there is an odd number of data points, say 15. In this case, the first quartile q_1, is the median of the 7 smallest numbers (those to the *left* of the median if the numbers are arranged in increasing order). The third quartile q_3 is the median of the 7 largest numbers (those to the *right* of the median if the numbers are arranged in increasing order). If there is an even number of data points, say 16, then the first quartile is the median of the smallest 8 numbers and the third quartile is the median of the largest 8 numbers. (Again, those 8 numbers are the numbers to the left and right, respectively, of the median of the data set when the numbers are arranged in increasing order.)

More formally, the way we define the first quartile, q_1, depends on whether we have an even or odd number of points. For an odd number of data points, say $2N + 1$, the first quartile is the median of the N smallest numbers. With an even number of data points, say $2N$, the first quartile is the median of the N smallest numbers. The third quartile, q_3, is defined like the first quartile, except it is the median of the N largest numbers. The interquartile range (IQR) is $q_3 - q_1$, which is a measure of the spread in the data. The quartiles themselves are measures of location of the data. The steps in the calculation of quartiles are summarized below.

Definition

QUARTILES AND THE INTERQUARTILE RANGE

To find the first, second, and third quartiles of a data set,

1. Arrange all the data in order from smallest to largest. Include any repeated value in the ordered list as often as it is repeated.
2. Determine the median of the data set. Let $q_2 = m =$ median of the entire data set.
3. If the number of data points is even, then go to step 4. If the number of data points is odd, then remove the middle data point from the ordered list from step 1 before going to step 4.
4. Divide the remaining data points into a lower and an upper half. The lower half consists of the first half of the remaining data points from the ordered list from step 1. The upper half consists of the second half of the remaining data points from the ordered list from step 1.
5. The **first quartile**, written q_1, is the median of the lower half of the data.
6. The **third quartile**, written q_3, is the median of the upper half of the data.
7. The **interquartile range**, written IQR, is the difference $q_3 - q_1$.

Now let's reconsider some examples we looked at earlier. Recall that we determined that the median of the set 4, 5, 5, 6, 6 was 5. Thus, we know $q_2 = 5$. The lower half of the data, not including the middle data point, is 4, 5, which has a median of 4.5. This is the first quartile, so $q_1 = 4.5$. The upper half of the data, not including the middle data point, is 6, 6, whose median is 6. This is the third quartile, so $q_3 = 6$. Thus, the interquartile range is IQR $= 6 - 4.5 = 1.5$.

For the data set 4, 5, 5, 6, 6, 10 we found that the median was 5.5. The lower half of the data set is 4, 5, 5, which has a median of 5, and the upper half of the data set is 6, 6, 10, which has a median of 6. Thus, $q_1 = 5$, $m = 5.5$, $q_3 = 6$, and IQR $= 6 - 5 = 1$. For large sets of data, it is usually true that approximately 25% of the data is between each of the quartile locations.

EXAMPLE 9.22 Consider the economics class test results from Section 8.1. Recall that the scores arranged in order were as follows:

26, 32, 54, 62, 67, 70, 71, 71, 74, 76, 80, 81, 84, 87, 87, 87, 89, 93, 95, 96.

Find the median, the first and third quartiles, and the interquartile range for this set of test scores.

SOLUTION There are 20 numbers in this data set, so no data point lies in the middle. Counting from the left, we see that the 10th data point is 76 and that the 11th data point is 80, so the median is $m = \dfrac{76 + 80}{2} = 78$.

Next we find the first quartile, q_1. This is the median of the lower half of the data set, which is 26, 32, 54, 62, 67, 70, 71, 71, 74, 76. There is also an even number of data points in this list, so the median of the lower half of the data is $\dfrac{67 + 70}{2} = 68.5$. Thus, $q_1 = 68.5$. Likewise, the third quartile is the median of the upper half of the data set, so $q_3 = \dfrac{87 + 87}{2} = 87$. The interquartile range is $q_3 - q_1 = 87 - 68.5 = 18.5$. Notice that 11 data points, or about 50% of the data, lie in the interval from 68.5 to 87, inclusive. ■

Five-Number Summaries and Box-and-Whisker Plots

Suppose we have a data set and we know that s is the smallest data point, L is the largest data point, m is the median, and q_1 and q_3 are the first and third quartiles. These five numbers have a special designation.

Definition

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a data set is the list s, q_1, m, q_3, L, where

s = the smallest value in the data set
q_1 = the first quartile,
m = the median of the data set,
q_3 = the third quartile, and
L = the largest value in the data set.

For example, the five-number summary of the data from the economics class in Example 9.22 is the set 26, 68.5, 78, 87, 96. Recall that q_2 is another name for the median m. If we let $q_0 = s$ and $q_4 = L$, then the five-number summary could also be represented as q_0, q_1, q_2, q_3, q_4.

The numbers in the five-number summary of a data set may be graphically represented in a **box-and-whisker plot** (also called a **box plot**) to give a picture of the data while omitting the details. A box-and-whisker plot for the economics class of Example 9.22 is shown Figure 9.18.

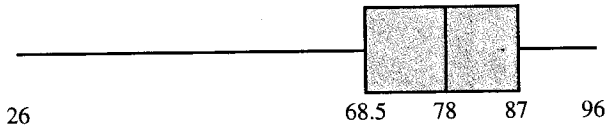


Figure 9.18

The box in the center portion of the graph shows the location of approximately 50% of the data that lies in the middle. In this case, they lie between the scores of 68.5 and 87, inclusive. The length of the box, $87 - 68.5 = 18.5$, is the IQR. The vertical line at 78 marks the median, or the middle of the data. The line that extends from 26 to 68.5 is one of the "whiskers." It indicates that 26 is the lowest score and that approximately 25% of the scores are between 26 and 68.5. The upper whisker shows that the highest score is 96 and that approximately 25% lie between 87 and 96. The fact that the upper whisker is much shorter than the lower whisker indicates that the approximately 25% of the data that is at the top of the scores are much closer together than the approximately 25% of the data that is at the bottom of the scores.

The relationship between the five-number summary s, q_1, m, q_3, L and the corresponding box-and-whisker plot is shown in Figure 9.19.

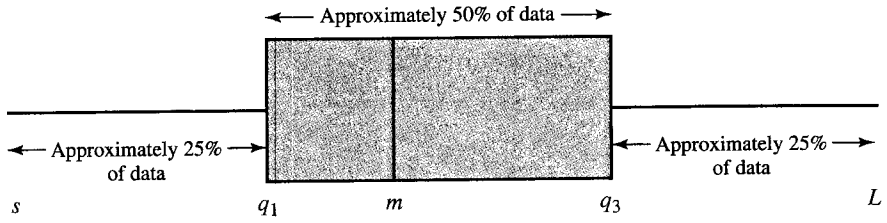


Figure 9.19

Notice that the first and third quartiles define the ends of the box, while the minimum and maximum values define the ends of the whiskers.

In this book, box-and-whisker plots are generally drawn horizontally, but you may also see them drawn vertically. Box-and-whisker plots give a quick and easy way to compare two data sets, as the next example illustrates.

EXAMPLE 9.23 Monthly rainfall data for two cities are given in Table 9.6.

Table 9.6

MONTHLY RAINFALL DATA (IN INCHES)												
	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
St. Louis, MO	2.21	2.31	3.26	3.74	4.12	4.10	3.29	2.96	3.20	2.64	2.64	2.23
Portland, OR	0.46	1.13	1.47	1.61	2.08	2.31	3.05	3.61	3.93	5.17	6.14	6.16

The first number is January's mean rainfall in St. Louis, the second is February's mean rainfall, and so on. Make box-and-whisker plots to compare the average rainfall in these communities. What conclusions can you draw?

SOLUTION We first arrange the rainfall amounts in order and then compute the five-number summary for each of the cities.

St. Louis, MO

Mean monthly rainfall (in inches): 2.21, 2.23, 2.31, 2.64, 2.64, 2.96, 3.20, 3.26, 3.29, 3.74, 4.10, 4.12

Median = 3.08 $q_1 = 2.475$ $q_3 = 3.515$

Five-number summary: 2.21, 2.475, 3.08, 3.515, 4.12

Interquartile range: $3.515 - 2.475 = 1.04$

Portland, OR

Mean monthly rainfall (in inches): 0.46, 1.13, 1.47, 1.61, 2.08, 2.31, 3.05, 3.61, 3.93, 5.17, 6.14, 6.16

Median = 2.68 $q_1 = 1.54$ $q_3 = 4.55$

Five-number summary: 0.46, 1.54, 2.68, 4.55, 6.16

Interquartile range: $4.55 - 1.54 = 3.01$

Now we can draw two box-and-whisker plots side by side for comparison (Figure 9.20).

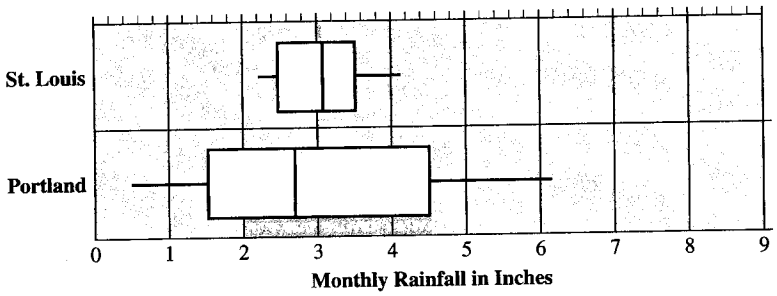


Figure 9.20

We can draw at least two conclusions from these two plots. Because the graph for the St. Louis data is much narrower than the Portland graph, we can see that the amount of rain varies less in St. Louis on a month-to-month basis. The wider spread of the plot for Portland indicates that Portland has some months that are drier than St. Louis's driest months and some that are far wetter than St. Louis's wettest months.

VARIANCE AND STANDARD DEVIATION

So far we have looked at two different ways to describe the variability of a data set—the range and quartiles. Next we look at another way to measure a data set's variability. Suppose that the organizer of a job training program is interested in the income level of its graduates 5 years after they have finished their training. The income data gathered in a small pilot study appears in Table 9.7.

Table 9.7

Annual Incomes		
$12,000	$29,400	$39,900
$25,600	$35,700	$41,900
$25,800	$36,100	$75,800

The set of incomes listed in Table 9.7 has a mean of $35,800. (Check this calculation.) Notice that none of the graduates whose incomes are listed actually has that mean income. In fact, many of the incomes listed seem to be spread rather far from the mean. Our goal here is to describe another quantitative measure of how data are spread out from the mean. This measure, called the standard deviation, has important applications in statistics.

The calculation of the standard deviation requires several steps. First, we must calculate the difference between each data point and the mean.

Definition

DEVIATION FROM THE MEAN

The difference between a data point x and the mean $\bar{x}$, namely $x - \bar{x}$, is called the **deviation from the mean** of data point.

EXAMPLE 9.24 Make a table showing the annual incomes from Table 9.7 together with the deviation from the mean of each data point.

SOLUTION We know the mean is \$35,800. We subtract this number from each of the incomes. The results are listed in Table 9.8.

Table 9.8

Data Point	Deviation from the Mean
\$12,000	12,000 − 35,800 = −\$23,800
25,600	25,600 − 35,800 = −10,200
25,800	25,800 − 35,800 = −10,000
29,400	29,400 − 35,800 = −6400
35,700	35,700 − 35,800 = −100
36,100	36,100 − 35,800 = 300
39,900	39,900 − 35,800 = 4100
41,900	41,900 − 35,800 = 6100
75,800	75,800 − 35,800 = 40,000

■

Notice that the deviations are negative when the income is less than the mean and positive when the income is greater than the mean. It might seem reasonable to find the average deviation from the mean. However, if we try to calculate the average deviation from the mean by first adding up all the deviations, we will find that the sum of all the deviations from the mean in Table 9.8 is zero (check this by adding all the numbers in the last column). This result is true for all data sets. Thus, summing the deviations from the mean will always give a value of zero no matter what the data set looks like. One way around this problem is to take the absolute value of the deviations from the mean. Although this strategy seems reasonable, the absolute value and related averages have no convenient algebraic properties that can be used for further analysis of the data. So, instead of summing the deviations from the mean, we sum the *squares* of the deviations from the mean. Then we can use the sum of the squares of the deviations from the mean to compute the sample variance as stated in the next definition.

Definition

SAMPLE VARIANCE

Given a sample of n measurements $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$, the **sample variance**, s^2, is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.$$

EXAMPLE 9.25 Compute the sample variance of the annual incomes given in Table 9.7.

SOLUTION The deviations from the mean have already been listed in Table 9.8. Now we add a column containing the squares of the deviations from the mean (Table 9.9).

Table 9.9

Data Point	Deviation from the Mean	(Deviation from the Mean)^2
\$12,000	−23,800	566,440,000
25,600	−10,200	104,040,000
25,800	−10,000	100,000,000
29,400	−6400	40,960,000
35,700	−100	10,000
36,100	300	90,000
39,900	4100	16,810,000
41,900	6100	37,210,000
75,800	40,000	1,600,000,000

The sample variance is found by dividing the sum of the squares of the deviations from the mean by $n - 1$. The sum of the numbers in the right hand column of Table 9.9 is 2,465,560,000, so the sample variance is $s^2 = \dfrac{2{,}465{,}560{,}000}{9 - 1} = 308{,}195{,}000.$ ■

The units of each of the squares of the deviations from the mean in Table 9.9, and on the sample variance, must be *square dollars* because we obtained these numbers by multiplying dollars times dollars. Obviously, none of us has ever seen a *square dollar* and we never will. But if we take the square root of the *square dollars* in the sample variance, then we will have the sensible units of dollars. The square root of the sample variance is called the sample standard deviation.

Definition

SAMPLE STANDARD DEVIATION

Given a sample of n measurements $x_1, x_1, x_2, \ldots, x_n$ with mean $\bar{x}$, the **sample standard deviation**, s, is

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}.$$

The sample standard deviation measures how the data are spread out, using the same units as the data. In Example 9.24, the sample standard deviation is $\sqrt{308{,}195{,}000} \approx$ \$17,555.

EXAMPLE 9.26 Shown next are two data sets. The first set of data gives the weights (in pounds) of five turkeys chosen from a flock of turkeys being sent to market, while the second set gives the weights (in pounds) of five dogs chosen from those at a dog show for small breeds. Find the sample mean, sample variance, and sample standard deviation of each of the sets of weights.

Turkeys: 17, 18, 19, 20, 21 Dogs: 13, 16, 19, 22, 25

SOLUTION The sample mean for the turkeys is $\bar{x} = \dfrac{17 + 18 + 19 + 20 + 21}{5} =$ 19 pounds; the sample mean for the dogs is $\bar{x} = \dfrac{13 + 16 + 19 + 22 + 25}{5} =$ 19 pounds. Although the sample means are equal, the turkey weights are bunched closer together than the dog weights, so the variance should reflect this difference. For the set of turkey weights, we compute the sample variance and the standard deviation by completing Table 9.10.

Table 9.10

Turkey Weights, in pounds	Deviation from the Mean	(Deviation from the Mean)^2
17	$17 - 19 = -2$	4
18	$18 - 19 = -1$	1
19	$19 - 19 = 0$	0
20	$20 - 19 = 1$	1
21	$21 - 19 = 2$	4

The sum of the 5 numbers in the right-hand column is 10, so the sample variance is $s^2 = \dfrac{10}{5 - 1} = \dfrac{10}{4} = 2.5$. The sample standard deviation is $s = \sqrt{2.5} \approx 1.58$ pounds.

For the set of dog weights, we compute the sample variance and the standard deviation by completing Table 9.11.

Table 9.11

Dog Weights, in pounds	Deviation from the Mean	(Deviation from the Mean)^2
13	$13 - 19 = -6$	36
16	$16 - 19 = 3$	9
19	$19 - 19 = 0$	0
22	$22 - 19 = 3$	9
25	$25 - 19 = 6$	36

The sum of the numbers in the right-hand column is 90, so the sample variance is $s^2 = \dfrac{90}{5 - 1} = \dfrac{90}{4} = 22.5$. The sample standard deviation is $s = \sqrt{22.5} \approx 4.74$ pounds. ■

Notice that in this example the sample variance and sample standard deviation are larger for the dog data than for the turkey data (4.74 pounds for the dogs and 1.58 pounds for the turkeys), which reflects the wider spread in the dog data. Also notice that for each of the data sets the sample standard deviation is measured in the same units (pounds) as the original data points.

If a data set represents the measurements of *all* the elements of a population, then you can compute the population variance and the population standard deviation using *all* the data points according to the following definition. We use N for the number of elements in the *population* to distinguish it from n, the number of elements in a *sample*.

Tidbit

The Greek letter μ (spelled "mu") is used to represent the population mean, and the letter σ (spelled "sigma," is used to represent the population standard deviation. These are only two examples of Greek letters commonly used in many areas of mathematics. For example, the letters α (spelled "alpha") and β (spelled "beta") are often used in geometry and trigonometry to represent angles of a triangle. The Greek letter Σ (uppercase sigma) is the standard symbol used to indicate a sum in statistics and calculus.

Definition

POPULATION VARIANCE AND POPULATION STANDARD DEVIATION

Suppose the set of all N measurements, $x_1, x_2, \ldots, x_N$, on a population of N elements is given. If the population mean is μ, then the **population variance**, σ^2, is

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

and the **population standard deviation**, σ, is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}}.$$

Notice that when computing the *sample* variance for n measurements, we divide the sum of the squares of the deviations from the mean by $n - 1$, whereas when computing the *population* variance for the N measurements from all the elements in the population, we divide by N. It might seem more reasonable to use the same form for the divisor in both cases. One reason we use $n - 1$ to calculate the sample variance is that we will be using the sample variance to estimate the population variance. Using n as the divisor in computing the sample variance tends to give an underestimate of the population variance.

Many scientific and graphing calculators have built-in statistical functions that can be used to find the mean, median, variance, standard deviation, and quartiles. Graphing calculators will even create box-and-whisker plots for given sets of data. Likewise, a spreadsheet can eliminate much of the tedious work involved in statistical calculations.

SOLUTION OF THE INITIAL PROBLEM

Suppose you have a choice of two stockbrokers. Each will build a portfolio of stocks for you from his or her recommended lists. Over the past year, the percentage gains of the first stockbroker's recommendations were 21%, −3%, 16%, 27%, 9%, 11%, 13%, 6%, and 17%. The second stockbroker's recommendations had percentage gains of 11%, 13%, 16%, 8%, 5%, 14%, 15%, 17%, and 18%. Your goal is to minimize your risk while maintaining a steady rate of growth. Which stockbroker should you choose?

SOLUTION We will first check the average rate of growth by computing the mean of each data set. The mean of the returns from the first portfolio is

$$\frac{21 + (-3) + 16 + 27 + 9 + 11 + 13 + 6 + 17}{9} = \frac{117}{9} = 13.$$

The mean of the returns from the second portfolio is

$$\frac{11 + 13 + 16 + 8 + 5 + 14 + 15 + 17 + 18}{9} = \frac{117}{9} = 13.$$

Notice that the average growth rates are the same. To minimize risk you might choose the broker whose choices have the least variability. To determine the variability, compute the standard deviation and interquartile range of each data set. The computations for the first data set are in Table 9.12.

Table 9.12

Data Point	Deviation from the Mean	(Deviation from the Mean)^2
21	$21 - 13 = 8$	64
-3	$-3 - 13 = -16$	256
16	$16 - 13 = 3$	9
27	$27 - 13 = 14$	196
9	$9 - 13 = -4$	16
11	$11 - 13 = -2$	4
13	$13 - 13 = 0$	0
6	$6 - 13 = -7$	49
17	$17 - 13 = 4$	16
	Total of Squared Deviations	610
$\text{Variance} = \frac{610}{9-1} = \frac{610}{8} = 76.25$		
$\text{Standard deviation} = \sqrt{76.25} \approx 8.73$		

The calculations for the second data set are in Table 9.13.

Table 9.13

Data Point	Deviation from the Mean	(Deviation from the Mean)^2
11	$11 - 13 = -2$	4
13	$13 - 13 = 0$	0
16	$16 - 13 = 3$	9
8	$8 - 13 = -5$	25
5	$5 - 13 = -8$	64
14	$14 - 13 = 1$	1
15	$15 - 13 = 2$	4
17	$17 - 13 = 4$	16
18	$18 - 13 = 5$	25
	Total of Squared Deviations	148
$\text{Variance} = \frac{148}{9-1} = \frac{148}{8} = 18.5$		
$\text{Standard deviation} = \sqrt{18.5} \approx 4.30$		

The standard deviation of the second portfolio is much smaller than the standard deviation of the first. To be thorough, we will also compare the interquartile ranges (IQRs). The percent returns in the first portfolio in order, are -3, 6, 9, 11, 13, 16, 17, 21, and 27. The first quartile is the median of -3, 6, 9, and 11, or $\frac{6+9}{2} = 7.5$. The third quartile is the median of 16, 17, 21, and 27, or $\frac{17+21}{2} = 19$. The IQR is $19 - 7.5 = 11.5$.

Similarly, the percent returns in the second portfolio, in order, are 5, 8, 11, 13, 14, 15, 16, 17, and 18. The first quartile is $\frac{8+11}{2} = 9.5$, and the third quartile is $\frac{16+17}{2} = 16.5$, so the IQR is $16.5 - 9.5 = 7$. The IQR for the second data set is smaller than the interquartile range of the first data set, again indicating that the first portfolio has greater variability. The second stockbroker should be chosen to minimize risk.

PROBLEM SET 9.3

1. The team roster for the NBA Boston Celtics basketball team (2003–2004 season) and the players' heights and weights are given in the following table.

Player	Height, in inches	Weight, in pounds
Marcus Banks	74	200
Mark Blount	84	250
Ricky Davis	79	195
Brandon Hunter	79	260
Mike James	74	188
Jumaine Jones	80	218
Raef LaFrentz	83	240
Walter McCarty	82	230
Chris Mihm	84	265
Chris Mills	79	220
Kendrick Perkins	82	280
Paul Pierce	78	230
Michael Stewart	82	230
Jiri Welsch	79	215

Source: www.espn.com.

a. Find the mean, median, and mode for the players' heights.

b. Find the mean, median, and mode for the players' weights.

c. Suppose Marcus Banks is cut from the team. Recalculate the mean, median, and mode for the height data. Which of these value(s) did not change? Explain why.

d. Suppose Kendrick Perkins is cut from the team. Recalculate the mean, median, and mode for the weight data. Which of these values changed the most? Explain why.

2. The team roster for the WNBA Los Angeles Sparks basketball team (2003–2004 season) and the players' heights and weights are given in the following table.

Player	Height, in inches	Weight, in pounds
Tamecka Dixon	69	148
Isabelle Fijalkowski	77	200
Jennifer Gillom	75	180
Chandra Johnson	75	185
Lisa Leslie	77	170
Mwadi Mabika	71	165
DeLisha Milton-Jones	73	172
Vanessa Nygaard	73	175
Lynn Pride	74	180
Nikki Teasley	72	169
Teresa Weatherspoon	68	161
Shaquala Williams	66	135
Sophia Witherspoon	70	145

Source: www.wnba.com.

a. Find the mean, median, and mode for the players' heights.

b. Find the mean, median, and mode for the players' weights.

c. Suppose a new 77-inch-tall player joins the team. Recalculate the mean, median, and mode for the height data. Did any of these three values stay the same? Explain why or why not.

d. Suppose Shaquala Williams is cut from the team. Recalculate the mean, median, and mode for the weight data. Which of these values changed the most? Explain why.

3. The following table contains homicide rates per 100,000 for the years from 1980 through 1999 in the United States.

Year	Homicide Rate per 100,000	Year	Homicide Rate per 100,000
1980	10.2	1990	9.4
1981	9.8	1991	9.8
1982	9.1	1992	9.3
1983	8.3	1993	9.5
1984	7.9	1994	9.0
1985	7.9	1995	8.2
1986	8.6	1996	7.4
1987	8.3	1997	6.8
1988	8.4	1998	6.3
1989	8.7	1999	5.7

Source: U.S. Bureau of Justice Statistics.

a. Find the mean and median of the homicide rates for the years 1980 through 1989.

b. Find the mean and median of the homicide rates for the years 1990 through 1999.

c. If you were a member of a political group and wanted to emphasize the idea that crime prevention policies were working, which would you report, the mean or the median, for each period in parts (a) and (b) and why?

4. The following table contains salaries for all players on two NBA basketball teams for the 2003–2004 season.

Annual Salary, Memphis Grizzlies	Annual Salary, Minnesota Timberwolves
$6,834,444	$27,995,000
$6,600,000	$13,475,000
$6,187,500	$8,000,000
$6,187,500	$5,250,000
$5,955,000	$4,917,000
$4,917,000	$4,498,514
$4,592,418	$2,475,000
$3,416,520	$1,500,000
$3,380,457	$938,679
$2,533,440	$938,679
$1,580,702	$709,224
$1,325,000	$698,800
$1,285,440	$638,679
$1,143,360	$184,355
$1,063,680	$134,541
$366,931	$114,727
	$83,464

Source: www.hoopshype.com/salaries.

a. Find the mean and median of the salaries for the Memphis Grizzlies.

b. Find the mean and median of the salaries for the Minnesota Timberwolves.

c. Which measure of central tendency is a better representation of a typical salary in each case? Explain your reasoning.

5. Students in a class took a 10-point quiz. The class received one grade of 5, three grades of 6, eight grades of 7, six grades of 8, five grades of 9, and three grades of 10.

a. Make a histogram of these quiz scores.

b. Find the mean, median, and mode of the quiz scores.

c. Is the data set symmetric, skewed right, skewed left, or none of these? Justify your answer.

6. Students fill out course evaluation forms near the end of a course. The course is given an overall rating by each student. Students rate the course 0, 1, 2, 3, or 4, with 4 being the highest possible rating. The

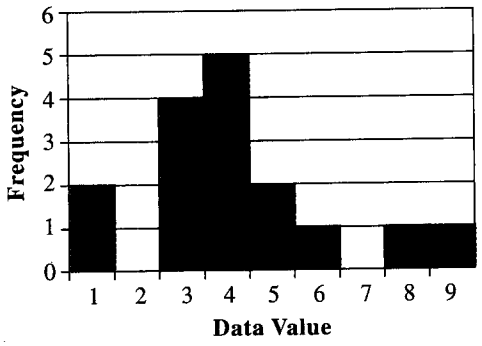
course received one evaluation with a rating of 0, one evaluation with a rating of 1, sixteen evaluations with a rating of 2, nine evaluations with a rating of 3, and eight evaluations with a rating of 4.

a. Make a histogram of the student ratings.

b. Find the mean, median, and mode of the student ratings.

c. Is the data set symmetric, skewed right, skewed left, or none of these? Justify your answer.

7. Consider the following histogram.

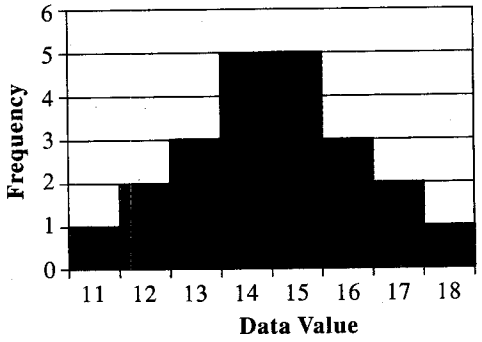


a. Find the mean, median, and mode of the data.

b. Compare the mean and the median. How could you have predicted which of the two would be larger just by looking at the histogram?

c. Is it possible to add a single number to the data set that causes the mean, median, and mode to change? If it is possible, give an example. If it is not, explain why.

8. Consider the following histogram.



a. Find the mean, median, and mode of the data.

b. How do the mean and the median compare? How could you have predicted the relationship between the two just by looking at the histogram?

c. Is it possible to add a single number to the data set that causes the mean, median, and mode to change? If it is possible, give an example. If it is not, explain why.

Problems 9 and 10

Identify whether the mean, median, or mode might be the most appropriate measure of central tendency to use in each case. Give a reason for your choice.

9. a. A medical study needs to record normal human body temperatures.

b. A shoe store manager must determine how many of each size of shoes to order.

10. a. A company lists its typical salary on its website to attract new employees.

b. The National Center for Health Statistics reports the life expectancy for females.

11. Use the data in Table 9.4 to determine the 2002 per-capita personal income for the following group of West Coast states: Hawaii, Alaska, Washington, Oregon, and California.

12. Use the data in Table 9.4 to determine the 2002 per-capita personal income for the following group of East Coast states: Maine, Massachusetts, Rhode Island, Connecticut, and New Jersey.

13. a. Create one example of a data set with five values such that the mean is 50, the median is 55, and the mode is 61.

b. Create one example of a data set with six values such that the range is 27, the mean is 14, the median is 12, and there is no mode.

14. a. Create one example of a data set with five values such that the mean is 19, the median is 15, and the two modes are 10 and 15.

b. Create one example of a data set with six values such that the range is 10, the mean is 10, the median is 9.5, and the smallest value is 5.

Problems 15 through 18

For each set of data shown, do the following.

a. Find the range.

b. Find the median.

c. Find the first and third quartiles.

d. Find the interquartile range.

e. Find the five-number summary.

f. Create a box-and-whisker plot.

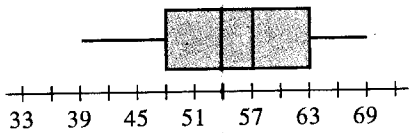
15. 10, 8, 9, 3, 12, 15, 4, 6, 1, 5, 11

16. 2, 5, 10, 20, 6, 4, 12, 15, 9, 8, 16

17. 10, 21, 13, 6, 12, 24, 14, 26, 9, 18

18. 7, 3, 5, 13, 20, 6, 4, 12, 15, 10, 9, 16

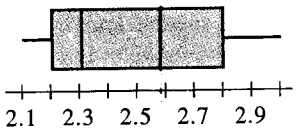
19. Consider the following box-and-whisker plot, which represents a data set.



a. List the five-number summary for the data set.

b. Is the distribution symmetric or skewed? If it is skewed, is it skewed right or skewed left?

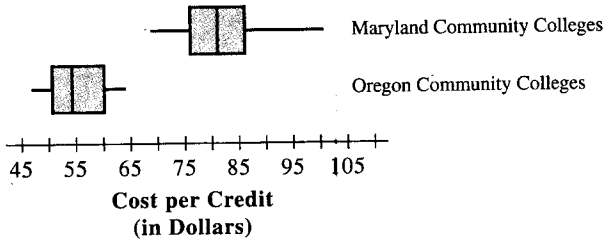
20. Consider the following box-and-whisker plot, which represents a data set.



a. List the five-number summary for the data set.

b. Is the distribution symmetric or skewed? If it is skewed, is it skewed right or skewed left?

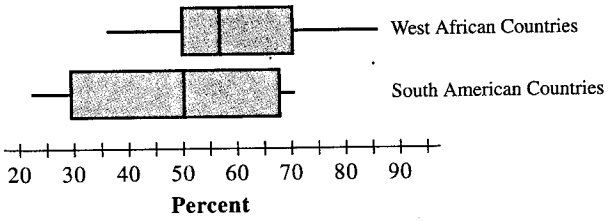
21. The following box-and-whisker plots summarize the cost per credit for community colleges in Maryland and Oregon. *Sources*: Maryland Association of Community Colleges and Oregon Community College Association.



a. Estimate the values for the five-number summary for each state.

b. Describe three conclusions you might draw about the per-credit cost of community colleges in Maryland and Oregon based on a comparison of the box-and-whisker plots.

22. The following box-and-whisker plots summarize the percentages of the population below the poverty line in 9 South American countries and in 10 West African countries. (*Source:* www.nationmaster.com.)



a. Estimate the values for the five-number summary for each plot.

b. Describe three conclusions you might draw about poverty in the West African countries and in the South American countries based on a comparison of the box-and-whisker plots.

Problems 23 and 24

The following table contains earthquake magnitudes and depths for earthquakes in Hawaii during February and March 2004.

Magnitude		Depth (km)	
2.0	2.7	5.9	8.7
2.0	2.2	0.2	7.4
2.3	2.0	3.4	6.9
2.0	2.2	2.7	30.0
2.0	2.2	0.9	8.8
2.3	2.4	1.7	43.2
2.1	3.5	3.1	41.8
2.5	3.2	4.0	9.5
2.3	2.5	33.1	10.0
2.0	2.2	47.2	8.4
2.1	2.9	4.2	8.6
2.0	2.4	2.2	33.6

Source: U.S. Geological Survey.

23. Consider the earthquake magnitude data.

a. Find the five-number summary for the 24 data points.

b. Draw the box-and-whisker plot for the data.

c. Is the distribution symmetric, skewed right, or skewed left?

24. Consider the earthquake depth data.

a. Find the five-number summary for the 24 data points.

b. Draw the box-and-whisker plot for the data.

c. Is the distribution symmetric, skewed right, or skewed left?

25. Find the five-number summary and draw a box-and-whisker plot for the salary data for the Memphis Grizzlies using the table in problem 4.

26. Find the five-number summary and draw a box-and-whisker plot for the salary data for the Minnesota Timberwolves using the table in problem 4.

27. The heights, in inches, of the players on the Denver Nuggets basketball team during the 2003–2004 season are 82, 80, 77, 81, 65, 83, 84, 83, 76, 74, 82, 76, 84, and 81. The heights, in inches, of the players on the Houston Rockets basketball team during the 2003–2004 season are 83, 75, 77, 76, 75, 90, 76, 81, 81, 79, 81, 79, and 71. (*Source:* www.espn.com.)

a. Find the five-number summary for the Denver Nuggets' heights.

b. Find the five-number summary for the Houston Rockets' heights.

c. Draw a box-and-whisker plot for the Denver Nuggets' data.

d. Draw a box-and-whisker plot for the Houston Rockets' data.

e. List three observations based on a comparison of the box-and-whisker plots from parts (c) and (d).

28. Yellowstone National Park is home to over 500 geysers, which are hot springs that erupt periodically. Old Faithful is a geyser that erupts more frequently than other geysers. However, it is not the largest or most regular geyser. A log book of geyser activity is kept by park rangers and others at the Old Faithful Visitors Center. Consider the following 1998 and 2003 data describing the duration of 27 Old Faithful eruptions for a 3-day period in August.

August 1998 Eruption Duration, Minutes:Seconds			August 2003 Eruption Duration, Minutes:Seconds		
3:35	4:10	4:10	4:34	1:56	4:12
4:00	4:15	4:49	4:10	3:53	1:48
4:06	4:06	4:04	1:50	4:23	4:48
2:54	4:23	4:30	4:24	1:51	2:01
4:28	4:10	4:24	4:28	5:00	4:45
3:54	4:03	4:21	2:20	4:08	1:53
4:30	1:47	4:09	4:45	4:06	4:31
3:47	4:36	4:15	2:13	4:30	1:47
3:52	4:10	4:25	4:36	4:22	4:48

Source: www.geyserstudy.org/g_logs.htm.

a. Find the five-number summary for the 1998 eruption data.

b. Find the five-number summary for the 2003 eruption data.

c. Draw a box-and-whisker plot for the 1998 data.

d. Draw a box-and-whisker plot for the 2003 data.

e. List three observations based on a comparison of the box-and-whisker plots from parts (c) and (d).

29. Create a data set with five values, not all the same, so that the mean, median, and the mode are all exactly the same.

30. Create a data set with five values so that the mean is greater than the median and the mode is less than the median.

31. A set of 10 scores from a test in psychology has a mean of 80.7. When the professor goes back to check the grades at a later date, she finds that one score was not recorded. The scores she recorded are 66, 72, 75, 76, 81, 86, 88, 90, and 94. What is the missing score?

32. An instructor records 12 tests in a college history class. The mean of those scores is 76.5. The instructor discovers later that a score of 86 was incorrectly recorded as 68. What should the correct mean be?

33. For the data set 4, 10, 7, 1, 5, do the following.

a. Find the mean of the data.

b. Find the deviation from the mean for each value in the set.

c. Square each deviation from the mean in part (b).

d. Add up the squared deviations from part (c) and divide the total by the number of data values. What is this value called?

34. For the data set 9.1, 7.4, 12.6, 15.6, 11.3, 10.9, do the following.

a. Find the mean of the data.

b. Find the deviation from the mean for each value in the set.

c. Square each deviation from the mean in part (b).

d. Add up the squared deviations from part (c) and divide the total by the number of data values. What is this value called?

35. Find the sample mean, sample variance, and sample standard deviation for each of the following data sets.

a. 4, 6, 7, 10, 13

b. 2, 2, 1, 2, 4, 12

c. 3, 4, 4, 4, 5, 5, 5, 6

36. Find the sample mean, sample variance, and sample standard deviation for each of the following data sets.

a. 3, 0, 4, 5, 14

b. 1, 2, 3, 10, 12, 17

c. 5, 5, 5, 6, 10, 11, 11, 11

37. Consider the data set 3, 10, 9, 7, 15.

a. Find the mean and the sample standard deviation.

b. Modify the data set by adding 5 to each data value, and then find the mean and the sample standard deviation of the new data set.

c. How do the means from parts (a) and (b) compare? Give an explanation of why this result will be true in general.

d. How do the sample standard deviations from parts (a) and (b) compare? Explain why this result will be true in general.

38. Consider the data set 6, 7, 9, 12, 15.

a. Find the mean and the sample standard deviation.

b. Modify the data set by subtracting 3 from each data value, and then find the mean and the sample standard deviation of the new data set.

c. How do the means from parts (a) and (b) compare? Explain why this result will be true in general.

d. How do the sample standard deviations from parts (a) and (b) compare? Explain why this result will be true in general.

39. Consider the data set 9, 7, 3, 10, 15.

a. Find the mean and the sample standard deviation.

b. Modify the data set by multiplying each data value by 3, and then find the mean and the sample standard deviation of the new data set.

c. How do the means from parts (a) and (b) compare? Explain why this result will be true in general.

d. How do the sample standard deviations from parts (a) and (b) compare? Explain why this result will be true in general.

40. Consider the data set 12, 9, 7, 15, 6.

a. Find the mean and the sample standard deviation.

b. Modify the data set by dividing each data value by 10, then find the mean and the sample standard deviation of the new data set.

c. How do the means from parts (a) and (b) compare? Explain why this result will be true in general.

d. How do the sample standard deviations from parts (a) and (b) compare? Explain why this result will be true in general.

41. Find the sample variance and sample standard deviation for the Boston Celtics' heights in Problem 1.

42. Find the sample variance and sample standard deviation for the Boston Celtics' weights in Problem 1.

43. For each of the following data sets, give an example or explain why it is not possible.

a. A set of five values such the population variance is zero.

b. A set of five values such that the population variance is negative.

44. For each of the following data sets, give an example or explain why it is not possible.

a. A set of five values such the sample standard deviation is zero.

b. A set of five values such that the sample variance is less than the sample standard deviation.

Extended Problems

Outliers and Modified Box-and-Whisker Plots

45. An **outlier** is a data point that appears to be not typical of the data as a whole because it is much bigger or much smaller than most data points. In general, this term is imprecise and subject to interpretation. In the case of data sets such as the ones we are studying, and in the context of ranked data, many statisticians have agreed to call any data point an outlier if it is more than 1.5 times the interquartile range, 1.5(IQR), below the first quartile, or if it is more than 1.5(IQR) above the third quartile. In box-and-whisker plots, outliers are usually denoted by an asterisk and the highest and lowest nonoutliers are plotted as the ends of whiskers.

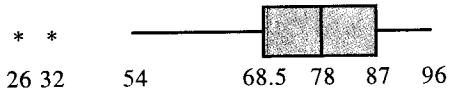
By this definition, there are two outliers in the economics class test scores that we considered in Example 9.22. The interquartile range for that data set was IQR = 18.5, so 1.5(IQR) must be 1.5(18.5), or 27.75. Thus, any scores less than

$$q_1 - 1.5(\text{IQR}) = 68.5 - 27.75 = 40.75$$

or greater than

$$q_3 + 1.5(\text{IQR}) = 87 + 27.75 = 114.75$$

are outliers. Two scores fit the bill, namely, 26 and 32. We will redraw our box-and-whisker plot below, using asterisks to indicate that these scores are outliers. We call this graph a **modified box-and-whisker plot**. Notice that the left whisker now ends at 54, which is the lowest score that is *not* an outlier.



a. Find the five-number summary and draw the box-and-whisker plot for the Minnesota Timberwolves' salary data given in problem 4. Identify any outliers and draw a modified box-and-whisker plot.

b. Find the five-number summary and draw the box-and-whisker plot for the earthquake depth data examined in problem 24. Identify any outliers and draw a modified box-and-whisker plot.

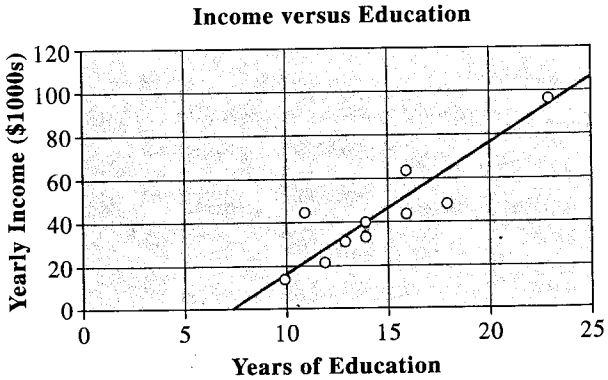
c. Consider the Old Faithful eruption data given in problem 28. What is the shortest eruption duration, in seconds, that would be considered an outlier for the 1998 data? For the 2003 data? Were there outliers in either data set?

Problems 46 through 49: Scatterplots and Regression Lines

Scatterplots are graphs used to display ordered pairs of numeric observations. Think about plotting a collection of points of the form (x, y). By creating a scatterplot, we may discover a relationship between two variables. For instance, sometimes when the ordered pairs are plotted, the points appear to lie roughly on a straight line. In this case, we say that there is a linear relationship between the variables. We will consider scatterplots in which the two variables are linearly related. For example, suppose you interview 10 people and ask them about their educational attainments and income levels with the results shown in the following table.

Person	Years of Education	Yearly Income ($1000s)	Ordered Pair to Plot
1	12	22	(12, 22)
2	16	63	(16, 63)
3	18	48	(18, 48)
4	10	14	(10, 14)
5	14	40	(14, 40)
6	14	34	(14, 34)
7	13	31	(13, 31)
8	11	45	(11, 45)
9	21	96	(21, 96)
10	16	44	(16, 44)

We can plot the ordered pairs as is done in the following graph and draw a line that mimics the trend in the data points and seems to "fit" the points. Although several lines may roughly match the points, the best-fitting line is called the **regression line**.



In this example, the regression line illustrates and quantifies the relationship between education level and income: the line lets you predict the amount of income a person will earn if you know how many years of education he or she has had. According to the regression line for this scat-

ter plot, a person with 20 years of education is predicted to have an annual income of approximately $76,000.

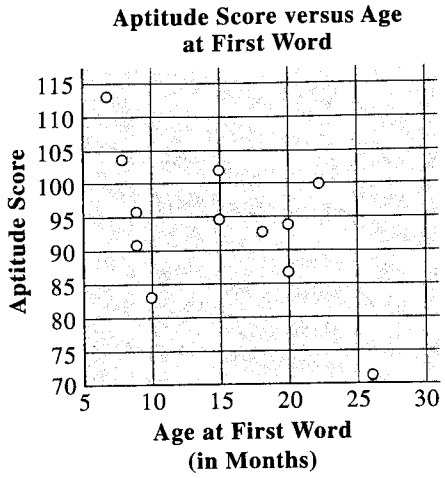
46. a. Suppose aerial surveys were made of a certain wooded area in Alaska on 10 different days. On each day, the wind velocity in miles per hour and the number of black bears sighted were recorded. Construct a scatterplot with wind velocity on the horizontal axis and number of black bears sighted on the vertical axis. Does there appear to be a linear trend in the scatterplot?

Wind Velocity (mph)	Black Bears Sighted
2.1	93
16.7	60
21.1	30
15.9	63
4.9	82
11.8	76
23.6	43
4.0	89
21.5	49
24.4	36

b. A company that assembles electronic parts would like to develop screening tests for new employees. Twelve employees are selected at random and their aptitude test results are listed together with their mean weekly output. Construct a scatterplot with aptitude test results on the horizontal axis and mean weekly output on the vertical axis. Does the scatterplot appear to have a linear trend?

Aptitude Test Results	Mean Weekly Output
6	30
9	49
5	32
8	42
7	39
5	28
8	41
10	46
9	44
11	50

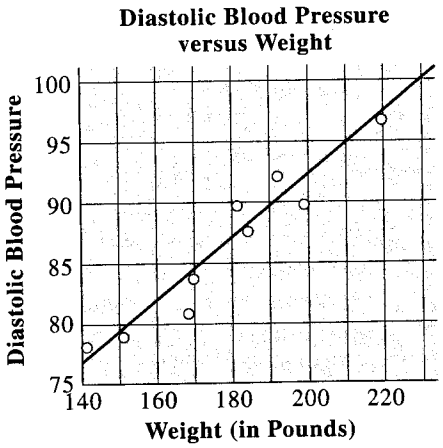
47. A study of cognitive development in young children recorded the age (in months) when they spoke their first word and the scores on an aptitude test taken much later. The data is contained in the following scatterplot.



a. Use a straightedge to draw an approximation to the regression line.

b. Use your line to predict the aptitude test score for a child who spoke his first word at 12 months and another who spoke hers at 25 months of age.

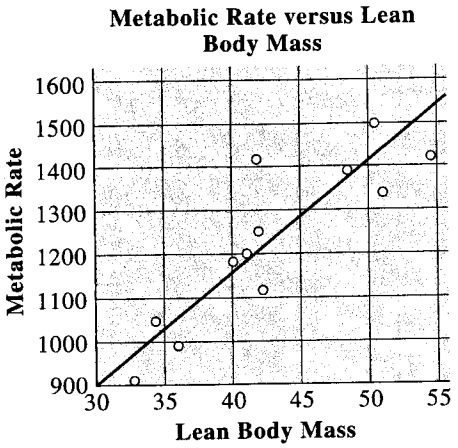
48. A doctor conducted a study to investigate the relationship between weight and diastolic blood pressure of males between 40 and 50 years of age. Consider the following scatterplot of diastolic blood pressure versus weight with the regression line drawn in.



a. Predict the diastolic blood pressure of a 45-year-old man who weighs 160 pounds.

b. If a 50-year-old man has a diastolic blood pressure of 100, use the regression line to predict the man's weight.

49. In a study on obesity involving 12 women, their lean body mass (in kilograms) was compared with their resting metabolic rate. Consider the following scatterplot of metabolic rate versus lean body mass with the regression line drawn in.



a. Predict the resting metabolic rate for a woman with a lean body mass of 40 kilograms.

b. Predict the lean body mass for a woman with a metabolic rate of 1000.

Problems 50 through 52

We can use a straightedge to draw an approximate regression line or the statistical features of a calculator to generate the equation of the regression line. Alternatively, we can use the formula to generate the equation of the best fitting regression line, which is $y - \bar{y} = m(x - \bar{x})$, where m is the slope of the line, $\bar{x}$ is the mean of the x-values, and $\bar{y}$ is the mean of the y-values. If n is the number of points, then m is calculated as follows.

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

The slope calculation may look difficult at first glance. However, the value can be obtained systematically. The symbol $\sum$ (sigma) means that the indicated values are added, as we will show next.

Consider a car manufacturer testing fuel efficiency in a new car model. A test car ran a race course at varying speeds. The following table shows the speed of the car (in miles per hour) and the corresponding fuel efficiency (in miles per gallon).

Speed, x	Miles per Gallon, y
30	34
35	31
40	32
45	30
50	29
55	30
60	28
65	27

The following table shows how to calculate the required values for the slope formula systematically. Shown at the bottom of each column is the sum of the numbers in the column. In addition, the first two columns also show the mean of the values in the column.

x	y	x^2	xy
30	34	900	1020
35	31	1225	1085
40	32	1600	1280
45	30	2025	1350
50	29	2500	1450
55	30	3025	1650
60	28	3600	1680
65	27	4225	1755
$\sum x = 380$	$\sum y = 241$	$\sum x^2 = 19{,}100$	$\sum xy = 11{,}270$
$\bar{x} = 47.5$	$\bar{y} = 30.125$		

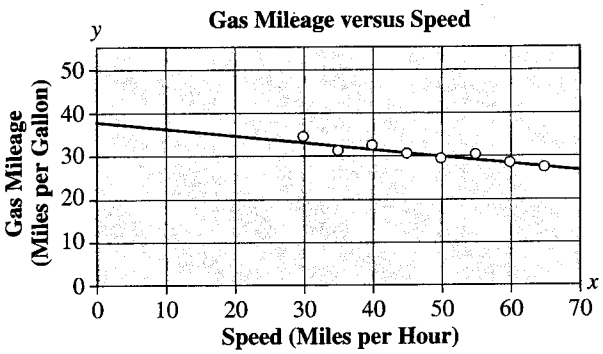
$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(8)(11{,}270) - (380)(241)}{(8)(19{,}100) - (380)^2} = \frac{-1420}{8400} \approx -0.17$$

The regression line has a slope of approximately -0.17; that is, for every mile per hour increase in speed, there is a loss of 0.17 mile per gallon in fuel efficiency. The linear regression equation is calculated as follows.

$$y - \bar{y} = m(x - \bar{x})$$
$$y - 30.125 = -0.17(x - 47.5)$$
$$y = -0.17x + 38.2$$

A linear regression line always passes through the point $(\bar{x}, \bar{y})$. This property allows us to quickly sketch the regression line on the scatterplot relating fuel efficiency

and speed of the race car. Draw the line so that it passes through the y-intercept (0, 38.2) and $(\bar{x}, \bar{y}) = (47.5, 30.125)$ as shown.



50. A geyser is a hot spring that erupts periodically. The following table contains data for Wyoming's famous Old Faithful Geyser. The variable x represents the duration of the eruption (in minutes). The variable y represents the waiting time (in minutes) until the next eruption. (Source: www.stat.duke.)

x	y
4.4	78
3.9	74
4.0	68
4.0	76
3.5	80
4.1	84
2.3	50

a. Make a scatterplot of the data.

b. Create a table like the one shown in the worked-out example prior to this problem and calculate the required values for the slope. Use the formula to determine the equation of the regression line and carefully graph it on the scatterplot.

c. Use the regression line from part (b) to predict the waiting time until the next eruption given that the duration of an eruption is 5 minutes.

51. A college admissions office uses high-school grade point average (GPA) as one of its selection criteria for admitting new students. At the end of the year, the admissions officer randomly selected 10 students from the freshman class and compared their high-school grade point averages and their college grade point averages. In the following table the variable x represents the high-school grade point average, and the variable y represents the college grade point average.

x	y
2.8	2.5
3.2	2.6
3.4	3.1
3.7	3.2
3.5	3.3
3.8	3.3
3.9	3.6
4.0	3.8
3.6	3.9
3.8	4.0

a. Make a scatterplot of the data.

b. Create a table like the one shown prior to problem 50 and calculate the required values for the slope. Use the formula to determine the equation of the regression line and carefully graph it on the scatterplot.

c. Use the regression line from part (b) to predict the college grade point average for a student with a high-school grade point average of 2.2.

d. What do the regression line and scatterplot reveal about the relationship between high-school grade point average and college grade point average?

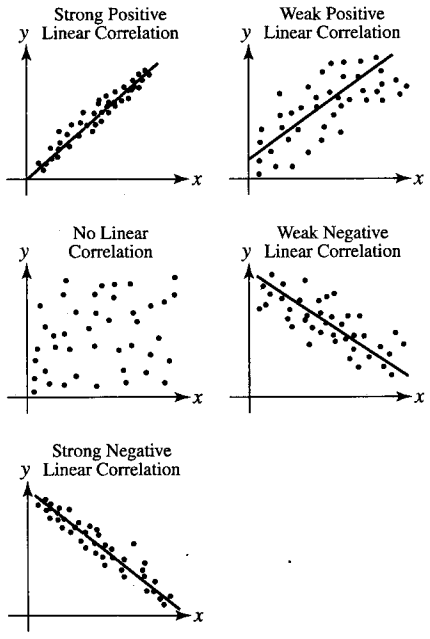
52. a. Determine the equation of the regression line for the aerial survey data from problem 46 (a).

b. Determine the equation of the regression line for the screening test data from problem 46 (b).

Linear Correlation Coefficient

53. Some scatterplots have a strong linear trend, while others have a weak linear trend or no trend at all. If there appears to be a linear relationship between two variables, then we can indicate the strength of that linear relationship by using a **linear correlation coefficient**, which is a number between -1 and 1. If the linear correlation coefficient is 1, then the points in the scatterplot will fall exactly onto a line with a positive slope. If the linear correlation coefficient is -1, then the points in the scatterplot will fall exactly onto a line with a negative slope. If the two variables have no linear relationship, then the linear correlation coefficient is 0.

Instead of calculating the exact value of the linear correlation coefficient, we will use a more intuitive description of correlation. A **strong positive correlation** will indicate that the points in a scatterplot closely follow a linear trend, meaning that as one variable increases, the other variable also increases. At the other extreme, a **strong negative correlation** will indicate that the points in a scatterplot closely follow a linear trend, but that as one variable increases, the other decreases. **Weak positive correlation** or **weak negative correlation** will indicate that the points seem to display a linear trend, but that the points are more scattered. If there is no apparent linear relationship between the variables, we will say they have **no linear correlation**. Consider the following scatterplots.



For each of the following, describe the correlation between the variables as a strong positive correlation, a weak positive correlation, no linear correlation, a weak negative correlation, or a strong negative correlation.

a. Describe the correlation between the variables in problem 47.

b. Describe the correlation between the variables in problem 48.

c. Describe the correlation between the variables in problem 49.

d. Describe the correlation between the variables in problem 50.

54. "Correlation is not causation" is a well-known proverb in statistics. It means that two variables can have a strong positive or strong negative correlation without one quantity causing the other. Tobacco companies have argued this point. While it appears that lung disease, heart disease, and other health problems are strongly correlated with smoking, the correlation does not prove that smoking causes these illnesses. Rather, it is possible that some other factor, perhaps genetic or environmental, leads a person to enjoy smoking and also causes disease. It is not possible to prove this argument wrong, although it may seem implausible given that ingredients in tobacco smoke have been shown to cause disease in animals.

Correlation and causation have played a central role in recent highly publicized lawsuits against large companies. Dow Corning produced silicone-gel breast implants; lawsuits filed against that company claimed the implants caused a host of physical ailments in women. Plaintiffs have sued McDonald's restaurants, claiming that fast food causes obesity. Many people have sued Philip Morris, a cigarette company, claiming that cigarette smoke causes lung cancer. Research the lawsuits against these three companies. In each case, describe the claims made by the plaintiffs. What evidence was offered? Did lawyers prove causation in any of the three cases? How were these lawsuits settled? Write a report to summarize your findings, and include numerical data, if possible.

Key Ideas and Questions

The following questions review the main ideas of this chapter. Write your answers to the questions and then refer to the pages listed by number to make certain that you have mastered these ideas.

1. What is the difference between a population and a sample? pgs. 565–566 What is meant by a representative sample? pg. 567
2. What are some common sources of bias in surveys? pg. 569 What is one kind of unbiased sample? pg. 569
3. What are two different ways to take a simple random sample? pg. 569 What does the 50% refer to in a 50% independent sample? pg. 579 What is one disadvantage to using independent sampling? pg. 579
4. How is a 1-in-10 systematic sample selected? pg. 580 What is one advantage and one disadvantage of using systematic sampling? pg. 581
5. Under what circumstances might a quota sample be taken? pg. 581 What is one advantage and one disadvantage of using quota sampling? pgs. 581–582
6. How is a stratified sample selected? pg. 582 What is a stratum? pg. 582 What is one advantage to using stratified random sampling? pg. 582
7. What is the motivation behind cluster sampling? pg. 583 How are cluster sampling and stratified sampling similar and how are they different? pg. 584
8. What are the three main measures of central tendency? pgs. 595, 596, 600 How is the mean calculated? pg. 595 How is the mean affected if one very large or very small value is added to the data set? pg. 595 What is the difference between the sample mean and the population mean? pg. 596
9. How is the median calculated if there is an odd number of data points? pg. 596 How is the median calculated if there is an even number of data points? pg. 596 What is one advantage to using the median rather than the mean? pg. 597 How does the relationship between the mean and the median provide information about how a distribution is skewed? pg. 599
10. Under what condition would a data set have no mode? pg. 600
11. When would a weighted mean be used? pg. 602
12. What are one advantage and one disadvantage to using the range as a measure of spread? pg. 604
13. How is each quartile calculated? pg. 605 What values make up the five-number summary for a set of data? pg. 606 What does a box-and-whisker plot represent? pg. 606
14. Why does it make sense to use the sample standard deviation rather than the sample variance? pg. 610

Vocabulary

Following is a list of key vocabulary for this chapter. Mentally review each of these terms, write down the meaning of each one in your own words, and use it in a sentence. Then refer to the page number following each term to review any material that you are unsure of before solving the Chapter 9 Review Problems.

SECTION 9.1

Population 565
Elements 565
Variable 566
Measure 566
Census 566
Sample 566
Quantitative Variable 567
Quantitative Data 567
Qualitative Data 567
Qualitative Variable 567
Ordinal Data 567
Nominal Data 567
Representative Sample 567
Statistical Inference 568
Bias 568
Simple Random Sample 569
Random-Number Generator 569
Random-Number Table 569

SECTION 9.2

Sample Survey Design 578
Independent Sampling 579
50% Independent Sample 579
Systematic Sampling 580
1-in-k Systematic Sampling 580
Quota Sampling 581
Statified Sampling 582
Stratum 582
Stratified Random Sample 582
Sampling Units/Clusters 583
Frame 583
Sample 583
Cluster Sampling 583

SECTION 9.3

Measures of Location 595
Measures of Central Tendency/Measures of Center 595
Mean/Arithmetic Mean 595
Sample Mean 596
Population Mean 596
Median 596
Symmetric Data Distribution 598
Skewed Data Distribution 599
Skewed Left/Right 599
Mode 600
Weight 602
Weighted Mean 602
Measures of Variability/Measures of Spread 603
Range 603
First Quartile 605
Third Quartile 605
Interquartile Range (IQR) 605
Five-Number Summary 606
Box-and-Whisker Plot/Box Plot 606
Deviation from the Mean 609
Sample Variance 609
Sample Standard Deviation 610
Population Variance 612
Population Standard Deviation 612

CHAPTER 9 REVIEW PROBLEMS

1. The city council would like to determine how local voters feel about eliminating metered parking downtown. A survey is taken of adults who are shopping downtown on one afternoon.
   a. What is the population in this case?
   b. What is the sample?
   c. What is the variable of interest?
   d. What sources of bias might there be in this sampling procedure?

2. The student union at a university wishes to raise student fees so that a new center for bowling and video games can be built. A group opposed to this plan takes a survey of students coming from the library.
   a. Discuss sources of bias in this survey.
   b. Describe a sampling method that would yield a representative sample.

3. Explain in detail how to choose three people in an unbiased way from a group of five people.

4. A study of heart disease follows the history of 100 people over the course of their lives. For bookkeeping purposes, these people are assigned numbers from 00 to 99. Using the table in Figure 9.4, choose a simple random sample of size 20 from this group. Begin in column 2 of row 109. Use the last two digits and read down the column. List the people included in the sample.

5. A potato chip manufacturer would like to sample bags of potato chips produced in a single day to see how many bags contain at least 12.5 ounces of potato chips. The manufacturer plans to take a sample of 25 packages.
   a. Give the population, the sample, and the variable of interest in this survey.
   b. If the manufacturer knows that 4993 bags of potato chips will be produced in a single day, plans to include the first and last bag in the sample, and takes a 1-in-k systematic sample, then find the values of r and k.
   c. Suppose the manufacturer has 1000 crates of bags of potato chips in a warehouse and plans to take a cluster sample of 10 crates. They will measure every bag of chips in each crate sampled. Label each crate from 000 to 999 and begin in the table in Figure 9.4 in column 2 of row 110. Use the first three digits, read down the column and list the crates that will be part of the sample.

6. The following table contains a list of the presidents of the United States and their ages at inauguration.
   a. Find the mean, median, and mode of the presidents' ages at inauguration.
   b. Give the five-number summary of the data.
   c. Construct a box-and-whisker plot for the age data.

President (age)			
01. Washington (57)	11. Polk (49)	22. Cleveland (47)	33. Truman (60)
02. J. Adams (61)	12. Taylor (64)	23. Harrison (55)	34. Eisenhower (62)
03. Jefferson (57)	13. Fillmore (50)	24. Cleveland (55)	35. Kennedy (43)
04. Madison (57)	14. Pierce (48)	25. McKinley (54)	36. L. B. Johnson (55)
05. Monroe (58)	15. Buchanan (65)	26. T. Roosevelt (42)	37. Nixon (56)
06. J. Q. Adams (57)	16. Lincoln (52)	27. Taft (51)	38. Ford (61)
07. Jackson (61)	17. A. Johnson (56)	28. Wilson (56)	39. Carter (52)
08. Van Buren (54)	18. Grant (46)	29. Harding (55)	40. Reagan (69)
09. Harrison (68)	19. Hayes (54)	30. Coolidge (51)	41. G. H. W. Bush (64)
10. Tyler (51)	20. Garfield (49)	31. Hoover (54)	42. Clinton (46)
	21. Arthur (50)	32. F. D. Roosevelt (51)	43. G. W. Bush (54)

Source: www.infoplease.com.

7. Find the population variance and population standard deviation for the presidents' ages at inauguration from problem 6.

8. a. Find a 1-in-6 systematic sample of the U.S. presidents. Suppose the digit 4 is randomly selected to start the systematic sample. Which presidents are included in the sample?

b. Find the sample mean and sample standard deviation for the sample of presidents found in part (a). Compare the results to the population mean and standard deviation found in problem 7. What do you notice?

9. Use the table in Figure 9.4 to find a 20% independent sample of the presidents from problem 6. Use column 2, beginning with the first digit in row 120, and read across the row. Let the digits 1 and 2 mean that the president is included in the sample. List the presidents that are included in the sample. Explain why the number of presidents included in the sample is not exactly 20% of the total number of presidents of the United States.

10. a. Compute the mean, the median, and the mode of the following data set.

6, 8, 6, 9, 7, 8, 7, 6, 8, 6, 7, 7, 10, 7, 8, 7, 9, 7, 7, 9, 8, 7, 8

b. Is the data symmetric, skewed right or skewed left? Justify your answer.

11. You record the number of minutes you exercise each day over a 2-week period, with one exception. You forgot to record the exercise time for the day 7 of week 2. In the following list of exercise times, the missing time is listed as x.

Week 1 = 40, 20, 27, 20, 30, 28, 45

Week 2 = 30, 22, 25, 25, 30, 27, x

a. Find the mean and median of the week 1 exercise times.

b. How many minutes would you have to exercise on day 7 of week 2 so that the mean exercise times for both weeks would be the same?

c. Is it possible to find a value for x for day 7 of week 2 so that your median exercise times for week 1 and week 2 are the same? If it is possible, give the value. If it is not, explain why not.

12. A student took two-credit, three-credit, and four-credit classes. He had 13 two-credit classes, for which he has a mean grade of $3.2\dot{}$. He had 22 three-credit classes, for which he has a mean grade of 3.6. He had 21 four-credit classes, for which his mean grade is 3.5. Use a weighted mean to compute his overall grade point average.

13. For each of the following data sets, do the following.

(i) Find the range.

(ii) Find the median.

(iii) Find the first and third quartiles.

(iv) Find the interquartile range

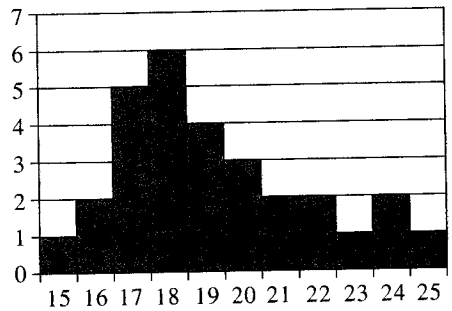
(v) Give the five-number summary.

(vi) Construct a box-and-whisker plot.

a. 4, 1, 2, 5, 8, 2, 6, 9, 4, 3, 1, 5, 10, 4

b. 22, 31, 38, 30, 25, 29, 31, 26, 40, 34, 26, 29

14. Consider the following histogram.



a. Find the mean, median, and mode of the data set.

b. Is the data set symmetric, skewed right, or skewed left? Justify your answer.

c. Give the five-number summary for the data set.

d. Construct a box-and-whisker plot for the data.

15. The following table contains the top 10 and bottom 10 states in terms of average teacher salary for the fall of 2003.

Top 10 Teacher Salaries	Bottom 10 Teacher Salaries
California $56,283	Arkansas $37,753
Connecticut $55,367	West Virginia $38,481
New Jersey $54,158	New Mexico $36,965
Michigan $53,798	Louisiana $37,300
New York $52,600	Nebraska $37,896
Pennsylvania $51,424	Oklahoma $34,877
Massachusetts $52,043	Montana $35,754
Rhode Island $51,076	Mississippi $34,555
Illinois $51,289	North Dakota $33,210
Alaska $49,685	South Dakota $32,416

Source: www.nea.org.

a. Find the five-number summary for the top 10 average teacher salaries and construct a box-and-whisker plot.

b. Find the five-number summary for the bottom 10 average teacher salaries and construct a box-and-whisker plot.

c. Compare the box-and-whisker plots from parts (a) and (b) and list three observations.

16. Consider the average teacher salary data from the previous problem.

a. Find the mean and range for the top 10 average teacher salaries.

b. Find the mean and range for the bottom 10 average teacher salaries.

17. a. Find the population standard deviation for the top 10 average teacher salaries from problem 15.

b. Find the population standard deviation for the bottom 10 average teacher salaries from problem 15.

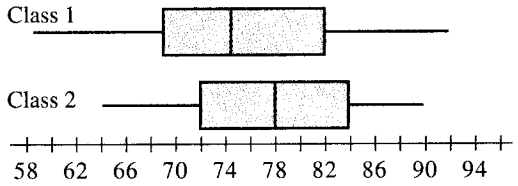
18. Take a stratified random sample of states listed for the average teacher salary data from problem 15. Label the 10 states with the top average salaries 0 to 9 and do the same for the 10 states with the bottom average salaries. Suppose that there are two strata: the top 10 states and the bottom 10 states in terms of average teacher salaries. Select five states from each stratum as directed below.

a. Use the table in Figure 9.4 to select the stratified sample. For the strata containing states with the top average salaries, begin in column 4, row 120, and read across the row. For the strata containing the states with the bottom average salaries, begin in column 5, row 125, and read across the row. List the states in the sample from each stratum.

b. Find the mean and sample standard deviation for the sample of states from the top average salary strata from part (a).

c. Find the mean and sample standard deviation for the sample of states from the bottom average salary strata from part (a).

19. Two classes took the same algebra exam, with test results illustrated next.



a. Give the five-number summary for each class.

b. Based on the box-and-whisker plots for the two classes, which class seems to have done better? Justify your answer.

GEORGE GALLUP (1901–1984), a pioneer of scientific public opinion polling, called polling "a new field of journalism." He earned a Ph.D. with a dissertation on systematic methods of gathering data on reader interest in the content of newspapers. After teaching journalism and advertising at Drake University and Northwestern University, he applied his skills to help his mother-in-law be elected Lieutenant Governor of Iowa. Gallup then took a position as head of an advertising agency research department. In 1935, he established the American Institute of Public Opinion, which became firmly established the next year by correctly predicting the outcome of the presidential election that pitted Franklin D. Roosevelt against Alf Landon. Later, faith in public opinion polls, and the Gallup Poll in particular, was damaged by the incorrect prediction that Thomas Dewey would defeat Harry S. Truman in the 1948 presidential election.

Source © Bettmann/Corbis

Gallup attributed the 1948 debacle mainly to the fact that his organization stopped the polling process 3 weeks before the election. While it may seem obvious that 3 weeks can make a significant difference in a presidential race, remember that television was still in its infancy in 1948, so a television-advertising blitz was impossible. Nonetheless, Truman was able to sway massive numbers of voters by touring the country in a train, making many stops to give speeches. This quaint method of politicking was called a "whistle-stop campaign."

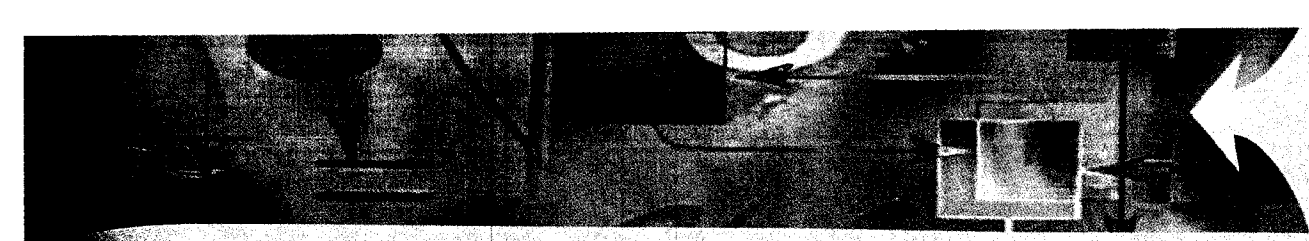
After the 1948 presidential election, the Gallup organization revised its polling so that it continued through the weekend before any presidential election. After years of successful predictions by polling groups, faith in public opinion polls is high, and the most reliable polling organizations make their predictions on a firm scientific basis.

ELMO ROPER (1900–1971) was also a pioneer in the field of public opinion surveys. In contrast to Dr. Gallup, Roper's background was more practical than academic. Although he attended college, he did not earn a degree. Roper spent 1921 through 1928 as a jewelry store owner in Creston, Iowa. After closing the store, he worked as a traveling salesman for 4 years until 1933, when he became a sales analyst for Traub Manufacturing, a jewelry company.

Source: Photo provided by the Roper Center for Public Opinion Research

Roper's first assignment for Traub Manufacturing was to determine why sales of the company's engagement rings were faltering. Roper's research revealed that the rings were not appealing to either of the main markets for engagement rings. The rings were too old-fashioned for the upscale stores and too expensive for the small stores. In fact, the experience must have been transforming because in 1934, Roper formed the market research firm of Cherington, Roper, and Wood along with his friend Richardson Wood and former Harvard Business School Professor Paul T. Cherington. Wood left the company in 1937, and Cherington and Roper split up in 1937. In 1935, while still with Cherington, Roper, and Wood, Wood was able to convince *Fortune* magazine to begin publishing a public opinion poll as a feature called the "Fortune Survey." The Fortune Survey was the first national public opinion poll.

The prestige of Roper's public opinion polls enjoyed the same upward surge as Gallup's when the reelection of President Roosevelt was correctly predicted in 1936, and it suffered the same collapse as Gallup's when President Truman's defeat was incorrectly predicted in 1948. However, Roper always maintained that market research was the bulk of his business, with public opinion polling accounting for about 5% of his company's work.

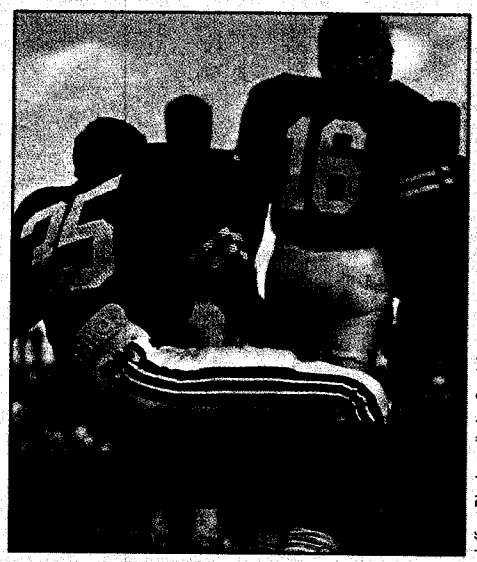


10

Probability

Run That Ball!

In the opening game of the season, the local college's quarterback suffered a broken collarbone, an injury that would cause him to miss most of the season. This football team runs an option offense. In the option offense, the quarterback is more exposed because he has the option of running the ball himself, handing the ball to a running back, or passing to a receiver. Fans noticed that quarterback injuries have occurred more frequently since the option offense was installed 2 years earlier. At the boosters club meeting the following week, the coach was asked if his quarterbacks would continue getting hurt. The coach responded, "Four quarterbacks in the country were injured on Saturday. One was an option quarterback and the others were standard drop-back passers. Just because you run the option does not mean that you are going to get hurt. There is risk in any sport."



Jeffrey Blackman/Index Stock Imagery

The coach's reasoning seems plausible: his quarterback was injured, but the quarterbacks of three other teams were injured as well. On closer inspection, however, we see a fallacy in his reasoning. The NCAA consists of just over 100 division 1-A football teams. Of these, only 4 were using the option offense. Thus, based on the results of the first week of play, the probability that one of the option quarterbacks would be injured is $\frac{1}{4}$, or 25%. By comparison, only 3 of approximately 100, or about 3%, of the rest of the teams in the country lost a quarterback. Looked at this way, the option offense seems hazardous to a quarterback's health.