CHAPTER

16

STATISTICS

Every four years, each major political party in the United States holds a convention to select the party's nominee for President of the United States. Before these conventions are held, each candidate assembles a staff whose job is to plan a successful campaign. This plan relies heavily on statistics: on the collection and organization of data, on the results of opinion polls, and on information about the factors that influence the way people vote. At the same time, newspaper reporters and television commentators assemble other data to keep the public informed on the progress of the candidates.

Election campaigns are just one example of the use of statistics to organize data in a way that enables us to use available information to evaluate the current situation and to plan for the future.

CHAPTER TABLE OF CONTENTS

16-1 COLLECTING DATA

In our daily lives, we often deal with problems that involve many related items of numerical information called **data**. For example, in the daily newspaper we can find data dealing with sports, with business, with politics, or with the weather.

Statistics is the study of numerical data. There are three typical steps in a statistical study:

STEP 1. The collection of data.

STEP 2. The organization of these data into tables, charts, and graphs.

STEP 3. The drawing of conclusions from an analysis of these data.

When these three steps, which describe and summarize the formation and use of a set of data, are included in a statistical study, the study is often called **descriptive statistics**. You will study these steps in this first course. In some cases, a fourth step, in which the analyzed data are used to predict trends and future events, is added.

Data can be either **qualitative** or **quantitative**. For example, a restaurant may ask customers to rate the meal that was served as excellent, very good, good, fair, or poor. This is a qualitative evaluation. Or the restaurant may wish to make a record of each customer tip at different times of the day. This is a quantitative evaluation, which lends itself more readily to further statistical analysis.

Data can be collected in a number of ways, including the following:

1. A written *questionnaire* or list of questions that a person can answer by checking one of several categories or supplying written responses. Categories to be checked may be either qualitative or quantitative. Written responses are usually qualitative.
2. An *interview*, either in person or by telephone, in which answers are given verbally and responses are recorded by the person asking the questions. Verbal answers are usually qualitative.
3. A *log* or a diary, such as a hospital chart or an hourly recording of the outdoor temperature, in which a person records information on a regular basis. This type of information is usually quantitative.

Note: Not all numerical data are quantitative data. For instance, a researcher wishes to investigate the eye color of the population of a certain island. The researcher assigns "blue" to 0, "black" to 1, "brown" to 2, and so on. The resulting data, although numerical, are qualitative since it represents eye color and the assignment was arbitrary.

Sampling

A statistical study may be useful in situations such as the following:

1. A doctor wants to know how effective a new medicine will be in curing a disease.
2. A quality-control team wants to know the expected life span of flashlight batteries made by its company.
3. A company advertising on television wants to know the most frequently watched TV shows so that its ads will be seen by the greatest number of people.

When a statistical study is conducted, it is not always possible to obtain information about every person, object, or situation to which the study applies. Unlike a **census**, in which every person is counted, some statistical studies use only a **sample**, or portion, of the items being investigated.

To find effective medicines, pharmaceutical companies usually conduct tests in which a sample, or small group, of the patients having the disease under study receive the medicine. If the manufacturer of flashlight batteries tested the life span of every battery made, the warehouse would soon be filled with dead batteries. The manufacturer tests only a sample of the batteries to determine their average life span. An advertiser cannot contact every person owning a TV set to determine which shows are being watched. Instead, the advertiser studies TV ratings released by a firm that conducts polls based on a small sample of TV viewers.

For any statistical study, whether based on a census or a sample, to be useful, data must be collected carefully and correctly. Poorly designed sampling techniques result in **bias**, that is, the tendency to favor a selection of certain members of the population which, in turn, produces unreliable conclusions.

Techniques of Sampling

We must be careful when choosing samples:

1. The sample must be fair or unbiased, to reflect the entire population being studied. To know what an apple pie tastes like, it is not necessary to eat the entire pie. Eating a sample, such as a piece of the apple pie, would be a fair way of knowing how the pie tastes. However, eating only the crust or only the apples would be an unfair sample that would not tell us what the entire pie tastes like.
2. The sample must contain a reasonable number of the items being tested or counted. If a medicine is generally effective, it must work for many people. The sample tested cannot include only one or two patients. Similarly, the manufacturer of flashlight batteries cannot make claims based on testing five or 10 batteries. A better sample might include 100 batteries.

3. Patterns of sampling or random selection should be employed in a study. The manufacturer of flashlight batteries might test every 1,000th battery to come off the assembly line. Or, the batteries to be tested might be selected at random.

These techniques will help to make the sample, or the small group, representative of the entire population of items being studied. From the study of the small group, reasonable conclusions can be drawn about the entire group.

EXAMPLE 1

To determine which television programs are the most popular in a large city, a poll is conducted by selecting people at random at a street corner and interviewing them. Outside of which location would the interviewer be most likely to find an unbiased sample?

(1) a ball park (2) a concert hall (3) a supermarket

Solution People outside a ball park may be going to a game or purchasing tickets for a game in the future; this sample may be biased in favor of sports programs. Similarly, those outside a concert hall may favor musical or cultural programs. The best (that is, the fairest) sample or cross section of people for the three choices given would probably be found outside a supermarket.

Answer (3)

Experimental Design

So far we have focused on data collection. In an **experiment**, a researcher imposes a treatment on one or more groups. The **treatment group** receives the treatment, while the **control group** does not.

For instance, consider an experiment of a new medicine for weight loss. Only the treatment group is given the medicine, and conditions are kept as similar as possible for both groups. In particular, both groups are given the same diet and exercise. Also, both groups are of large enough size and are chosen such that they are comprised of representative samples of the general population.

However, it is often not enough to have just a control group and a treatment group. The researcher must keep in mind that people often tend to respond to *any* treatment. This is called the **placebo effect**. In such cases, subjects would report that the treatment worked even when it is ineffective. To account for the placebo effect, researchers use a group that is given a **placebo** or a dummy treatment.

Of course, subjects in the experimental and placebo groups should not know which group they are in (otherwise, psychology will again confound the results). The practice of not letting people know whether or not they have been given the real treatment is called **blinding**, and experiments using blinding are said to be **single-blind experiments**. When the variable of interest is hard to measure or

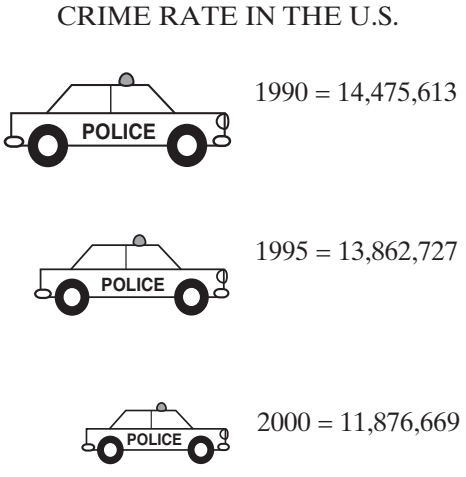
define, **double-blind experiments** are needed. For example, consider an experiment measuring the effectiveness of a drug for attention deficit disorder. The problem is that "attention deficiency" is difficult to define, and so a researcher with a bias towards a particular conclusion may interpret the results of the placebo and treatment groups differently. To avoid such problems, the researchers working directly with the test subjects are not told which group a subject belongs to.

Interpreting Graphs of Data

Oftentimes embellishments to graphs distort the perception of the data, and so you must exercise care when interpreting graphs of data.

1. Two- and three-dimensional figures.

As the graph on the right shows, graphs using two- or three-dimensional figures can distort small changes in the data. The *lengths* show the decrease in crime, but since our eyes tend to focus on the *areas*, the total decrease appears greater than it really is. The reason is because linear changes are increased in higher dimensions. For instance, if a length doubles in value, say from x to $2x$, the area of a square with sides of length x will increase by



$$x^2 \to (2x)^2 = 4x^2,$$

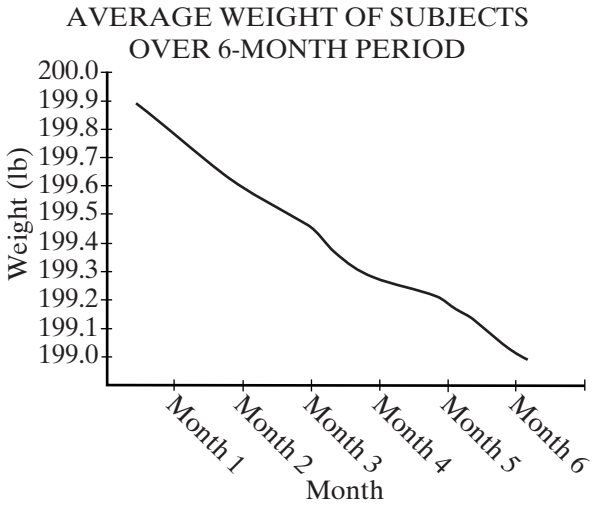
a four-fold increase. Similarly, the volume of a cube with edges of length x will increase by

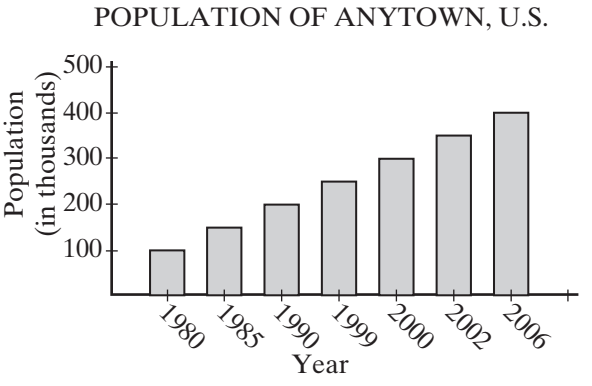
$$x^3 \to (2x)^3 = 8x^3,$$

an eight-fold increase!

2. Horizontal and vertical scales.

The scales used on the vertical and horizontal axes can exaggerate, diminish, and/or distort the nature of the change in the data. For instance, in the graph on the left of the following page, the total change in weight is less than a pound, which is negligible for an adult human. However, the scale used apparently amplifies this amount. While on the right, the unequal horizontal scale makes the population growth appear linear.





EXERCISES

Writing About Mathematics

1. A census attempts to count every person. Explain why a census may be unreliable.

2. A sample of a new soap powder was left at each home in a small town. The occupants were asked to try the powder and return a questionnaire evaluating the product. To encourage the return of the questionnaire, the company promised to send a coupon for a free box of the soap powder to each person who responded. Do you think that the questionnaires that were returned represent a fair sample of all of the persons who tried the soap? Explain why or why not.

Developing Skills

In 3–10, determine if each variable is quantitative or qualitative.

3. Political affiliation

4. Opinions of students on a new music album

5. SAT scores

6. Nationality

7. Cholesterol level

8. Class membership (freshman, sophomore, etc.)

9. Height

10. Number of times the word "alligator" is used in an essay.

In 11–18, in each case a sample of students is to be selected and the height of each student is to be measured to determine the average height of a student in high school. For each sample:

a. Tell whether the sample is biased or unbiased.

b. If the sample is biased, explain how this might affect the outcome of the survey.

11. The basketball team

12. The senior class

13. All 14-year-old students

14. All girls

15. Every tenth person selected from an alphabetical list of all students

16. Every fifth person selected from an alphabetical list of all boys

17. The first three students who report to the nurse on Monday

18. The first three students who enter each homeroom on Tuesday

In 19–24, in each case the Student Organization wishes to interview a sample of students to determine the general interests of the student body. Two questions will be asked: "Do you want more pep rallies for sports events? Do you want more dances?" For each location, tell whether the Student Organization would find an unbiased sample at that place. If the sample is biased, explain how this might influence the result of the survey.

19. The gym, after a game

20. The library

21. The lunchroom

22. The cheerleaders' meeting

23. The next meeting of the Junior Prom committee

24. A homeroom section chosen at random

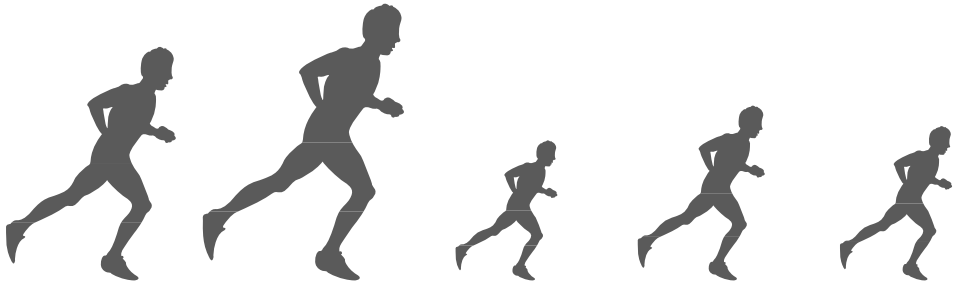
25. A statistical study is useful when reliable data are collected. At times, however, people may exaggerate or lie when answering a question. Of the six questions that follow, find the *three* questions that will most probably produce the largest number of *unreliable* answers.

(1) What is your height? (2) What is your weight?
(3) What is your age? (4) In which state do you live?
(5) What is your income? (6) How many people are in your family?

26. List the three steps necessary to conduct a statistical study.

27. Explain why the graph below is misleading.

SUMMER OLYMPIC GAMES CHAMPIONS
100-METER RACE



1988	1992	1996	2000	2004
Carl Lewis	Linford Christie	Donovan Bailey	Maurice Green	Justin Gatlin
(USA)	(GBR)	(CAN)	(USA)	(USA)
9.92 sec	9.96 sec	9.84 sec	9.87 sec	9.85 sec

28. Investigators at the University of Kalamazoo were interested in determining whether or not women can determine a man's preference for children based on the way that he looks. Researchers asked a group of 20 male volunteers whether or not they liked children. The researchers then showed photographs of the faces of the men to a group of 10 female volunteers and asked them to pick out which men they thought liked children. The women correctly identified over 90% of the men who said they liked children. The researchers concluded that women could identify a man's preference for children based on the way that he looks. Identify potential problems with this experiment.

Hands-On Activity

Collect *quantitative* data for a statistical study.

1. Decide the topic of the study. What data will you collect?

2. Decide how the data will be collected. What will be the source(s) of that data?

a. Questionnaires

b. Personal interviews

c. Telephone interviews

d. Published materials from sources such as almanacs or newspapers.

3. Collect the data. How many values are necessary to obtain reliable information?

Keep the data that you collect to use as you learn more about statistical studies.

16-2 ORGANIZING DATA

Data are often collected in an unorganized and random manner. For example, a teacher recorded the number of days each of 25 students in her class was absent last month. These absences were as follows:

$$0, 3, 1, 0, 4, 2, 1, 3, 5, 0, 2, 0, 0, 0, 4, 0, 1, 1, 2, 1, 0, 7, 3, 1, 0$$

How many students were absent fewer than 2 days? What was the number of days for which the most students were absent? How many students were absent more than 5 days? To answer questions such as these, we find it helpful to organize the data.

One method of organizing data is to write it as an ordered list. In order from least to greatest, the absences become:

$$0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 7$$

We can immediately observe certain facts from this ordered list: more students were absent 0 days than any other number of days, the same number of students were absent 5 and 7 days. However, for more a quantitative analysis, it is useful to make a table.

Preparing a Table

In the left column of the accompanying table, we list the data values (in this case the number of absences) in order. We start with the largest number, 7, at the top and go down to the smallest number, 0.

For each occurrence of a data value, we place a **tally** mark, |, in the row for that number. For example, the first data value in the teacher's list is 0, so we place a tally in the 0 row; the second value is 3, so we place a tally in the 3 row. We follow this procedure until a tally for each data value is recorded in the proper row. To simplify counting, we write every fifth tally as a diagonal mark passing through the first four tallies: 𝍸.

Absences	Tally
7	|
6	
5	|
4	||
3	|||
2	|||
1	𝍸 |
0	𝍸 ||||

Once the data have been organized, we can count the number of tally marks in each row and add a column for the **frequency**, that is, the number of times that a value occurs in the set of data. When there are no tally marks in a row, as for the row showing 6 absences, the frequency is 0. The sum of all of the frequencies is called the **total frequency**. In this case, the total frequency is 25. (It is always wise to check the total frequency to be sure that no data value was overlooked or duplicated in tallying.) From the table, called a **frequency distribution table**, it is now easy to see that 15 students were absent fewer than 2 days, that more students were absent 0 days (9) than any other number of days, and that 1 student was absent more than 5 days.

Absences	Tally	Frequency
7	|	1
6		0
5	|	1
4	||	2
3	|||	3
2	|||	3
1	𝍸 |	6
0	𝍸 ||||	9
Total frequency		25

Grouped Data

A teacher marked a set of 32 test papers. The grades or scores earned by the students were as follows:

90, 85, 74, 86, 65, 62, 100, 95, 77, 82, 50, 83, 77, 93, 73, 72,
98, 66, 45, 100, 50, 89, 78, 70, 75, 95, 80, 78, 83, 81, 72, 75

Because of the large number of different scores, it is convenient to organize these data into **groups** or **intervals**, which must be equal in size. Here we will use six intervals: 41–50, 51–60, 61–70, 71–80, 81–90, 91–100. Each interval has a length of 10, found by subtracting the starting point of an interval from the starting point of the next higher interval.

For each test score, we now place a tally mark in the row for the interval that includes that score. For example, the first two scores in the list above are 90 and 85, so we place two tally marks in the interval 81–90. The next score is 74, so we place a tally mark in the interval 71–80. When all of the scores have been tallied, we write the frequency for each interval.

This table, containing a set of intervals and the corresponding frequency for each interval, is an example of **grouped data**.

Interval	Tally	Frequency
91–100	𝍸 |	6
81–90	𝍸 |||	8
71–80	𝍸 𝍸 |	11
61–70	||||	4
51–60		0
41–50	|||	3

When unorganized data are grouped into intervals, we must follow certain rules in setting up the intervals:

1. The intervals must cover the complete range of values. The **range** is the difference between the highest and lowest values.
2. The intervals must be equal in size.
3. The number of intervals should be between 5 and 15. The use of too many or too few intervals does not make for effective grouping of data. We usually use a large number of intervals, for example, 15, only when we have a very large set of data, such as hundreds of test scores.
4. Every data value to be tallied must fall into one and only one interval. Thus, the intervals should not overlap. When an interval ends with a counting number, the following interval begins with the next counting number.
5. The intervals must be listed in order, either highest to lowest or lowest to highest.

These rules tell us that there are many ways to set up tables, all of them correct, for the same set of data. For example, here is another correct way to group the 32 unorganized test scores given at the beginning of this section. Note that the length of the interval here is 8.

Interval	Tally	Frequency
93–100	𝍸 |	6
85–92	||||	4
77–84	𝍸 ||||	9
69–76	𝍸 ||	7
61–68	|||	3
53–60		0
45–52	|||	3

Constructing a Stem-and-Leaf Diagram

Another method of displaying data is called a **stem-and-leaf diagram**. The stem-and-leaf diagram groups the data without losing the individual data values.

A group of 30 students were asked to record the length of time, in minutes, spent on math homework yesterday. They reported the following data:

38, 15, 22, 20, 25, 44, 5, 40, 38, 22, 20, 35, 20, 0, 36,
27, 37, 26, 33, 25, 17, 45, 22, 30, 18, 48, 12, 10, 24, 27

To construct a stem-and-leaf diagram for the lengths of time given, we begin by choosing part of the data values to be the stem. Since every score is a one- or two-digit number, we will choose the tens digit as a convenient stem. For the one-digit numbers, 0 and 5, the stem is 0; for the other data values, the stem is 1, 2, 3, or 4. Then the units digit will be the leaf. We construct the diagram as follows:

STEP 1. List the stems, starting with 4, under one another to the left of a vertical line beneath a crossbar.

Stem	Leaf
4	
3	
2	
1	
0	

STEP 2. Enter each score by writing its leaf (the units digit) to the right of the vertical line, following the appropriate stem (its tens value). For example, enter 38 by writing 8 to the right of the vertical line, after stem 3.

Stem	Leaf
4	
3	8
2	
1	
0	

STEP 3. Add the other scores to the diagram until all are entered.

Stem	Leaf
4	4 0 5 8
3	8 8 5 6 7 3 0
2	2 0 5 2 0 0 7 6 5 2 4 7
1	5 7 8 2 0
0	5 0

STEP 4. Arrange the leaves in order after each stem.

Stem	Leaf
4	0 4 5 8
3	0 3 5 6 7 8 8
2	0 0 0 2 2 2 4 5 5 6 7 7
1	0 2 5 7 8
0	0 5

STEP 5. Add a key to demonstrate the meaning of each value in the diagram.

Key: 3 | 0 = 30

EXAMPLE 1

The following data consist of the weights, in kilograms, of a group of 30 students:

70, 43, 48, 72, 53, 81, 76, 54, 58, 64, 51, 53, 75, 62, 84,
67, 72, 80, 88, 65, 60, 43, 53, 42, 57, 61, 55, 75, 82, 71

a. Organize the data in a table. Use five intervals starting with 40–49.

b. Based on the grouped data, which interval contains the greatest number of students?

c. How many students weigh less than 70 kilograms?

Solution **a.**

Interval	Tally	Frequency (number)
80–89	𝍸	5
70–79	𝍸 ||	7
60–69	𝍸 |	6
50–59	𝍸 |||	8
40–49	||||	4

b. The interval 50–59 contains the greatest number of students, 8. *Answer*

c. The three lowest intervals, namely 40–49, 50–59, and 60–69, show weights less than 70 kilograms. Add the frequencies in these three intervals:
$4 + 8 + 6 = 18$ *Answer*

EXAMPLE 2

Draw a stem-and-leaf diagram for the data in Example 1.

Solution Let the tens digit be the stem and the units digit the leaf.

(1) Enter the data values in the given order:

Stem	Leaf
8	1 4 0 8 2
7	0 2 6 5 2 5 1
6	4 2 7 5 0 1
5	3 4 8 1 3 3 7 5
4	3 8 3 2

(2) Arrange the leaves in numerical order after each stem:

Stem	Leaf
8	0 1 2 4 8
7	0 1 2 2 5 5 6
6	0 1 2 4 5 7
5	1 3 3 3 4 5 7 8
4	2 3 3 8

(3) Add a key indicating unit of measure:

Key: 5 | 1 = 51 kg

EXERCISES

Writing About Mathematics

1. Of the examples given above, which gives more information about the data: the table or the stem-and-leaf diagram? Explain your answer.

2. A set of data ranges from 2 to 654. What stem can be used for this set of data when drawing a stem-and-leaf diagram? What leaves would be used with this stem? Explain your choices.

Developing Skills

3. a. Copy and complete the table to group the data, which represent the heights, in centimeters, of 36 students:

162, 173, 178, 181, 155, 162, 168, 147, 180,
171, 168, 183, 157, 158, 180, 164, 160, 171,
183, 174, 166, 175, 169, 180, 149, 170, 150,
158, 162, 175, 171, 163, 158, 163, 164, 177

Interval	Tally	Frequency
180–189		
170–179		
160–169		
150–159		
140–149		

b. Use the grouped data to answer the following questions:

(1) How many students are less than 160 centimeters in height?

(2) How many students are 160 centimeters or more in height?

(3) Which interval contains the greatest number of students?

(4) Which interval contains the least number of students?

c. Display the data in a stem-and-leaf diagram. Use the first two digits of the numbers as the stems.

d. What is the range of the data?

e. How many students are taller than 175 centimeters?

4. a. Copy and complete the table to group the data, which gives the lifespan, in hours, of 50 flashlight batteries:

73, 81, 92, 80, 108, 76, 84, 102, 58, 72,
82, 100, 70, 72, 95, 105, 75, 84, 101, 62,
63, 104, 97, 85, 106, 72, 57, 85, 82, 90,
54, 75, 80, 52, 87, 91, 85, 103, 78, 79,
91, 70, 88, 73, 67, 101, 96, 84, 53, 86

Interval	Tally	Frequency
50–59		
60–69		
70–79		
80–89		
90–99		
100–109		

b. Use the grouped data to answer the following questions:

(1) How many flashlight batteries lasted for 80 or more hours?

(2) How many flashlight batteries lasted fewer than 80 hours?

(3) Which interval contains the greatest number of batteries?

(4) Which interval contains the least number of batteries?

c. Display the data in a stem-and-leaf diagram. Use the digits from 5 through 10 as the stems.

d. What is the range of the data?

e. What is the probability that a battery selected at random lasted more than 100 hours?

5. The following data consist of the hours spent each week watching television, as reported by a group of 38 teenagers:

13, 20, 17, 36, 25, 21, 9, 32, 20, 17, 12, 19, 5, 8, 11, 28, 25, 18,
19, 22, 4, 6, 0, 10, 16, 3, 27, 31, 15, 18, 20, 17, 3, 6, 19, 25, 4, 7

a. Construct a table to group these data, using intervals of 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, and 35–39.

b. Construct a table to group these data, using intervals of 0–7, 8–15, 16–23, 24–31, and 32–39.

c. Display the data in a stem-and-leaf diagram.

d. What is the range of the data?

e. What is the probability that a teenager, selected at random from this group, spends less than 4 hours watching television each week?

6. The following data show test scores for 30 students:

90, 83, 87, 71, 62, 46, 67, 72, 75, 100, 93, 81, 74, 75, 82,
83, 83, 84, 92, 58, 95, 98, 81, 88, 72, 59, 95, 50, 73, 93

a. Construct a table, using intervals of length 10 starting with 91–100.

b. Construct a table, using intervals of length 12 starting with 89–100.

c. For the grouped data in part **a**, which interval contains the greatest number of students?

d. For the grouped data in part **b**, which interval contains the greatest number of students?

e. Do the answers for parts **c** and **d** indicate the same general region of test scores, such as "scores in the eighties"? Explain your answer.

7. For the ungrouped data from Exercise 5, tell why each of the following sets of intervals is not correct for grouping the data.

a.

Interval
25–38
13–24
0–12

b.

Interval
30–39
20–29
10–19
5–9
0–4

c.

Interval
32–40
24–32
16–24
8–16
0–8

d.

Interval
33–40
25–32
17–24
9–16
1–8

Hands-On Activity

Organize the data that you collected in the Hands-On Activity for Section 16-1.

1. Use a stem-and-leaf diagram.

a. Decide what will be used as stems.

b. Decide what will be used as leaves.

c. Construct the diagram.

d. Check that the number of leaves in the diagram equals the number of values in the data collected.

2. Use a frequency table.

a. How many intervals will be used?

b. What will be the length of each interval?

c. What will be the starting and ending points of each interval? Check that the intervals do not overlap, are equal in size, and that every value falls into only one interval.

d. Tally the data.

e. List the frequency for each interval.

f. Check that the total frequency equals the number of values in the data collected.

3. Decide which method of organization is better for your data. Explain your choice.

Keep your organized data to work with as you learn more about statistics.

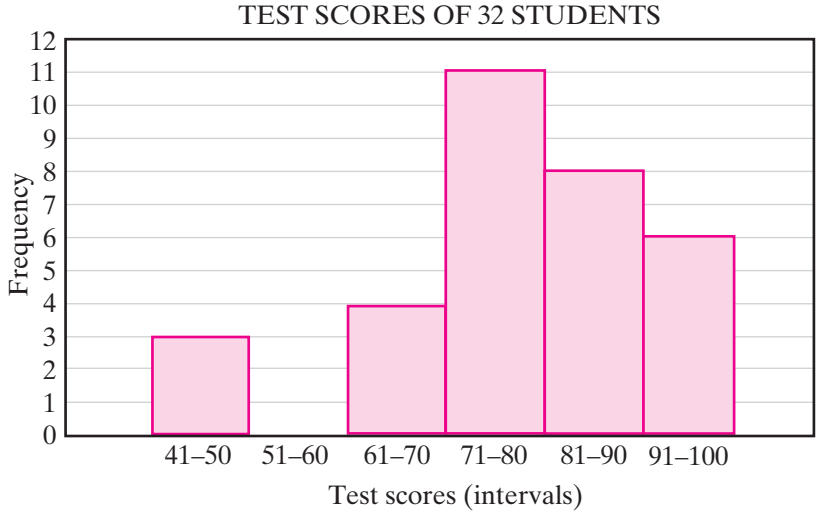
16-3 THE HISTOGRAM

In Section 16-2 we organized data by grouping them into intervals of equal length. After the data have been organized, a graph can be used to visualize the intervals and their frequencies.

The table below shows the distribution of test scores for 32 students in a class. The data have been organized into six intervals of length 10.

Test Scores (Intervals)	Frequency (Number of Scores)
91–100	6
81–90	8
71–80	11
61–70	4
51–60	0
41–50	3

We can use a histogram to display the data graphically. A **histogram** is a vertical bar graph in which each interval is represented by the width of the bar and the frequency of the interval is represented by the height of the bar. The bars are placed next to each other to show that, as one interval ends, the next interval begins.



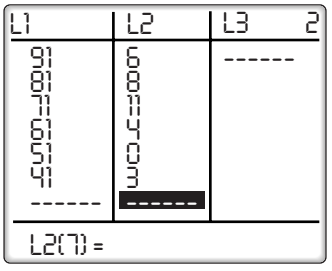
In the above histogram, the intervals are listed on the horizontal axis in the order of increasing scores, and the frequency scale is shown on the vertical axis. The first bar shows that 3 students had test scores in the interval 41–50. Since no student scored in the interval 51–60, there is no bar for this interval. Then, 4 students scored between 61 and 70; 11 between 71 and 80; 8 between 81 and 90; and 6 between 91 and 100.

Except for an interval having a frequency of 0, the interval 51–60 in this example, there are no gaps between the bars drawn in a histogram. Since the histogram displays the frequency, or number of data values, in each interval, we sometimes call this graph a **frequency histogram**.

A graphing calculator can display a frequency histogram from the data on a frequency distribution table.

(1) Clear L_1 and L_2 with the ClrList function by pressing STAT 4 2nd L1 , 2nd L2 ENTER.

(2) Press STAT 1 to edit the lists. L_1 will contain the minimum value of each interval. Move the cursor to the first entry position in L_1. Type the value and then press ENTER. Type the next value and then press ENTER. Repeat this process until all the minimum values of the intervals have been entered.

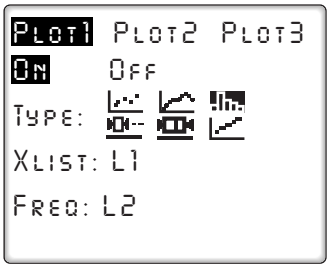


(3) Repeat the process to enter the frequencies that correspond to each interval in L_2.

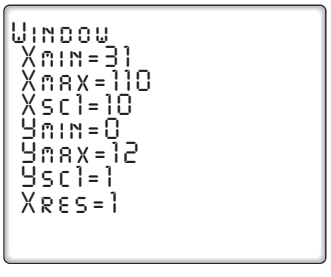
(4) Clear any functions in the Y= menu.

(5) Turn on Plot1 from the STAT PLOT menu, and configure it to graph a histogram. Make sure to also set Xlist to L_1 and Freq to L_2.

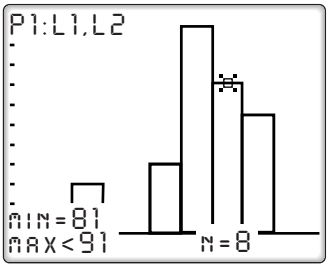
ENTER: 2nd STAT PLOT 1 ENTER
▼ ▶ ▶ ENTER ▶ 2nd
L1 ▼ 2nd L2



(6) In the WINDOW menu, accessed by pressing WINDOW, enter Xmin as 31, the length of one interval less than the smallest interval value and Xmax as 110, the length of one interval more than the largest interval value. Enter Xscl as 10, the length of the interval. The Ymin is 0 and Ymax is 12 to be greater than the largest frequency.



(7) Press GRAPH to draw the graph. We can view the frequency (n) associated with each interval by pressing TRACE. Use the left and right arrow keys to move between intervals.



EXAMPLE 1

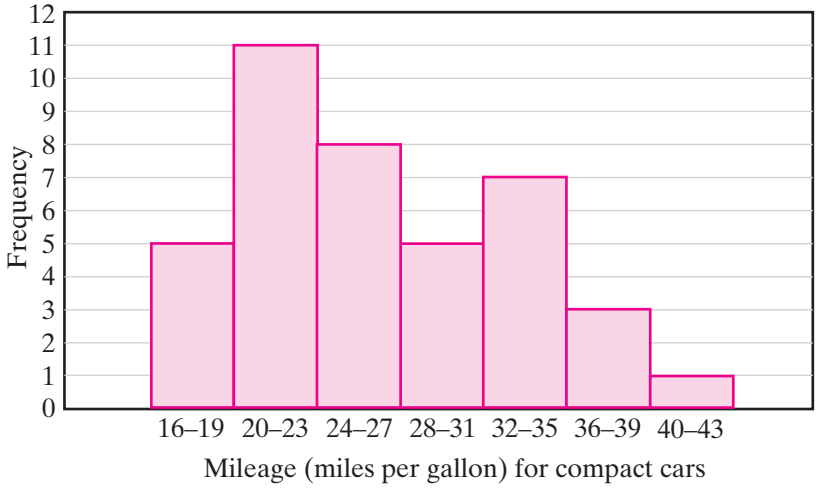
The table on the right represents the number of miles per gallon of gasoline obtained by 40 drivers of compact cars. Construct a frequency histogram based on the data.

Interval	Frequency
16–19	5
20–23	11
24–27	8
28–31	5
32–35	7
36–39	3
40–43	1

Solution (1) Draw and label a vertical scale to show frequencies. The scale starts at 0 and increases to include the highest frequency in any one interval (here, it is 11).

(2) Draw and label intervals of equal length on a horizontal scale. Label the horizontal scale, telling what the numbers represent.

(3) Draw the bars vertically, leaving no gaps between the intervals.



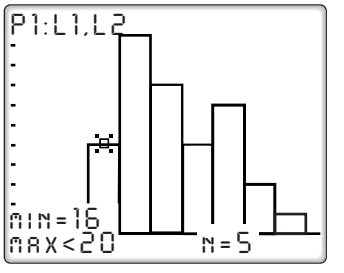
Calculator Solution

(1) Press STAT 1 to edit the lists and enter the minimum value of each interval into L_1: 16, 20, 24, 28, 32, 36, 40. Use the arrow key to move into L_2, and enter the corresponding frequencies: 5, 11, 8, 5, 7, 3, 1.

(2) Go to the STAT PLOT menu and choose Plot1 by pressing 2nd STAT PLOT 1. Move the cursor with the arrow keys, then press ENTER to select On and the histogram. Type 2nd L1 into Xlist and 2nd L2 into Freq.

(3) Set the Window. Each interval has length 4, so set Xmin to 12 (4 less than the smallest interval value), Xmax to 44 (4 more than the largest interval value), and Xscl to 4. Make Ymin 0 and Ymax 12 to be greater than the largest frequency.

(4) Draw the graph by pressing GRAPH. Press TRACE and use the right and left arrow keys to show the frequencies, the heights of the vertical bars.



EXAMPLE 2

Use the histogram constructed in Example 1 to answer the following questions:

a. In what interval is the greatest frequency found?

b. What is the number (or frequency) of cars reporting mileages between 28 and 31 miles per gallon?

c. For what interval are the fewest cars reported?

d. How many of the cars reported mileage greater than 31 miles per gallon?

e. What percent of the cars reported mileage from 24 to 27 miles per gallon?

Solution

a. 20–23

b. 5

c. 40–43

d. Add the frequencies for the three highest intervals. The interval 32–35 has a frequency of 7; 36–39 a frequency of 3; 40–43 a frequency of 1: $7 + 3 + 1 = 11$.

e. The interval 24–27 has a frequency of 8. The total frequency for this survey is 40. $\frac{8}{40} = \frac{1}{5} = 20\%$.

Answers **a.** 20–23 **b.** 5 **c.** 40–43 **d.** 11 **e.** 20%

EXERCISES

Writing About Mathematics

1. Compare a stem-and-leaf diagram with a frequency histogram. In what ways are they alike and in what ways are they different?

2. If the data in Example 1 had been grouped into intervals with a lowest interval of 16–20, what would be the endpoints for the other intervals? Would you be able to determine the frequency for each new interval? Explain why or why not.

Developing Skills

In 3–5, in each case, construct a frequency histogram for the grouped data. Use graph paper or a graphing calculator.

3.

Interval	Frequency
91–100	5
81–90	9
71–80	7
61–70	2
51–60	4

4.

Interval	Frequency
30–34	5
25–29	10
20–24	10
15–19	12
10–14	0
5–9	2

5.

Interval	Frequency
1–3	24
4–6	30
7–9	28
10–12	41
13–15	19
16–18	8

6. For the table of grouped data given in Exercise 5, answer the following questions:

a. What is the total frequency in the table?

b. What interval contains the greatest frequency?

c. The number of data values reported for the interval 4–6 is what percent of the total number of data values?

d. How many data values from 10 through 18 were reported?

Applying Skills

7. Towering Ted McGurn is the star of the school's basketball team. The number of points scored by Ted in his last 20 games are as follows:

36, 32, 28, 30, 33, 36, 24, 33, 29, 30, 30, 25, 34, 36, 34, 31, 36, 29, 30, 34

a. Copy and complete the table to find the frequency for each interval.

Interval	Tally	Frequency
35–37		
32–34		
29–31		
26–28		
23–25		

b. Construct a frequency histogram based on the data found in part **a**.

c. Which interval contains the greatest frequency?

d. In how many games did Ted score 32 or more points?

e. In what percent of these 20 games did Ted score fewer than 26 points?

8. Thirty students on the track team were timed in the 200-meter dash. Each student's time was recorded to the *nearest tenth* of a second. Their times are as follows:

29.3, 31.2, 28.5, 37.6, 30.9, 26.0, 32.4, 31.8, 36.6, 35.0,
38.0, 37.0, 22.8, 35.2, 35.8, 37.7, 38.1, 34.0, 34.1, 28.8,
29.6, 26.9, 36.9, 39.6, 29.9, 30.0, 36.0, 36.1, 38.2, 37.8

a. Copy and complete the table to find the frequency in each interval.

b. Construct a frequency histogram for the given data.

c. Determine the number of students who ran the 200-meter dash in under 29 seconds.

d. If a student on the track team is chosen at random, what is the probability that he or she ran the 200-meter dash in fewer than 29 seconds?

Interval	Tally	Frequency
37.0–40.9		
33.0–36.9		
29.0–32.9		
25.0–28.9		
21.0–24.9		

Hands-On-Activity

Construct a histogram to display the data that you collected and organized in the Hands-On Activities for Sections 16-1 and 16-2.

1. Draw the histogram on graph paper.

2. Follow the steps in this section to display the histogram on a graphing calculator.

16-4 THE MEAN, THE MEDIAN, AND THE MODE

In a statistical study, after we have collected the data, organized them, and presented them graphically, we then analyze the data and summarize our findings. To do this, we often look for a representative, or typical, score.

Averages in Arithmetic

In your previous study of arithmetic, you learned how to find the average of two or more numbers. For example, to find the average of 17, 25, and 30:

STEP 1. Add these three numbers: $17 + 25 + 30 = 72$.

STEP 2. Divide this sum by 3 since there are three numbers: $72 \div 3 = 24$.

The average of the three numbers is 24.

Averages in Statistics

The word **average** has many different meanings. For example, there is an *average* of test scores, a batting *average*, the *average* television viewer, an *average* intelligence, and the *average* size of a family. These averages are *not* found by

the same rule or procedure. Because of this confusion, in statistics we speak of **measures of central tendency**. These measures are numbers that usually fall somewhere in the center of a set of organized data.

We will discuss three measures of central tendency: the mean, the median, and the mode.

The Mean

In statistics, the arithmetic average previously studied is called the **mean** of a set of numbers. It is also called the **arithmetic mean** or the **numerical average**. The mean is found in the same way as the arithmetic average is found.

Procedure

To find the mean of a set of *n* numbers, add the numbers and divide the sum by *n*. The symbol used for the mean is $\bar{x}$.

For example, if Ralph's grades on five tests in science during this marking period are 93, 80, 86, 72, and 94, he can find the mean of his test grades as follows:

STEP 1. Add the five data values: $93 + 80 + 86 + 72 + 94 = 425$.

STEP 2. Divide this sum by 5, the number of tests: $425 \div 5 = 85$.

The mean (arithmetic average) is 85.

Let us consider another example. In a car wash, there are seven employees whose ages are 17, 19, 20, 17, 46, 17, and 18. What is the mean of the ages of these employees?

Here, we add the seven ages to get a sum of 154. Then, $154 \div 7 = 22$. While the mean age of 22 is the correct answer, this measure does *not* truly represent the data. Only one person is older than 22, while six people are under 22. For this reason, we will look at another measure of central tendency that will eliminate the extreme case (the employee aged 46) that is *distorting* the data.

The Median

The **median** is the middle value for a set of data arranged in numerical order. For example, the median of the ages 17, 19, 20, 17, 46, 17, and 18 for the car-wash employees can be found in the following manner:

STEP 1. Arrange the ages in numerical order: 17, 17, 17, 18, 19, 20, 46

STEP 2. Find the middle number: 17, 17, 17, 18, 19, 20, 46

↑

The median is 18 because there are three ages less than 18 and three ages greater than 18. The median, 18, is a better indication of the typical age of the

employees than the mean, 22, because there are so many younger people working at the car wash.

Now, let us suppose that one of the car-wash employees has a birthday, and her age changes from 17 to 18. What is now the median age?

STEP 1. Arrange the ages in numerical order: 17, 17, 18, 18, 19, 20, 46

STEP 2. Find the middle number: 17, 17, 18, 18, 19, 20, 46

↑

The median, or middle value, is again 18. We can no longer say that there are three ages less than 18 because one of the three youngest employees is now 18.

We can say, however, that:

1. the median is 18 because there are three ages less than or equal to 18 and three ages greater than or equal to 18; or
2. the median is 18 because, when the data values are arranged in numerical order, there are three values below this median, or middle number, and three values above it.

Recently, the car wash hired a new employee whose age is 21. The data now include eight ages, an even number, so there is no middle value. What is now the median age?

STEP 1. Arrange the ages in numerical order: 17, 17, 18, 18, 19, 20, 21, 46

STEP 2. There is no single middle number. Find the *two* middle numbers: 17, 17, 18, 18, 19, 20, 21, 46

↑ ↑

STEP 3. Find the mean (arithmetic average) of the two middle numbers: $\frac{18 + 19}{2} = 18\frac{1}{2}$

The median is now $18\frac{1}{2}$. There are four ages less than this center value of $18\frac{1}{2}$ and four ages greater than $18\frac{1}{2}$.

Procedure

To find the median of a set of *n* numbers:

1. Arrange the numbers in numerical order.
2. If *n* is odd, find the middle number. This number is the median.
3. If *n* is even, find the mean (arithmetic average) of the *two* middle numbers. This average is the median.

The Mode

The **mode** is the data value that appears most often in a given set of data. It is usually best to arrange the data in numerical order before finding the mode.

Let us consider some examples of finding the mode:

1. The ages of employees in a car wash are 17, 17, 17, 18, 19, 20, 46. The mode, which is the number appearing most often, is 17.
2. The number of hours each of six students spent reading a book are 6, 6, 8, 11, 14, 21. The mode, or number appearing most frequently, is 6. In this case, however, the mode is not a useful measure of central tendency. A better indication is given by the mean or the median.
3. The number of photographs printed from each of Renee's last six rolls of film are 8, 8, 9, 11, 11, and 12. Since 8 appears twice and 11 appears twice, we say that there are two modes: 8 and 11. We do not take the average of these two numbers since the mode tells us where most of the scores appear. We simply report both numbers. When *two* modes appear within a set of data, we say that the data are **bimodal**.
4. The number of people living in each house on Meryl's street are 2, 2, 3, 3, 4, 5, 5, 6, 8. These data have *three* modes: 2, 3, and 5.
5. Ralph's test scores in science are 72, 80, 86, 93, and 94. Here, every number appears the same number of times, once. Since *no* number appears more often than the others, we define such data as having no mode.

Procedure

To find the mode for a set of data, find the number or numbers that occur most often.

1. If one number appears most often in the data, that number is the mode.
2. If two or more numbers appear more often than all other data values, and these numbers appear with the same frequency, then each of these numbers is a mode.
3. If each number in a set of data occurs with the same frequency, there is no mode.

KEEP IN MIND Three measures of central tendency are:

1. The mean, or mean average, found by adding n data values and then dividing the sum by n.
2. The median, or middle score, found when the data are arranged in numerical order.
3. The mode, or the value that appears most often.

A graphing calculator can be used to arrange the data in numerical order and to find the mean and the median. The calculator solution in the following example lists the keystrokes needed to do this.

EXAMPLE 1

The weights, in pounds, of five players on the basketball team are 195, 168, 174, 182, and 181. Find the average weight of a player on this team.

Solution The word *average*, by itself, indicates the *mean*. Therefore:

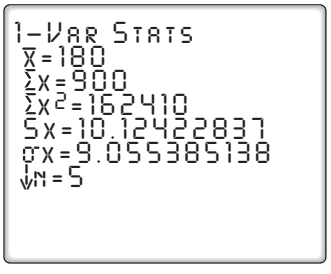
(1) Add the five weights: $195 + 168 + 174 + 182 + 181 = 900$.

(2) Divide the sum by 5, the number of players: $900 \div 5 = 180$.

Calculator Solution Enter the data into list L_1. Then use 1-Var Stats from the STAT CALC menu to display information about this set of data.

ENTER: STAT ▶ ENTER ENTER

DISPLAY:



The first value given is $\bar{x}$, the mean.

Answer 180 pounds ■

The second value given is $\Sigma x = 900$. The symbol Σ represents a sum and $\Sigma x = 900$ can be read as "The sum of the values of x is 900." The list shows other values related to this set of data. The arrow at the bottom of the display indicates that more entries follow what appears on the screen. These can be displayed by pressing the down arrow. One of these is the median (Med = 181). The display also shows that there are 5 data values ($n = 5$). Others we will use in later sections in this chapter and in more advanced courses.

EXAMPLE 2

Renaldo has marks of 75, 82, and 90 on three mathematics tests. What mark must he obtain on the next test to have an average of exactly 85 for the four math tests?

Solution The word *average*, by itself, indicates the *mean*.

Let x = Renaldo's mark on the fourth test.

The sum of the four test marks divided by 4 is 85.

$$\frac{75 + 82 + 90 + x}{4} = 85$$

$$\frac{247 + x}{4} = 85$$

$$247 + x = 340$$

$$x = 93$$

Check

$$\frac{75 + 82 + 90 + 93}{4} \stackrel{?}{=} 85$$

$$\frac{340}{4} \stackrel{?}{=} 85$$

$$85 \stackrel{?}{=} 85 \checkmark$$

Answer Renaldo must obtain a mark of 93 on his fourth math test. ■

EXAMPLE 3

Find the median for each distribution.

a. 4, 2, 5, 5, 1 **b.** 9, 8, 8, 7, 4, 3, 3, 2, 0, 0

Solution **a.** Arrange the data in numerical order: 1, 2, 4, 5, 5

The median is the middle value: 1, 2, 4, 5, 5

Answer median = 4

b. Since there is an even number of values, there are two middle values. Find the mean (average) of these two middle values:

$$9, 8, 8, 7, \underset{\uparrow}{4}, \underset{\uparrow}{3}, 3, 2, 0, 0$$

$$\frac{4 + 3}{2} = \frac{7}{2} = 3\frac{1}{2}$$

Answer median = $3\frac{1}{2}$ or 3.5 ■

EXAMPLE 4

Find the mode for each distribution.

a. 2, 9, 3, 7, 3 **b.** 3, 4, 5, 4, 3, 7, 2 **c.** 1, 2, 3, 4, 5, 6, 7

Solution **a.** Arrange the data in numerical order: 2, 3, 3, 7, 9.

The mode, or most frequent value, is 3.

b. Arrange the data in numerical order: 2, 3, 3, 4, 4, 5, 7. Both 3 and 4 appear twice. There are two modes.

c. Every value occurs the same number of times in the data set 1, 2, 3, 4, 5, 6, 7. There is no mode.

Answers **a.** The mode is 3. **b.** The modes are 3 and 4. **c.** There is no mode. ■

Linear Transformations of Data

Multiplying each data value by the same constant or adding the same constant to each data value is an example of a **linear transformation of a set of data**.

Let us start by examining additive transformations. For instance, consider the data 2, 2, 3, 4, 5. If 10 is added to each data value, the data set becomes:

$$12, 12, 13, 14, 15$$

Notice that every measure of central tendency has been shifted to the right by 10 units:

old mean $= \frac{2+2+3+4+5}{5} = 3.2$	new mean $= \frac{12+12+13+14+15}{5} = 13.2$
old median $= 3$	new median $= 13$
old mode $= 2$	new mode $= 12$

In fact, this result is valid for any additive transformation of a data set. In general:

► **If $\bar{x}$, d, and o are the mean, median, and mode of a set of data and the constant c is added to each data value, then $\bar{x} + c$, $d + c$, and $o + c$ are the mean, median, and mode of the transformed data.**

It can be also shown that a similar result holds for multiplicative transformations, that is:

► **If $\bar{x}$, d, and o are the mean, median, and mode of a set of data and each data value is multiplied by the *nonzero* constant c, then $c\bar{x}$, cd, and co are the mean, median, and mode of the transformed data.**

EXAMPLE 5

In Ms. Huan's Algebra class, the average score on the most recent quiz was 65. Being in a generous mood, Ms. Huan decided to curve the quiz by adding 10 points to each quiz score. What will be the new average score for the class?

Answer $65 + 10 = 75$ points

EXERCISES

Writing About Mathematics

1. On her first two math tests, Rene received grades of 67 and 79. Her mean (average) grade for these two tests was 73. On her third test she received a grade of 91. Rene found the mean of 73 and 91 and said that her mean for the three tests was 82. Do you agree with Rene? Explain why or why not.

2. Carlos said that when there are n numbers in a set of data and n is an odd number, the median is the $\frac{n+1}{2}$th number when the data are arranged in order. Do you agree with Carlos? Explain why or why not.

Developing Skills

3. For each set of data, find the mean.

a. 7, 3, 5, 11, 9

b. 22, 38, 18, 14, 22, 30

c. $5\frac{1}{2}, 2\frac{3}{4}, 7\frac{1}{2}, 5\frac{3}{4}, 4\frac{1}{2}$

d. 1.00, 0.01, 1.10, 0.12, 1.00, 1.03

4. Find the median for each set of data.

a. 1, 2, 5, 3, 4

b. 2, 9, 2, 9, 7

c. 3, 8, 12, 7, 1, 0, 4

d. 80, 83, 97, 79, 25

e. 3.2, 8.7, 1.4

f. 2.00, 0.20, 2.20, 0.02, 2.02

g. 21, 24, 23, 22, 20, 24, 23, 21, 22, 23

h. 5, 7, 9, 3, 8, 7, 5, 6

5. What is the median for the digits $1, 2, 3, \ldots, 9$?

6. What is the median for the counting numbers from 1 through 100?

7. Find the mode for each distribution.

a. 2, 2, 3, 4, 8

b. 2, 2, 3, 8, 8

c. 2, 2, 8, 8, 8

d. 2, 3, 4, 7, 8

e. 2, 2, 3, 8, 8, 9, 9

f. 1, 2, 1, 2, 1, 2, 1

g. 1, 2, 3, 2, 1, 2, 3, 2, 1

h. 3, 19, 21, 75, 0, 6

i. 3, 2, 7, 6, 2, 7, 3, 1, 4, 2, 7, 5

j. 19, 21, 18, 23, 19, 22, 18, 19, 20

8. A set of data consists of six numbers: 7, 8, 8, 9, 9, and x. Find the mode for these six numbers when:

a. $x = 9$ **b.** $x = 8$ **c.** $x = 7$ **d.** $x = 6$

9. A set of data consists of the values 2, 4, 5, x, 5, 4. Find a possible value of x such that:

a. there is no mode because all scores appear an equal number of times

b. there is only one mode

c. there are two modes

10. For the set of data 5, 5, 6, 7, 7, which statement is true?

(1) mean = mode (3) mean = median

(2) median = mode (4) mean < median

11. For the set of data 8, 8, 9, 10, 15, which statement is true?

(1) mean < median (3) median < mode

(2) mean > mode (4) mean = median

12. When the data consists of 3, 4, 5, 4, 3, 4, 5, which statement is true?
(1) mean > median (3) median < mode
(2) mean > mode (4) mean = median

13. For which set of data is there no mode?
(1) 2, 1, 3, 1, 2 (3) 1, 2, 4, 3, 5
(2) 1, 2, 3, 3, 3 (4) 2, 2, 3, 3, 3

14. For which set of data is there more than one mode?
(1) 8, 7, 7, 8, 7 (3) 8, 7, 5, 7, 6, 5
(2) 8, 7, 4, 5, 6 (4) 1, 2, 2, 3, 3, 3

15. For which set of data does the median equal the mode?
(1) 3, 3, 4, 5, 6 (3) 3, 3, 4
(2) 3, 3, 4, 5 (4) 3, 4

16. For which set of data will the mean, median, and mode all be equal?
(1) 1, 2, 5, 5, 7 (3) 1, 1, 1, 2, 5
(2) 1, 2, 5, 5, 8, 9 (4) 1, 1, 2

17. The median of the following data is 11:

2, 5, 9, 11, 40, 3, 4, 5, 10, 45, 32, 40, 67, 7, 11, 9, 20, 34, 5, 1, 8, 15, 16, 19, 39

a. If 4 is subtracted from each data value, what is the median of the transformed data set?

b. If the largest data value is doubled and the smallest data value is halved, what is the median of the new data set?

18. The mean of the following data is 37.625:

3, 0, 1, 7, 8, 11, 31, 15, 99, 98, 92, 81, 85, 87, 55, 54, 34, 27, 26, 21, 14, 17, 19, 18

If each data value is multiplied by 2 and increased by 5, what is the mean of the transformed data set?

19. Three consecutive integers can be represented by x, $x + 1$, and $x + 2$. The average of these consecutive integers is 32. What are the three integers?

20. Three consecutive even integers can be represented by x, $x + 2$, and $x + 4$. The average of these consecutive even integers is 20. Find the integers.

21. The mean of three numbers is 31. The second is 1 more than twice the first. The third is 4 less than 3 times the first. Find the numbers.

Applying Skills

22. Sid received grades of 92, 84, and 70 on three tests. Find his test average.

23. Sarah's grades were 80 on each of two of her tests and 90 on each of three other tests. Find her test average.

24. Louise received a grade of x on each of two of her tests and of y on each of three other tests. Represent her average for all the tests in terms of x and y.

25. Andy has grades of 84, 65, and 76 on three social studies tests. What grade must he obtain on the next test to have an average of exactly 80 for the four tests?

26. Rosemary has grades of 90, 90, 92, and 78 on four English tests. What grade must she obtain on the next test so that her average for the five tests will be 90?

27. The first three test scores are shown below for each of four students. A fourth test will be given and averages taken for all four tests. Each student hopes to maintain an average of 85. Find the score needed by each student on the fourth test to have an 85 average, or explain why such an average is not possible.

a. Pat: 78, 80, 100

b. Bernice: 79, 80, 81

c. Helen: 90, 92, 95

d. Al: 65, 80, 80

28. The average weight of Sue, Pam, and Nancy is 55 kilograms.

a. What is the total weight of the three girls?

b. Agnes weighs 60 kilograms. What is the average weight of the four girls: Sue, Pam, Nancy, and Agnes?

29. For the first 6 days of a week, the average rainfall in Chicago was 1.2 inches. On the last day of the week, 1.9 inches of rain fell. What was the average rainfall for the week?

30. If the heights, in centimeters, of a group of students are 180, 180, 173, 170, and 167, what is the mean height of these students?

31. What is the median age of a family whose members are 42, 38, 14, 13, 10, and 8 years old?

32. What is the median age of a class in which 14 students are 14 years old and 16 students are 15 years old?

33. In a charity collection, ten people gave amounts of \$1, \$2, \$1, \$1, \$3, \$1, \$2, \$1, \$1, and \$1.50. What was the median donation?

34. The test scores for an examination were 62, 67, 67, 70, 90, 93, and 98. What is the median test score?

35. The weekly salaries of six employees in a small firm are \$440, \$445, \$445, \$450, \$450, and \$620.

a. For these six salaries, find: (1) the mean (2) the median (3) the mode

b. If negotiations for new salaries are in session and you represent management, which measure of central tendency will you use as the average salary? Explain your answer.

c. If negotiations are in session and you represent the labor union, which measure of central tendency will you use as an average salary? Explain your answer.

36. In a certain school district, bus service is provided for students living at least $1\frac{1}{2}$ miles from school. The distances, rounded to the nearest half mile, from school to home for ten students are $0, \frac{1}{2}, \frac{1}{2}, 1, 1, 1, 1, 1\frac{1}{2}, 3\frac{1}{2}$, and 10 miles.

a. For these data, find: (1) the mean (2) the median (3) the mode

b. How many of the ten students are entitled to bus service?

c. Explain why the mean is not a good measure of central tendency to describe the average distance between home and school for these students.

37. Last month, a carpenter used 12 boxes of nails each of which contained nails of only one size. The sizes marked on the boxes were:

$$\tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, \tfrac{3}{4}\text{ in.}, 1\text{ in.}, 1\text{ in.}, 2\text{ in.}, 2\text{ in.}$$

a. For these data, find: (1) the mean (2) the median (3) the mode

b. Describe the average-size nail used by the carpenter, using at least one of these measures of central tendency. Explain your answer.

Hands-On Activity

Find the mean, the median, and the mode for the data that you collected in the Hands-On Activity for Section 16-1. It may be necessary to go back to your original data to do this.

16-5 MEASURES OF CENTRAL TENDENCY AND GROUPED DATA

Intervals of Length 1

In a statistical study, when the range is small, we can use intervals of length 1 to group the data. For example, each member of a class of 25 students reported the number of books he or she read during the first half of the school year. The data are as follows:

5, 3, 5, 3, 1, 8, 2, 4, 2, 6, 3, 8, 8, 5, 3, 4, 5, 8, 5, 3, 3, 5, 6, 2, 3

These data, for which the values range from 1 to 8, can be organized into a table such as the one shown at the right, with each value representing an interval.

Since 25 students were included in this study, the total frequency, N, is 25. We can use this table, with intervals of length 1, to find the mode, median, and mean for these data.

Interval	Frequency
8	4
7	0
6	2
5	6
4	2
3	7
2	3
1	1
	$N = 25$

Mode of a Set of Grouped Data

Since the greatest frequency, 7, appears for interval 3, the mode for the data is 3. In general:

► **For a set of grouped data, the mode is the value of the interval that contains the greatest frequency.**

Median of a Set of Grouped Data

We have learned that the median for a set of data in numerical order is the middle value.

For these 25 numbers, there are 12 numbers greater than or equal to the median, and 12 numbers less than or equal to the median. Therefore, when the numbers are written in numerical order, the median is the 13th number from either end.

1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 8, 8, 8, 8
↑ The median is 4.

When the data are grouped in the table shown earlier, a simple counting procedure can be used to find the median, the 13th number. When we add the frequencies of the first four intervals, starting at the top, we find that these intervals include data for:

$$4 + 0 + 2 + 6 = 12 \text{ students}$$

Therefore, the next lower interval (with frequency greater than 0) must include the median, the value for the 13th student. This is the interval for the data value 4.

When we add the frequencies of the first three intervals, starting at the bottom, we find that these intervals include data for:

$$1 + 3 + 7 = 11 \text{ students}$$

The next higher interval contains two scores, one for the 12th student and one that is the median, or the value for the 13th student. Again this is the interval for the data value 4.

In general:

► **For a set of grouped data, the median is the value of the interval that contains the middle data value.**

Mean of a Set of Grouped Data

By adding the four 8's in the ungrouped data, we see that four students, reading eight books each, have read $8 + 8 + 8 + 8$ or 32 books. We can arrive at this same number by using the grouped intervals in the table: we multiply the four 8's by the frequency 4. Thus, $(4)(8) = 32$. Applying this multiplication shortcut to each row of the table, we obtain the third column of the following table:

Interval	Frequency	(Interval) × (Frequency)
8	4	$8 \times 4 = 32$
7	0	$7 \times 0 = 0$
6	2	$6 \times 2 = 12$
5	6	$5 \times 6 = 30$
4	2	$4 \times 2 = 8$
3	7	$3 \times 7 = 21$
2	3	$2 \times 3 = 6$
1	1	$1 \times 1 = 1$
	$N = 25$	Total = 110

The total (110) represents the sum of all 25 pieces of data. We can check this by adding the 25 scores in the unorganized data.

Finally, to find the mean, we divide the total number, 110, by the number of items, 25. Thus, the mean for the data is: $110 \div 25 = 4.4$.

Procedure

To find the mean for *N* values in a table of grouped data when the length of each interval is 1:

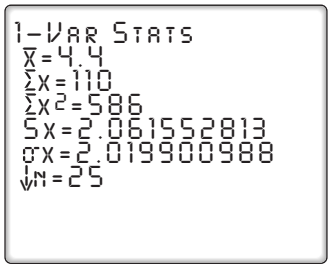
1. For each interval, multiply the interval value by its corresponding frequency.
2. Find the sum of these products.
3. Divide this sum by the total frequency, *N*.

Calculator Solution for Grouped Data

The calculator can be used to find the mean and median for the grouped data shown above. Enter the number of books read by each student into L_1 and the frequency for each number of books into L_2. Then use the 1-Var Stats from the STAT CALC menu to display information about the data.

ENTER: STAT ▶ ENTER 2nd L1 , 2nd L2 ENTER

DISPLAY:



The display shows that the mean, $\bar{x}$, is 4.4, the sum of the number of books read is 110, and the number of students, the total frequency, N, is 25. Use the down arrow to display the median, Med = 4.

Intervals Other Than Length 1

There are specific mathematical procedures to find the mean, median, and mode for grouped data with intervals other than length 1, but we will not study them at this time. Instead, we will simply identify the intervals that contain some of these measures of central tendency. For example, a small industrial plant surveyed 50 workers to find the number of miles each person commuted to work. The commuting distances were reported, to the nearest mile, as follows:

0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7, 7, 9,
10, 10, 10, 10, 10, 10, 10, 10, 12, 12, 14, 15, 17, 17,
18, 22, 23, 25, 28, 30, 32, 32, 33, 34, 34, 36, 37, 37, 52

These data are organized into a table with intervals of length 10, as follows:

Interval (commuting distance)	Frequency (number of workers)
50–59	1
40–49	0
30–39	9
20–29	4
10–19	15
0–9	21
	$N = 50$

Modal Interval

In the table, interval 0–9 contains the greatest frequency, 21. We say that interval 0–9 is the **group mode**, or **modal interval**, because this group of numbers has the greatest frequency. The modal interval is *not* the same as the mode. The modal interval is a group of numbers; the mode is usually a single number. For this example, the original data (before being placed into the table) show that the number appearing most often is 10. Hence, the mode is 10. The modal interval, which is 0–9, tells us that, of the six intervals in the table, the most frequently occurring commuting distance is 0 to 9 miles.

Both the mode and the modal interval depend on the concept of greatest frequency. For the mode, we look for a single number that has the greatest frequency. For the modal interval, we look for the interval that has the greatest frequency.

Interval Containing the Median

To find the interval containing the median, we follow the procedure described earlier in this section. For 50 numbers, the median, or middle number, will be at a point where 25 numbers are at or above the median and 25 are at or below it.

Count the frequencies in the table from the uppermost interval and move downward. We add $1 + 0 + 9 + 4 = 14$. Since there are 15 numbers in the next lower interval, and $14 + 15 = 29$, we see that the 25th number will be reached somewhere in that interval, 10–19.

Count from the bottom interval and move up. We have 21 numbers in the first interval. Since there are 15 numbers in the next higher interval, and $21 + 15 = 36$, we see that 25th number will be reached somewhere in that interval, 10–19. This is the same result that we obtained when we moved downward. The interval containing the median for this grouping is 10–19.

In this course, we will not deal with problems in which the median is not found in any interval.

Interval Containing the Mean

When data are grouped using intervals of length other than 1, there is no simple procedure to identify the interval containing the mean. However, the mean can be *approximated* by assuming that the data are equally distributed throughout each interval. The mean is then found by using the midpoint of each interval as the value of each entry in the interval. This problem is studied in higher-level courses.

EXAMPLE 1

In the table, the data indicate the heights, in inches, of 17 basketball players. For these data find:
a. the mode **b.** the median **c.** the mean

Height (inches)	Frequency (number)
77	2
76	0
75	5
74	3
73	4
72	2
71	1

Solution

a. The greatest frequency, 5, occurs for the height of 75 inches. The mode, or height appearing most often, is 75.

b. For 17 players, the median is the 9th number, so there are 8 heights greater than or equal to the median and 8 heights less than or equal to the median. Counting the frequencies going down, we have $2 + 0 + 5 = 7$. Since the frequency of the next interval is 3, the 8th, 9th, and 10th heights are in this interval, 74.

Counting the frequencies going up, we have $1 + 2 + 4 = 7$. Again, the frequency of the next interval is 3, and the 8th, 9th, and 10th heights are in this interval. The 9th height, the median, is 74.

c. (1) Multiply each height by its corresponding frequency:

$77 \times 2 = 154$ $\quad 76 \times 0 = 0$ $\quad 75 \times 5 = 375$ $\quad 74 \times 3 = 222$

$73 \times 4 = 292$ $\quad 72 \times 2 = 144$ $\quad 71 \times 1 = 71$

(2) Find the total of these products:

$$154 + 0 + 375 + 222 + 292 + 144 + 71 = 1{,}258$$

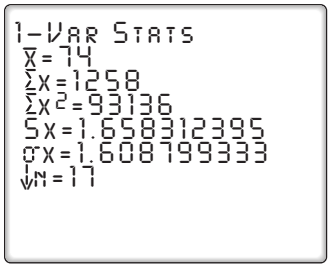
(3) Divide this total, 1,258, by the total frequency, 17 to obtain the mean:

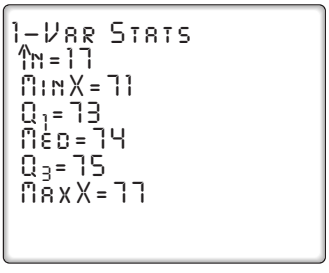
$$1258 \div 17 = 74$$

Calculator Solution Clear any previous data that may be stored in L_1 and L_2. Enter the heights of the players into L_1 and the frequencies into L_2. Then use 1-Var Stats from the STAT CALC menu to display information about the data. The screen will show the mean, $\bar{x}$. Press the down arrow key to display the median.

ENTER: STAT ▶ ENTER 2nd L1 , 2nd L2 ENTER

DISPLAY:





Answers **a.** mode = 75 **b.** median = 74 **c.** mean = 74

EXERCISES

Writing About Mathematics

1. The median for a set of 50 data values is the average of the 25th and 26th data values when the data is in numerical order. What must be true if the median is equal to one of the data values? Explain your answer.

2. What must be true about a set of data if the median is *not* one of the data values? Explain your answer.

Developing Skills

In 3–5, the data are grouped in each table in intervals of length 1.
Find: **a.** the total frequency **b.** the mean **c.** the median **d.** the mode

3.

Interval	Frequency
10	1
9	2
8	3
7	3
6	4
5	3

4.

Interval	Frequency
15	3
16	2
17	4
18	1
19	5
20	6

5.

Interval	Frequency
25	4
24	0
23	3
22	2
21	4
20	5
19	2

In 6–8, the data are grouped in each table in intervals other than length 1. Find: **a.** the total frequency **b.** the interval that contains the median **c.** the modal interval

6.

Interval	Frequency
55–64	3
45–54	8
35–44	7
25–34	6
15–24	2

7.

Interval	Frequency
4–9	12
10–15	13
16–21	9
22–27	12
28–33	15
34–39	10

8.

Interval	Frequency
126–150	4
101–125	6
76–100	6
51–75	3
26–50	7
1–25	2

Applying Skills

9. On a test consisting of 20 questions, 15 students received the following scores:

17, 14, 16, 18, 17, 19, 15, 15, 16, 13, 17, 12, 18, 16, 17

a. Make a frequency table for these students listing scores from 12 to 20.
b. Find the median score.
c. Find the mode.
d. Find the mean.

10. A questionnaire was distributed to 100 people. The table shows the time taken, in minutes, to complete the questionnaire.

a. For this set of data, find: (1) the mean (2) the median (3) the mode

b. How are the three measures found in part **a** related for these data?

Interval	Frequency
6	12
5	20
4	36
3	20
2	12

11. A storeowner kept a tally of the sizes of suits purchased in the store, as shown in the table.

a. For this set of data, find:
(1) the total frequency (2) the mean
(3) the median (4) the mode

b. Which measure of central tendency should the storeowner use to describe the average suit sold?

Size of Suit (interval)	Number Sold (frequency)
48	1
46	1
44	3
42	5
40	3
38	8
36	2
34	2

12. Test scores for a class of 20 students are as follows:

93, 84, 97, 98, 100, 78, 86, 100, 85, 92, 72, 55, 91, 90, 75, 94, 83, 60, 81, 95

a. Organize the data in a table using 51–60 as the smallest interval.

b. Find the modal interval.

c. Find the interval that contains the median.

13. The following data consist of the weights, in pounds, of 35 adults:

176, 154, 161, 125, 138, 142, 108, 115, 187, 158, 168, 162
135, 120, 134, 190, 195, 117, 142, 133, 138, 151, 150, 168
172, 115, 148, 112, 123, 137, 186, 171, 166, 166, 179

a. Organize the data in a table, using 100–119 as the smallest interval.

b. Construct a frequency histogram based on the grouped data.

c. In what interval is the median for these grouped data?

d. What is the modal interval?

16-6 QUARTILES, PERCENTILES, AND CUMULATIVE FREQUENCY

Quartiles

When the values in a set of data are listed in numerical order, the median separates the values into two equal parts. The numbers that separate the set into four equal parts are called **quartiles**.

To find the quartile values, we first divide the set of data into two equal parts and then divide each of these parts into two equal parts.

The heights, in inches, of 20 students are shown in the following list. The median, which is the average of the 10th and 11th data values, is shown here enclosed in a box.

Lower half: 53, 60, 61, 63, 64, 65, 65, 65, 65, 66,
Upper half: 66, 67, 67, 68, 69, 70, 70, 71, 71, 73

$\boxed{66}$

Median

Ten heights are listed in the lower half, 53–66. The middle value for these 10 heights is the average of the 5th and 6th values from the lower end, or 64.5. This value separates the lower half into two equal parts.

Ten heights are also listed in the upper half, 66–73. The middle value for these 10 heights is the average of the 5th and 6th values from the upper end, or 69.5. This value separates the upper half into two equal parts.

The 20 data values are now separated into four equal parts, or quarters.

53, 60, 61, 63, 64, | 65, 65, 65, 65, 66, | 66, 67, 67, 68, 69, | 70, 70, 71, 71, 73

$\boxed{64.5}$ $\boxed{66}$ $\boxed{69.5}$

First quartile — Median / Second quartile — Third quartile

The numbers that separate the data into four equal parts are the quartiles. For this set of data:

1. Since one quarter of the heights are less than or equal to 64.5 inches, 64.5 is the **lower quartile**, or **first quartile**.
2. Since two quarters of the heights are less than or equal to 66 inches, 66 is the **second quartile**. The second quartile is always the same as the median.
3. Since three quarters of the heights are less than or equal to 69.5 inches, 69.5 is the **upper quartile**, or **third quartile**.

Note: The quartiles are sometimes denoted Q_1, Q_2, and Q_3.

Procedure

To find the quartile values for a set of data:

1. Arrange the data in ascending order from left to right.
2. Find the median for the set of data. The median is the second quartile value.
3. Find the middle value for the lower half of the data. This number is the first, or lower, quartile value.
4. Find the middle value for the upper half of the data. This number is the third, or upper, quartile value.

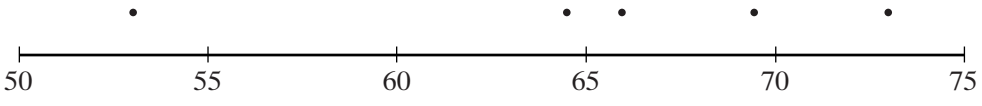
Note that when finding the first quartile, use all of the data values less than or equal to the median, but do not include the median in the calculation. Similarly, when finding the third quartile, use all of the data values greater than or equal to the median, but do not include the median in the calculation.

Constructing a Box-and-Whisker Plot

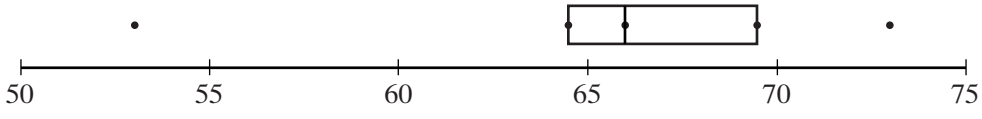
A **box-and-whisker plot** is a diagram that uses the quartile values, together with the maximum and minimum values, to display information about a set of data. To draw a box-and-whisker plot, we use the following steps.

STEP 1. Draw a scale with numbers from the minimum to the maximum value of a set of data. For example, for the set of heights of the 20 students, the scale should include the numbers from 53 to 73.

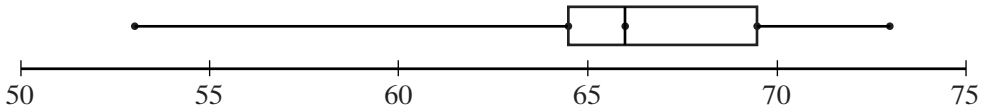
STEP 2. Above the scale, place dots to represent the five numbers that are the **statistical summary** for this set of data: the minimum value, the first quartile, the median, the third quartile, and the maximum value. For the heights of the 20 students, these numbers are 53, 64.5, 66, 69.5 and 73.



STEP 3. Draw a box between the dots that represent the lower and upper quartiles, and a vertical line in the box through the point that represents the median.

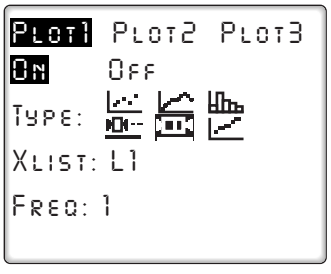


STEP 4. Add the whiskers by drawing a line segment joining the dots that represent the minimum data value and the lower quartile, and a second line segment joining the dots that represent the maximum data value and the upper quartile.



The box indicates the ranges of the middle half of the set of data. The long whisker at the left shows us that the data are more scattered at the lower than at the higher end.

A graphing calculator can display a box-and-whisker plot. Enter the data in L_1, then go to the STAT PLOT menu to select the type of graph to draw.

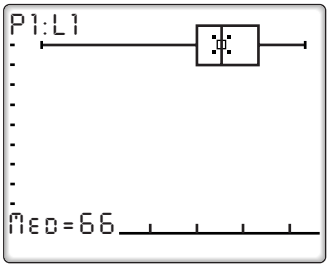


ENTER: [2nd] [STAT PLOT] [1] [ENTER]

[▼] [▶] [▶] [▶] [▶] [ENTER] [▼]

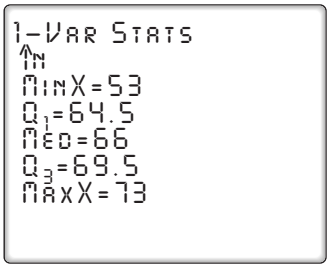
[2nd] [L1] [▼] [ALPHA] [1]

Now display the box-and-whisker plot by entering [ZOOM] [9].



We can press [TRACE] and the right and left arrow keys to display the minimum value, first quartile, median, third quartile, and maximum value.

The five statistical summary can also be displayed in 1-Var Stats. Scroll down to the last five values.



ENTER: [STAT] [▶] [ENTER] [ENTER]

EXAMPLE 1

Find the five statistical summary for the following set of data:

8, 5, 12, 9, 6, 2, 14, 7, 10, 17, 11, 8, 14, 5

Solution (1) Arrange the data in numerical order:

$$2, 5, 5, 6, 7, 8, 8, 9, 10, 11, 12, 14, 14, 17$$

We can see that 2 is the minimum value and 17 is the maximum value.

(2) Find the median. Since there are 14 data values in the set, the median is the average of the 7th and 8th values.

$$\text{Median} = \frac{8+9}{2} = 8.5$$

Therefore, 8.5 is the second quartile.

(3) Find the first quartile. There are seven values less than 8.5. The middle value is the 4th value from the lower end of the set of data, 6. Therefore, 6 is the first, or lower, quartile.

(4) Find the third quartile. There are seven values greater than 8.5. The middle value is the 4th value from the upper end of the set of data, 12. Therefore, 12 is the third, or upper, quartile.

Answer The minimum is 2, first quartile is 6, the second quartile is 8.5, the third quartile is 12, and the maximum is 17.

Note: The quartiles 6, 8.5, and 12 separate the data values into four equal parts even though the original number of data values, 14, is not divisible by 4:

$$\overline{2, 5, 5,}\ \boxed{6,}\ \overline{7, 8, 8,}\ \overline{9, 10, 11,}\ \boxed{12,}\ \overline{14, 14, 17}$$

$$\uparrow$$

$$\boxed{8.5}$$

The first and third quartile values, 6 and 12, are data values. If we think of each of these as a half data value in the groups that they separate, each group contains $3\frac{1}{2}$ data values, which is 25% of the total.

Percentiles

A **percentile** is a number that tells us what percent of the total number of data values lies at or below a given measure.

Let us consider again the set of data values representing the heights of 20 students. What is the percentile rank of 65? To find out, we separate the data into the values that are less than or equal to 65 and those that are greater than or equal to 65, so that the four 65's in the set are divided equally between the two groups:

$$\boxed{53, 60, 61, 63, 64, 65, 65,}\quad \boxed{65, 65, 66, 66, 67, 67, 68, 69, 70, 70, 71, 71, 73}$$

Half of 4, or 2, of the 65's are in the lower group and half are in the upper group.

Since there are seven data values in the lower group, we find what percent 7 is of 20, the total number of values:

$$\frac{7}{20} = 0.35 = 35\%$$

Therefore, 65 is at the 35th percentile.

To find the percentile rank of 69, we separate the data into the values that are less than or equal to 69 and those that are greater or equal to 69:

53, 60, 61, 63, 64, 65, 65, 65, 65, 66, 66, 67, 67, 68, 69, | 70, 70, 71, 71, 73

Because 69 occurs only once, we will include it as half of a data value in the lower group and half of a data value in the upper group. Therefore, there are $14\frac{1}{2}$ or 14.5 data values in the lower group.

$$\frac{14.5}{20} = 0.725 = 72.5\%$$

Because percentiles are usually not written using fractions, we say that 69 is at the 73rd percentile.

EXAMPLE 2

Find the percentile rank of 87 in the following set of 30 marks:

56, 65, 65, 67, 72, 73, 75, 77, 77, 78, 78, 78, 80, 80, 80,
82, 83, 85, 85, 85, 86, 87, 87, 87, 88, 90, 92, 93, 95, 98

Solution (1) Find the sum of the number of marks less than 87 and half of the number of 87's:

Number of marks less than 87	= 21
Half of the number of 87's (0.5×3)	= 1.5
	22.5

(2) Divide the sum by the total number of marks:

$$\frac{22.5}{30} = 0.75$$

(3) Change the decimal value to a percent: $0.75 = 75\%$.

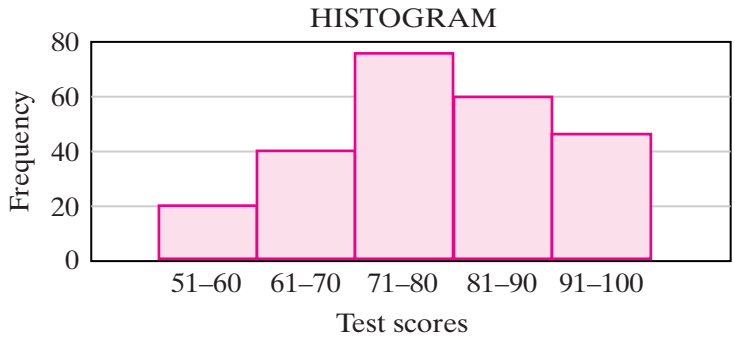
Answer: A mark of 87 is at the 75th percentile.

Note: 87 is also the upper quartile mark.

Cumulative Frequency

In a school, a final examination was given to all 240 students taking biology. The test grades of these students were then grouped into a table. At the same time, a histogram of the results was constructed, as shown below.

Interval (test scores)	Frequency (number)
91–100	45
81–90	60
71–80	75
61–70	40
51–60	20



From the table and the histogram, we can see that 20 students scored in the interval 51–60, 40 students scored in the interval 61–70, and so forth. We can use these data to construct a new type of histogram that will answer the question, "How many students scored *below* a certain grade?"

By answering the following questions, we will gather some information before constructing the new histogram:

1. How many students scored 60 or less on the test?

From the lowest interval, 51–60, we know that 20 students scored 60 or less.

2. How many students scored 70 or less on the test?

By adding the frequencies for the two lowest intervals, 51–60 and 61–70, we see that 20 + 40, or 60, students scored 70 or less.

3. How many students scored 80 or less on the test?

By adding the frequencies for the three lowest intervals, 51–60, 61–70, and 71–80, we see that 20 + 40 + 75, or 135, students scored 80 or less.

4. How many students scored 90 or less on the test?

Here, we add the frequencies in the four lowest intervals. Thus, 20 + 40 + 75 + 60, or 195, students scored 90 or less.

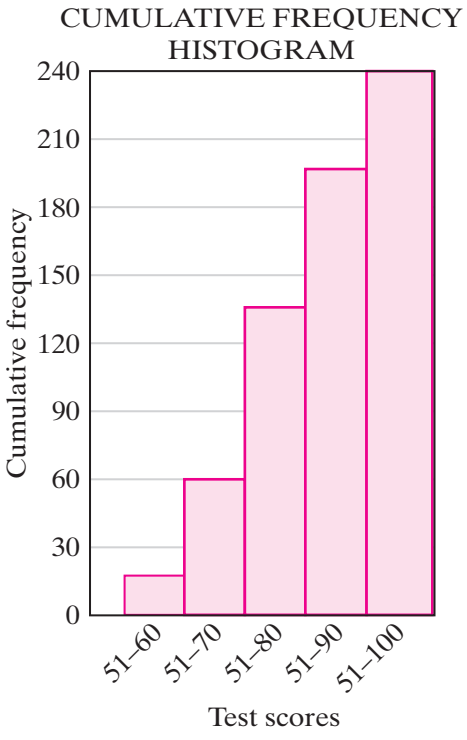
5. How many students scored 100 or less on the test?

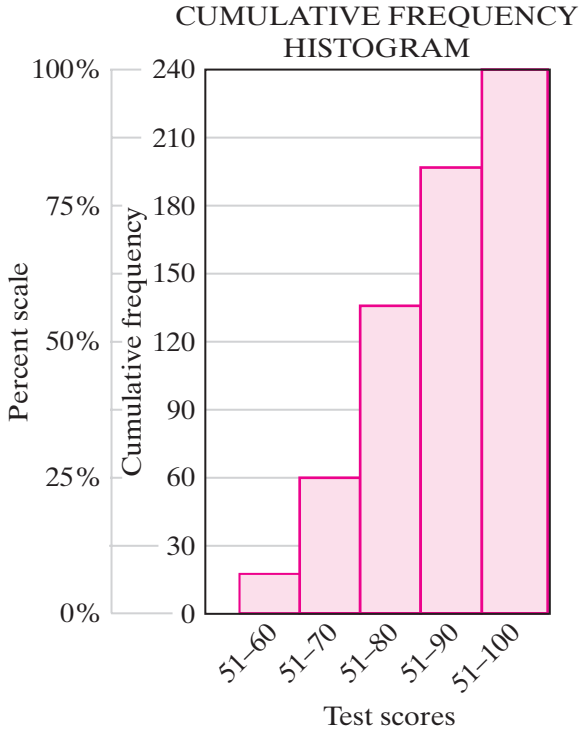
By adding the five lowest frequencies, 20 + 40 + 75 + 60 + 45, we see that 240 students scored 100 or less. This result makes sense because 240 students took the test and all of them scored 100 or less.

Constructing a Cumulative Frequency Histogram

The answers to the five questions we have just asked were found by adding, or *accumulating*, the frequencies for the intervals in the grouped data to find the **cumulative frequency**. The accumulation of data starts with the lowest interval of data values, in this case, the lowest test scores. The histogram that displays these accumulated figures is called a **cumulative frequency histogram**.

Interval (test scores)	Frequency (number)	Cumulative Frequency
91–100	45	240
81–90	60	195
71–80	75	135
61–70	40	60
51–60	20	20





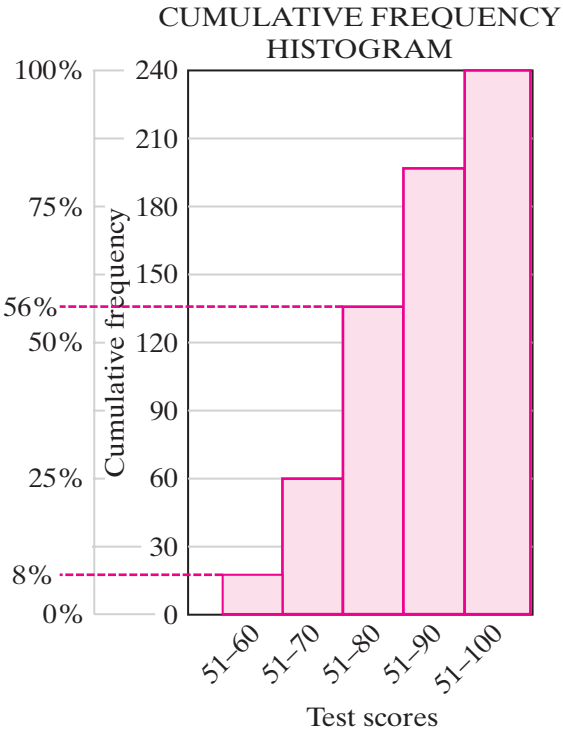
To find the cumulative frequency for each interval, we add the frequency for that interval to the frequencies for the intervals with lower values. To draw a cumulative frequency histogram, we use the cumulative frequencies to determine the heights of the bars.

For our example of the 240 biology students and their scores, the frequency scale for the cumulative frequency histogram goes from 0 to 240 (the total frequency for all of the data). We can replace the scale of the cumulative frequency histogram shown above with a different one that expresses the cumulative frequency in percents. Since 240 students represent 100% of the students taking the biology test, we write 100% to correspond to a cumulative frequency of 240. Similarly, since 0 students represent 0% of the students taking the biology test, we write 0% to correspond to a cumulative frequency of 0.

If we divide the percent scale into four equal parts, we can label the three added divisions as 25%, 50%, and 75%.

Thus the graph relates each cumulative frequency to a percent of the total number of biology students. For example, 120 students (half of the total number) corresponds to 50%.

Let us use the percent scale to answer the question, "What percent of the students scored 70 or below on the test?" The height of each bar represents both the number of students and the percent of the students who had scores at or below the largest number in the interval represented by that bar. Since 25%, or a quarter, of the scores were 70 or below, we say that 70 is an approximate value for the lower quartile, or the 25th percentile.



From the histogram, we can see that about 56% of the students had scores at or below 80. Thus, the second quartile, the median, is in the 51–80 interval. For these data, the upper quartile is in the 51–90 interval.

From the histogram, we can also conveniently read the approximate percentiles for the scores that are the end values of the intervals. For example, to find the percentile for a score of 60, the right-end score of the first interval, we draw a horizontal line segment from the height of the first interval to the percent scale, as shown by the dashed line in the histogram above. The fact that the horizontal line crosses the percent scale at about one-third the distance between 0% and 25% tells us that approximately 8% of the students scored 60 or below 60. Thus, the 8th percentile is a good estimate for a score of 60.

EXAMPLE 3

A reporter for the local newspaper is preparing an article on the ice cream stores in the area. She listed the following prices for a two-scoop cone at 15 stores.

$2.48, $2.57, $2.30, $2.79, $2.25, $3.00, $2.82, $2.75,
$2.55, $2.98, $2.53, $2.40, $2.80, $2.50, $2.65

a. List the data in a stem-and-leaf diagram.

b. Find the median.

c. Find the first and third quartiles.

d. Construct a box-and-whisker plot.

e. Draw a cumulative frequency histogram.

f. Find the percentile rank of a price of $2.75.

Solution

a. The first two digits in each price will be the stem. The lowest price is $2.25 and the highest price is $3.00.

Stem	Leaf
3.0	0
2.9	8
2.8	0 2
2.7	5 9
2.6	5
2.5	0 3 5 7
2.4	0 8
2.3	0
2.2	5

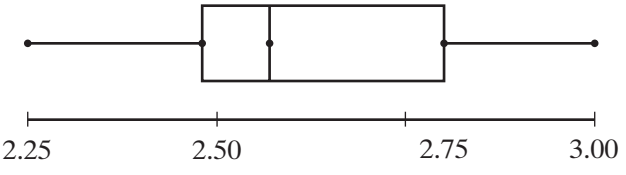
Key: 2.9 | 8 = $2.98

b. Since there are 15 prices, the median is the 8th from the top or from the bottom. The median is $2.57.

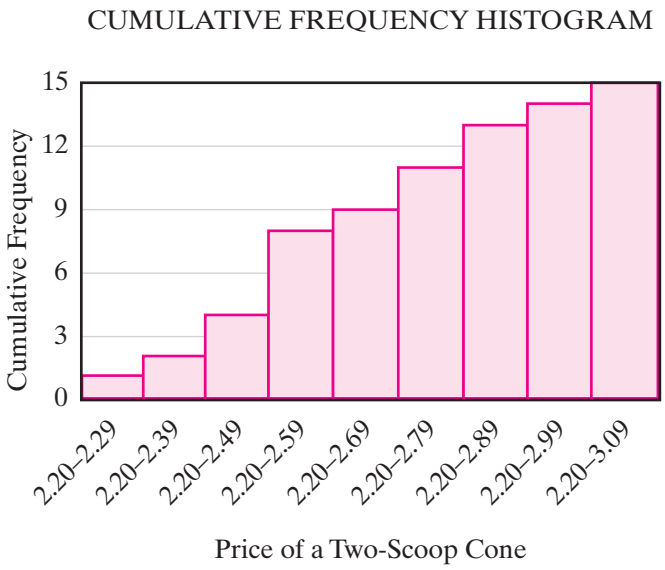
c. The middle value of the set of numbers below the median is the first quartile. That price is $2.48.

The middle value of the set of numbers above the median is the third quartile. That price is $2.80.

d. Use a scale from $2.25 to $3.00. Place dots at $2.48, $2.57, and $2.80 for the first quartile, the median, and the third quartile. Draw the box around the quartiles with a vertical line through the median. Add the whiskers.



Interval	Frequency	Cumulative Frequency
3.00–3.09	1	15
2.90–2.99	1	14
2.80–2.89	2	13
2.70–2.79	2	11
2.60–2.69	1	9
2.50–2.59	4	8
2.40–2.49	2	4
2.30–2.39	1	2
2.20–2.29	1	1



e. Make a cumulative frequency table and draw the histogram. Use 2.20–2.29 as the smallest interval.

f. There are 9 data values below \$2.75. Add $\frac{1}{2}$ for the data value \$2.75.

$$\text{Percentile rank: } \frac{9\frac{1}{2}}{15} = 0.6\overline{3} \approx 63\%$$

Answers **a.** Diagram **b.** median = \$2.57 **c.** first quartile = \$2.48; third quartile = \$2.80 **d.** Diagram **e.** Diagram **f.** 63rd percentile

Note: A cumulative frequency histogram can be drawn on a calculator just like a regular histogram. In list L_2, where we previously entered the frequencies for each individual interval, we now enter each cumulative frequency.

EXERCISES

Writing About Mathematics

1. a. Is it possible to determine the percentile rank of a given score if the set of scores is arranged in a stem-and-leaf diagram? Explain why or why not.

b. Is it possible to determine the percentile rank of a given score if the set of scores is shown on a cumulative frequency histogram? Explain why or why not.

2. A set of data consisting of 23 consecutive numbers is written in numerical order from left to right.

a. The number that is the first quartile is in which position from the left?

b. The number that is the third quartile is in which position from the left?

Developing Skills

In 3–6, for each set of data: **a.** Find the five numbers of the statistical summary **b.** Draw a box-and-whisker plot.

3. 12, 17, 20, 21, 25, 27, 29, 30, 32, 33, 33, 37, 40, 42, 44

4. 67, 70, 72, 77, 78, 78, 80, 84, 86, 88, 90, 92

5. 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 5, 7, 9, 9

6. 3.6, 4.0, 4.2, 4.3, 4.5, 4.8, 4.9, 5.0

In 7–9, data are grouped into tables. For each set of data:

a. Construct a cumulative frequency histogram.

b. Find the interval in which the lower quartile lies.

c. Find the interval in which the median lies.

d. Find the interval in which the upper quartile lies.

7.

Interval	Frequency
41–50	8
31–40	5
21–30	2
11–20	5
1–10	4

8.

Interval	Frequency
25–29	3
20–24	1
15–19	3
10–14	9
5–9	9

9.

Interval	Frequency
1–4	4
5–8	3
9–12	7
13–16	2
17–20	2

10. For the data given in the table:

a. Construct a cumulative frequency histogram.

b. In what interval is the median?

c. The value 10 occurs twice in the data. What is the percentile rank of 10?

Interval	Frequency
21–25	5
16–20	4
11–15	6
6–10	3
1–5	2

11. For the data given in the table:

a. Construct a cumulative frequency histogram.

b. In what interval is the median?

c. In what interval is the upper quartile?

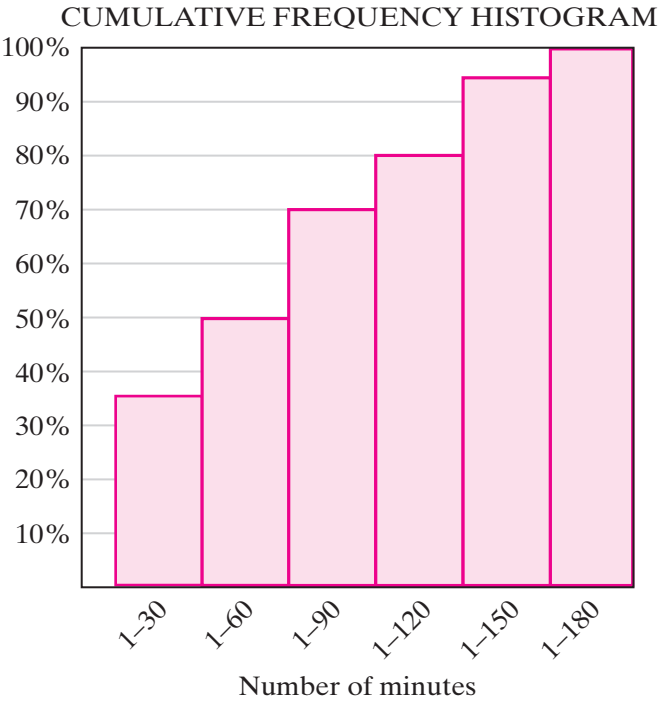
d. What percent of scores are 17 or less?

e. In what interval is the 25th percentile?

Interval	Frequency
33–37	4
28–32	3
23–27	7
18–22	12
13–17	8
8–12	5
3–7	1

Applying Skills

12. A group of 400 students were asked to state the number of minutes that each spends watching television in 1 day. The cumulative frequency histogram shown below summarizes the responses as percents.



a. What percent of the students questioned watch television for 90 minutes or less each day?

b. How many of the students watch television for 90 minutes or less each day?

c. In what interval is the upper quartile?

d. In what interval is the lower quartile?

e. If one of these students is picked at random, what is the probability that he or she watches 30 minutes or less of television each day?

13. A journalism student was doing a study of the readability of the daily newspaper. She chose several paragraphs at random and listed the number of letters in each of 88 words. She prepared the following chart.

a. Copy the chart, adding a column that lists the cumulative frequency

b. Find the median.

c. Find the first and third quartiles.

d. Construct a box-and whisker plot.

e. Draw a cumulative frequency histogram.

f. Find the percentile rank of a word with 7 letters.

Number of letters	Frequency
1	4
2	14
3	20
4	20
5	3
6	18
7	5
8	2
9	1
10	1

14. Cecilia's average for 4 years is 86. Her average is the upper quartile for her class of 250 students. At most, how many students in her class have averages that are less than Cecilia's?

15. In the table at the right, data are given for the heights, in inches, of 22 football players.

a. Copy and complete the table.

b. Draw a cumulative frequency histogram.

c. Find the height that is the lower quartile.

d. Find the height that is the upper quartile.

Height (inches)	Frequency	Cumulative Frequency
77	2	
76	2	
75	7	
74	5	
73	3	
72	2	
71	1	

16. The lower quartile for a set of data was 40. These data consisted of the heights, in inches, of 680 children. At most, how many of these children measured more than 40 inches?

In 17 and 18, select, in each case, the numeral preceding the correct answer.

17. On a standardized test, Sally scored at the 80th percentile. This means that

(1) Sally answered 80 questions correctly.

(2) Sally answered 80% of the questions correctly.

(3) Of the students who took the test, about 80% had the same score as Sally.

(4) Of the students who took the test, at least 80% had scores that were less than or equal to Sally's score.

18. For a set of data consisting of test scores, the 50th percentile is 87. Which of the following could be *false*?

(1) 50% of the scores are 87.

(2) 50% of the scores are 87 or less.

(3) Half of the scores are at least 87.

(4) The median is 87.

16-7 BIVARIATE STATISTICS

We have been studying **univariate statistics** or statistics that involve a single set of numbers. Statistics are often used to study the relationship between two different sets of values. For example, a dietician may want to study the relationship between the number of calories from fat in a person's diet and the level of cholesterol in that person's blood, or a merchant may want to study the relationship between the amount spent on advertising and gross sales. Although these examples involving **two-valued statistics** or **bivariate statistics** require complex statistical methods, we can investigate some of the properties of similar but simpler problems by looking at graphs and by using a graphing calculator. A graph that shows the pairs of values in the data as points in the plane is called a **scatter plot**.

Correlation

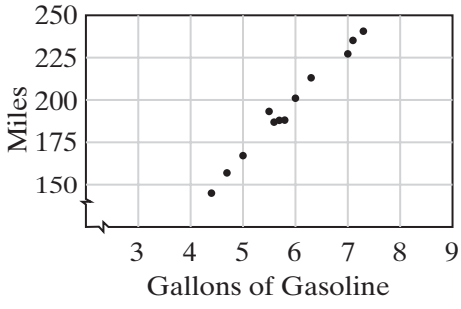
We will consider five cases of two-valued statistics to investigate the relationship or **correlation** between the variables based on their scatter plots.

CASE 1 *The data has positive linear correlation. The points in the scatter plot approximate a straight line that has a positive slope.*

A driver recorded the number of gallons of gasoline used and the number of miles driven each time she filled the tank. In this example, there is both correlation and **causation** since the increase in the number of miles driven causes the number of gallons of gasoline needed to increase.

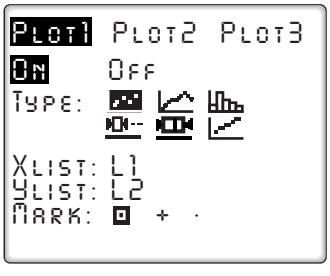
Gallons	7.2	5.8	7.0	5.5	5.6	7.1	6.0	4.4	5.0	6.2	4.7	5.7
Miles	240	188	226	193	187	235	202	145	167	212	154	188

This scatter plot can be duplicated on your graphing calculator. Enter the number of miles as L_1 and the number of gallons of gasoline as corresponding entries in L_2. The miles will be graphed as x-values and the gallons of gasoline as y-values. First, turn on Plot 1:



ENTER: 2nd STAT PLOT 1 ENTER ▼ ENTER ▼ 2nd L1 ▼ 2nd L2

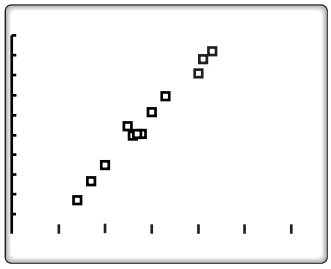
DISPLAY:



Now use ZoomStat from the ZOOM menu to construct a window that will include all values of x and y.

ENTER: ZOOM 9

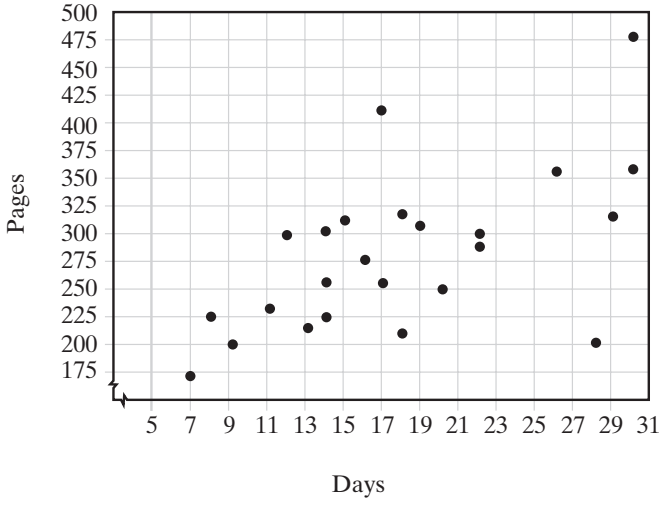
DISPLAY:



CASE 2 *The data has moderate positive correlation. The points in a scatter plot do not lie in a straight line but there is a general tendency for the values of y to increase as the values of x increase.*

Last month, each student in an English class was required to choose a book to read. The teacher recorded, for each student in the class, the number of days spent reading the book and the number of pages in the book.

Days	8	14	12	26	9	17	28	13	15	30	18	20
Pages	225	300	298	356	200	412	205	215	310	357	209	250
Days	29	22	17	14	11	14	22	19	16	7	18	30
Pages	314	288	256	225	232	256	300	305	276	172	318	480



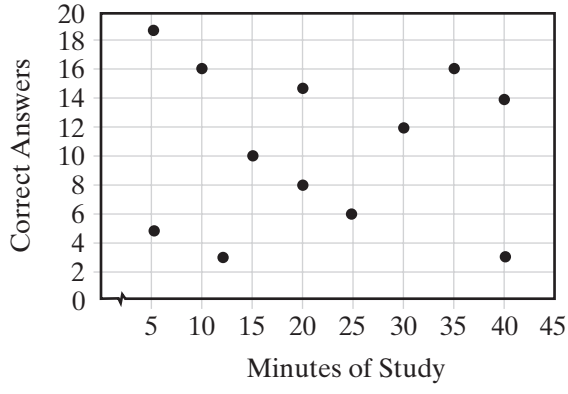
While the books with more pages may have required more time, some students read more rapidly and some spent more time each day reading. The graph shows that, in general, as the number of days needed to read a book increased, the number of pages that were read also increased.

CASE 3 *The data has no correlation.*

Before giving a test, a teacher asked each student how many minutes each had spent the night before preparing for the test. After correcting the test, she prepared the table below which compares the number of minutes of study to the number of correct answers.

Minutes of Study	20	15	40	5	10	25	30	12	5	20	35	40
Correct Answers	15	10	3	19	16	6	12	3	5	8	16	14

The graph shows that there is *no* correlation between the time spent studying just before the test and the number of correct answers on the test.

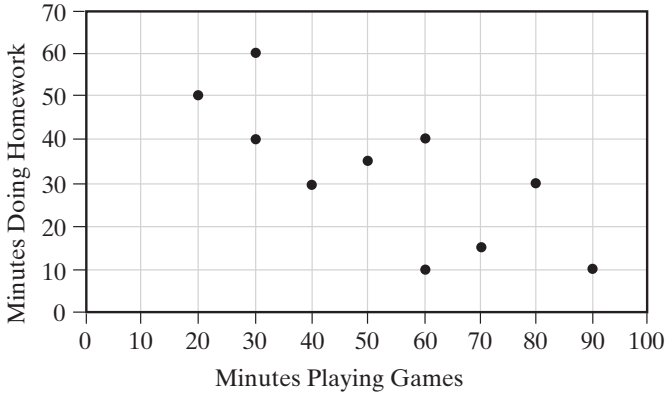


CASE 4 *The data has moderate negative correlation. The points in a scatter plot do not lie in a straight line but there is a general tendency for the values of y to decrease as the values of x increase.*

A group of children go to an after-school program at a local youth club. The director of the program keeps a record, shown below, of the time, in minutes, each student spends playing video games and doing homework.

Games	20	30	90	60	30	50	70	40	80	60
Homework	50	60	10	40	40	35	15	30	30	10

In this instance, the unit of measure, minutes, is found in the problem rather than in the table. To create meaningful graphs, always include a unit of measure on the horizontal and vertical axes.

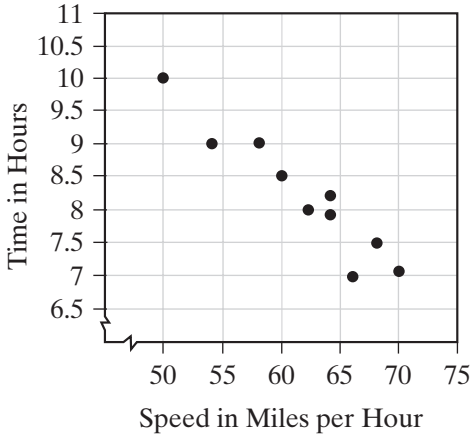


The graph shows that, in general, as the number of minutes spent playing video games increases, the number of minutes spent doing homework decreases.

CASE 5 *The data has negative linear correlation. The points in the scatter plot approximate a straight line that has a negative slope.*

A long-distance truck driver travels 500 miles each day. As he passes through different areas on his trip, his average speed and the length of time he drives each day vary. The chart below shows a record of average speed and time for a 10-day period.

Speed	50	64	68	60	54	66	70	62	64	58
Time	10	7.9	7.5	8.5	9.0	7.0	7.1	8.0	8.2	9.0



In this case, the increase in the average speed causes the time required to drive a fixed distance to decrease. This example indicates both negative correlation and causation.

It is important to note that correlation is not the same as causation. Correlation is an indication of the strength of the linear relationship or association between the variables, but it does not mean that changes in one variable are the cause of changes in the other. For example, suppose a study found there was a strong positive correlation between the number of pages in the daily newspaper and the number of voters who turn out for an election. One would not be correct in concluding that a greater number of pages causes a greater turnout. Rather, it is likely that the urgency of the issues is reflected in the increase of both the size of the newspaper and the size of the turnout.

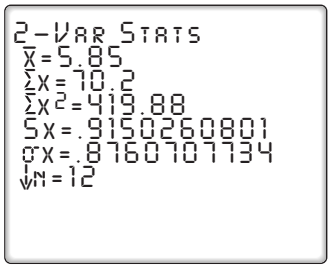
Another example where there is no causation occurs in **time series** or data that is collected at regular intervals over time. For instance, the population of the U.S. recorded every ten years is an example of a time series. In this case, we cannot say that time causes a change in the population. All we can do is note a general trend, if any.

Line of Best Fit

When it makes sense to consider one variable as the independent variable and the other as the dependent variable, and the data has a linear correlation (even if it is only moderate correlation), the data can be represented by a **line of best fit**. For example, we can write an equation for the data in Case 1. Enter the data into L_1 and L_2 if it is not already there. Find the mean values for x, the number of miles driven, and for y, the number of gallons of gasoline used. Then use 2-Var Stats from the STAT CALC menu:

ENTER: STAT ▶ 2 ENTER

DISPLAY:



The calculator gives $\bar{x} = 5.85$ and, by pressing the down arrow key, $\bar{y} = 194.75$. We will use these mean values, (5.85, 194.75), as one of the points on our line. We will choose one other data point, for example (7.1, 235), as a second point and write the equation of the line using the slope-intercept form $y = mx + b$. First find the slope:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{194.75 - 235}{5.85 - 7.1} = \frac{-40.25}{-1.25} = 32.2$$

Now use one of the points to find the y-intercept:

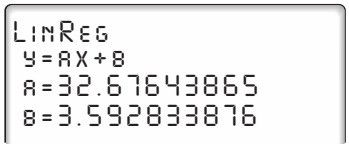
$$\begin{aligned} 194.75 &= 32.2(5.85) + b \\ 194.75 &= 188.37 + b \\ 6.38 &= b \end{aligned}$$

Round the values to three significant digits. A possible equation for a line of best fit is $y = 32.2x + 6.38$.

The calculator can also be used to find a line called the **regression line** to fit a bivariate set of data. Use the LinReg(ax+b) function in the STAT CALC menu.

ENTER: STAT ▸ 4 ENTER

DISPLAY:



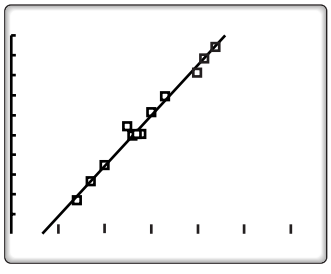
If we round the values to three significant digits, the equation of the regression line is $y = 32.7x + 3.59$. In this case, the difference between these two equations is negligible. *However*, *this is not always the case*. The regression line is a special line of best fit that minimizes the square of the vertical distances to each data point.

We can compare these two equations with the actual data. Graph the scatter plot of the data using ZoomStat. Then write the two equations in the Y= menu.

ENTER: Y= 32.2 X,T,Θ,*n* + 6.38 ENTER

32.7 X,T,Θ,*n* + 3.59 GRAPH

DISPLAY:



Notice that the lines are very close and do approximate the data.

Note 1: The equation of the line of best fit is very sensitive to rounding. Try to round the coefficients of the line of best fit to at least three significant digits or to whatever the test question asks.

Note 2: A line of best fit is appropriate only for data that exhibit a linear pattern. In more advanced courses, you will learn how to deal with nonlinear patterns.

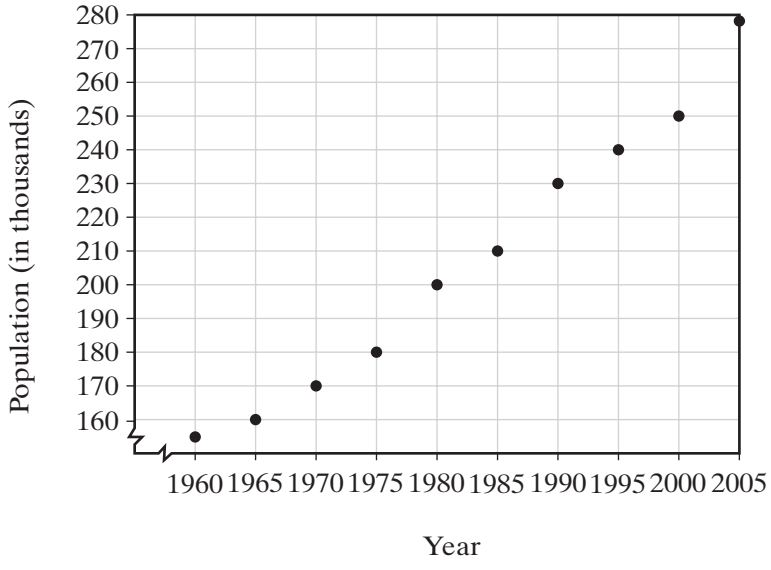
These equations can be used to predict values. For example, if the driver has driven 250 miles before filling the tank, how many gallons of gasoline should be needed? We will use the equation from the calculator.

$$\begin{aligned} y &= 32.7x + 3.59 \\ 250 &= 32.7x + 3.59 \\ 246.41 &= 32.7x \\ 7.535474006 &= x \end{aligned}$$

It is reasonable to say that the driver can expect to need about 7.5 gallons of gasoline.

What we just did is called **extrapolation**, that is, using the line of best fit to make a prediction outside of the range of data values. Using the line of best fit to make a prediction *within* the given range of data values is called **interpolation**.

In general, interpolation is usually safe, while care should be taken when extrapolating. The observed correlation pattern may not be valid outside of the given range of data values. For example, consider the scatter plot of the population of a town shown below. The population grew at a constant rate during the years in which the data was gathered. However, we do not expect the population to continue to grow forever, and thus, it may not be possible to extrapolate far into the future.



Keep in Mind In general, when a given relationship involves two sets of data:

1. In some cases a straight line, a line of best fit, can be drawn to approximate the relationship between the data sets.
2. If a line of best fit has a positive slope, the data has positive linear correlation.
3. If the line of best fit has a negative slope, the data has negative linear correlation.
4. A line of best fit can be drawn through $(\overline{x}, \overline{y})$, the point whose coordinates are the means of the given data. Any data point that appears to lie on or near the line of best fit can be used as a second point to write the equation.

5. A calculator can be used to find the regression line as the line of best fit.
6. When the graphed data points are so scattered that it is not possible to draw a straight line that approximates the given relationship, the data has no correlation.

To study bivariate data without using a graphing calculator:

1. Make a table that lists the data.
2. Plot the data as points on a graph.
3. Find the mean of each set of data and locate the point $(\overline{x}, \overline{y})$ on the graph.
4. Draw a line that best approximates the data.
5. Choose the point $(\overline{x}, \overline{y})$ and one other point or any two points that are on or close to the line that you drew. Use these points to write an equation of the line.
6. Use the equation of the line to predict related outcomes.

To study bivariate data using a graphing calculator:

1. Enter the data into L_1 and L_2 or any two lists in the memory of the calculator.
2. Use **STAT PLOT** to turn on a plot and to choose the type of plot needed. Enter the names of the lists in which the data is stored and choose the mark to be used for each data point.
3. Use ZoomStat to choose a viewing window that shows all of the data points.
4. Find the regression line using LinReg(ax+b) from the STAT CALC menu.
5. Enter the equation of the regression line in the Y= menu and use **GRAPH** to show the relationship between the data and the regression line.
6. Use the equation of the line to predict related outcomes.

In this course, we have found a line of best fit by finding a line that seems to represent the data or by using a calculator. In more advanced courses in statistics, you will learn detailed methods for finding the line of best fit.

EXAMPLE 1

The table below shows the number of calories and the number of grams of carbohydrates in a half-cup serving of ten different canned or frozen vegetables.

Carbohydrates	9	23	4	5	19	8	12	7	13	17
Calories	45	100	20	25	110	35	50	30	70	80

a. Draw a scatter plot on graph paper. Let the horizontal axis represent grams of carbohydrates and the vertical axis represent the number of calories.

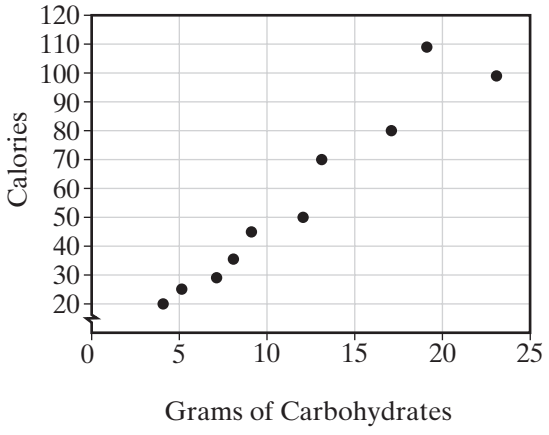
b. Find the mean number of grams of carbohydrates in a serving of vegetables and the mean number of calories in a serving of vegetables.

c. On the graph, draw a line that approximates the data in the table, and determine its equation.

d. Enter the data in L_1 and L_2 on your calculator and find the linear regression equation, LinReg(ax+b).

e. Use each equation to find the expected number of calories in a serving of vegetables with 20 grams of carbohydrates. Compare the answers.

Solution **a.**



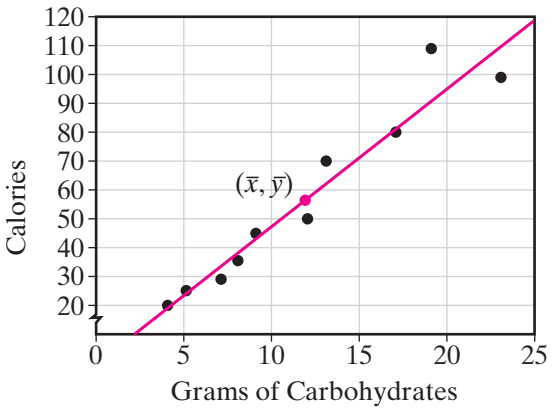
b. Enter the number of grams of carbohydrates in L_1 and the number of calories in L_2. Find $\bar{x}$ and $\bar{y}$, using 2-Var Stats.

ENTER: STAT ▶ 2 ENTER

The means of the x- and y-coordinates are $\bar{x} = 11.7$ and $\bar{y} = 56.5$.

Locate the point (11.7, 56.5) on the graph.

c.



The line we have drawn seems to go through the point $(4, 20)$. We will use this point and the point with the mean values, $(11.7, 56.5)$, to write an equation of a line of best fit.

$$m = \frac{56.5 - 20}{11.7 - 4} = \frac{36.5}{7.7} \approx 4.74$$

$$\begin{aligned} y &= mx + b \\ 20 &= \tfrac{36.5}{7.7}(4) + b \\ 1.05 &\approx b \end{aligned}$$

An equation of a best fit line is $y = 4.74x + 1.05$.

d. The data are in L_1 and L_2.

ENTER: STAT ▶ 4 ENTER

DISPLAY:

```
LinReg
y=ax+b
a=4.885506842
b=-.6604300475
```

The equation of the regression line is $y = 4.89x - 0.660$.

e. Let $x = 20$.

Use the equation from **c**.	Use the equation from **d**.
$y = 4.74x + 1.05$	$y = 4.89x - 0.660$
$y = 4.74(20) + 1.05$	$y = 4.89(20) - 0.660$
$y = 95.85$	$y = 97.14$

The two equations give very similar results. It would be reasonable to say that we could expect the number of calories to be about 96 or close to 100.

EXERCISES

Writing About Mathematics

1. a. Give an example of a set of bivariate data that has negative correlation.

b. Do you think that the change in the independent variable in your example causes the change in the dependent variable?

2. Explain the purpose of finding a line of best fit.

Applying Skills

3. When Gina bought a new car, she decided to keep a record of how much gas she uses. Each time she puts gas in the car, she records the number of gallons of gas purchased and the number of miles driven since the last fill-up. Her record for the first 2 months is as follows:

Gallons of gas	10	12	9	6	11	10	8	12	10	7
Miles driven	324	375	290	190	345	336	250	375	330	225

a. Draw a scatter plot of the data. Let the horizontal axis represent the number of gallons of gas and the vertical axis represent the number of miles driven.

b. Does the data have positive, negative, or no correlation?

c. Is this a causal relationship?

d. Find the mean number of gallons of gasoline per fill-up.

e. Find the mean number of miles driven between fill-ups.

f. Locate the point that represents the mean number of gallons of gasoline and the mean number of miles driven. Use $(0, 0)$ as a second point. Draw a line through these two points to approximate the data in the table.

g. Use the line drawn in part **d** to approximate the number of miles Gina could drive on 3 gallons of gasoline.

4. Gemma made a record of the cost and length of each of the 14 long-distance telephone calls that she made in the past month. Her record is given below.

Minutes	3.7	1.0	19.6	0.8	4.3	34.8	2.9
Cost	$0.35	$0.11	$2.12	$0.09	$0.47	$3.78	$0.24
Minutes	2.5	7.1	10.9	5.8	1.5	1.4	8.0
Cost	$0.27	$0.79	$1.21	$0.65	$0.20	$0.17	$0.89

a. Draw a scatter plot of the data on graph paper. Let the horizontal axis represent the number of minutes, and the vertical axis represent the cost of the call.

b. Does the data have positive, negative, or no correlation?

c. Is this a causal relationship?

d. Find the mean number of minutes per call.

e. Find the mean cost of the calls.

f. On the graph, draw a line of best fit for the data in the table and write its equation.

g. Use a calculator to find the equation of the regression line.

h. Approximate the cost of a call that lasted 14 minutes using the equation written in **d**.

i. Approximate the cost of a call that lasted 14 minutes using the equation written in **e**.

5. A local store did a study comparing the cost of a head of lettuce with the number of heads sold in one day. Each week, for five weeks, the price was changed and the average number of heads of lettuce sold per day was recorded. The data is shown in the chart below.

Cost per Head of Lettuce	$1.50	$1.25	$0.90	$1.75	$0.50
No. of Heads Sold	48	52	70	42	88

a. Draw a scatter plot of the data. Let the horizontal axis represent the cost of a head of lettuce and the vertical axis represent the number of heads sold.

b. Does the data have positive, negative, or no correlation?

c. Is this a causal relationship?

d. Find the mean cost per head.

e. Find the mean number of heads sold.

f. On the graph, draw a line that approximates the data in the table.

g. What appears to be the result of raising the price of a head of lettuce?

6. The chart below shows the recorded heights in inches and weights in pounds for the last 24 persons who enrolled in a health club.

Height	**Weight**	**Height**	**Weight**	**Height**	**Weight**	**Height**	**Weight**
69	160	75	180	66	145	71	165
67	160	76	155	66	130	66	155
63	135	70	175	68	160	67	140
73	185	73	170	68	140	78	210
71	215	68	190	72	170	72	160
79	225	74	190	77	195	69	145

a. Draw a scatter plot on graph paper to display the data.

b. Does the data have positive, negative, or no linear correlation?

c. Is this a causal relationship?

d. Draw and find the equation of a line of best fit. Use (77, 195) as a second point.

e. Use a calculator to find the linear regression line.

f. According to the equation written in **d.**, if the next person who enrolls in the health club is 62 inches tall, what would be the expected weight of that person?

g. According to the equation written in **e.**, if the next person who enrolls in the health club weighs 200 pounds, what would be the expected height of that person?

7. The chart below shows the number of millions of cellular telephones in use in the United States by year from 1994 to 2003.

Year	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03
Phones	24.1	33.8	44.0	55.3	69.2	86.0	109.5	128.3	140.8	158.7

Let L_1 be the number of years after 1990: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13.

a. Draw a scatter plot on graph paper to display the data.

b. Does the data have positive, negative, or no linear correlation?

c. Is this a causal relationship?

d. Draw and find the equation of a line of best fit.

e. On the graph, draw a line that approximates the data in the table, and determine its equation. Use (6, 44.0) as a second point.

f. If the line of best fit is approximately correct for years beyond 2003, estimate how many cellular phones will be in use in 2007.

8. The chart below shows, for the last 20 Supreme Court Justices to have left the court before 2000, the age at which the judge was nominated and the number of years as a Supreme Court judge.

Age	47	64	62	62	59	55	54	45	43	56
Years	15	16	24	17	24	4	3	31	23	5
Age	50	56	62	59	50	56	57	49	49	62
Years	33	16	16	7	18	7	13	6	12	1

a. Draw a scatter plot on graph paper to display the data.

b. Does the data have positive, negative, or no linear correlation?

c. Is there a causal relationship?

d. Draw and find an equation of a line of best fit.

e. Use a calculator to find the linear regression line.

f. Do you think that the data, the line of best fit, the regression line, or none of these could be used to approximate the number of years as a Supreme Court justice for the next person to retire from that office?

9. A cook was trying different recipes for potato salad and comparing the amount of dressing with the number of potatoes given in the recipe. The following data was recorded.

Number of Potatoes	7	4	2	8	6	7	5	4
Cups of Dressing	$1\frac{1}{2}$	$\frac{7}{8}$	$\frac{3}{4}$	$1\frac{1}{4}$	1	$1\frac{3}{4}$	$1\frac{1}{8}$	$\frac{3}{4}$

a. Draw a scatter plot on graph paper to display the data.

b. Does the data have positive, negative, or no linear correlation?

c. Draw and find the equation of a line of best fit. Use $\left(4, \frac{7}{8}\right)$ as a second point.

d. Use a calculator to find the linear regression line.

e. According to the equation written in **c.**, if the cook needs to use 10 potatoes to have enough salad, approximately how many cups of dressing are needed?

CHAPTER SUMMARY

Statistics is the study of numerical data. In a statistical study, data are collected, organized into tables and graphs, and analyzed to draw conclusions.

Data can either be quantitative or qualitative. **Quantitative data** represents counts or measurements. **Qualitative data** represents categories or qualities.

In an **experiment**, a researcher imposes a treatment on one or more groups. The **treatment group** receives the treatment, while the **control group** does not.

Tables and **stem-and-leaf diagrams** are used to organize data. A table should have between five and fifteen **intervals** that include all data values, are of equal size, and do not overlap.

A **histogram** is a bar graph in which the height of a bar represents the **frequency** of the data values represented by that bar.

A **cumulative frequency histogram** is a bar graph in which the height of the bar represents the total frequency of the data values that are less than or equal to the upper endpoint of that bar.

The mean, median, and mode are three **measures of central tendency**. The **mean** is the sum of the data values divided by the total frequency. The **median** is the middle value when the data values are placed in numerical order. The **mode** is the data value that has the largest frequency.

Quartile values separate the data into four equal parts. A **box-and-whisker plot** displays a set of data values using the **minimum**, the **first quartile**, the median, the **third quartile**, and the **maximum** as significant measures. The **percentile** rank tells what percent of the data values lie at or below a given measure.

In **two-valued statistics** or **bivariate statistics**, a relation between two different sets of data is studied. The data can be graphed on a **scatter plot**. The data may have positive, negative, or no correlation. Data that has positive or negative linear correlation can be represented by a **line of best fit**.

The line of best fit can be used to predict values not in the included data set. **Interpolation** is predicting within the given data range. **Extrapolation** is predicting outside of the given data range.

VOCABULARY

16-1 Data • Statistics • Descriptive statistics • Qualitative data • Quantitative data • Census • Sample • Bias • Experiment • Treatment group • Control group • Placebo effect • Placebo • Blinding • Single-blind experiment • Double-blind experiment

16-2 Tally • Frequency • Total frequency • Frequency distribution table • Group • Interval • Grouped data • Range • Stem-and-leaf diagram •

16-3 Histogram • Frequency histogram

16-4 Average • Measures of central tendency • Mean • Arithmetic mean • Numerical average • Median • Mode • Bimodal • Linear transformation of a data set

16-5 Group mode • Modal interval

16-6 Quartile • Lower quartile • First quartile • Second quartile • Upper quartile • Third quartile • Box-and-whisker plot • Five statistical summary • Percentile • Cumulative frequency • Cumulative frequency histogram

16-7 Univariate statistics • Two-valued statistics • Bivariate statistics • Scatter plot • Correlation • Causation • Time series • Line of best fit • Regression line • Extrapolation • Interpolation

REVIEW EXERCISES

1. Courtney said that the mean of a set of consecutive integers is the same as the median and that the mean can be found by adding the smallest and the largest numbers and dividing the sum by 2. Do you agree with Courtney? Explain why or why not.

2. A set of data contains N numbers arranged in numerical order.

a. When is the median one of the numbers in the set of data?

b. When is the median not one of the numbers in the set of data?

3. For each of the following sets of data, find: **a.** the mean **b.** the median **c.** the mode (if one exists)

(1) 3, 4, 3, 4, 3, 5

(2) 1, 3, 5, 7, 1, 2, 4

(3) 9, 3, 2, 8, 3, 3

(4) 9, 3, 2, 3, 8, 2, 7

4. Express, in terms of y, the mean of $3y - 2$ and $7y + 18$.

5. For the following data:

78, 91, 60, 65, 81, 72, 78, 80, 65, 63, 59, 78, 78, 54, 87, 75, 77

a. Use a stem-and-leaf diagram to organize the data.

b. Draw a histogram, using 50–59 as the lowest interval.

c. Draw a cumulative frequency histogram.

d. Draw a box-and-whisker plot.

6. The weights, in kilograms, of five adults are 53, 72, 68, 70, and 72.

a. Find: (1) the mean (2) the median (3) the mode

b. If each of the adults lost 5 kilograms, find, for the new set of weights: (1) the mean (2) the median (3) the mode

7. Steve's test scores are 82, 94, and 91. What grade must Steve earn on a fourth test so that the mean of his four scores will be exactly 90?

8. From Monday to Saturday of a week in May, the recorded high temperature readings were 72°, 75°, 79°, 83°, 83°, and 88°. For these data, find:

a. the mean **b.** the median **c.** the mode

9. Paul worked the following numbers of hours each week over a 20-week period:

15, 3, 7, 6, 2, 14, 9, 25, 8, 12, 8, 8, 15, 0, 8, 12, 28, 10, 14, 10

a. Organize the data in a frequency table, using 0–5 as the lowest interval.

b. Draw a frequency histogram of the data.

c. In what interval does the median lie?

d. Which interval contains the lower quartile?

10. The table shows the scores of 25 test papers.

Score	Frequency	Cumulative Frequency
60	1	
70	9	
80	8	
90	2	
100	5	

a. Is the data univariate or bivariate?

b. Find the mean score.

c. Find the median score.

d. Find the mode.

e. Copy and complete the table.

f. Draw a cumulative frequency histogram.

g. Find the percentile rank of 90.

h. What is the probability that a paper chosen at random has a score of 80?

11. The electoral votes cast for the winning presidential candidate in elections from 1900 to 2004 are as follows:

292, 336, 321, 435, 277, 404, 382, 444, 472, 523, 449, 432, 303, 442,
457, 303, 486, 301, 520, 297, 489, 525, 426, 370, 379, 271, 286

a. Organize the data in a stem-and-leaf diagram. (Use the first digit as the stems, and the last two digits as the leaves.)

b. Find the median number of electoral votes cast for the winning candidate.

c. Find the first-quartile and third-quartile values.

d. Draw a box-and-whisker plot to display the data.

12. The ages of 21 high school students are shown in the table at the right.

Age	Frequency
18	1
17	4
16	2
15	7
14	2
13	5

a. What is the median age?

b. What is the percentile rank of age 15?

c. When the ages of these 21 students are combined with the ages of 20 additional students, the median age remains unchanged. What is the smallest possible number of students under 16 in the second group?

13. For each variable, determine if it is qualitative or quantitative.

a. Major in college

b. GPA in college

c. Wind speed of a hurricane

d. Temperature of a rodent

e. Yearly profit of a corporation

f. Number of students late to class

g. Zip code

h. Employment status

14. Researchers looked into a possible relationship between alcoholism and pneumonia. They conducted a study of 100 current alcoholics, 50 former alcoholics, and 1,000 non-alcoholics who were hospitalized for a mild form of pneumonia. The researchers found that 30% of alcoholics and 30% of former alcoholics, versus only 15% of the non-alcoholics developed a more dangerous form of pneumonia. The researchers concluded that alcoholism raises the risk for developing pneumonia.

Discuss possible problems with this study.

15. Aurora buys oranges every week. The accompanying table lists the weights and the costs of her last 10 purchases of oranges.

Weight (lb)	2.2	1.2	3.6	4.5	1.0	2.5	1.8	5.0	3.5	1.7
Cost ($)	1.22	0.60	1.04	1.58	0.50	0.89	0.95	1.88	1.46	0.70

a. Is the data univariate or bivariate?

b. Draw a scatter plot of the data on graph paper. Let the horizontal axis represent the weights of the oranges and the vertical axis the costs.

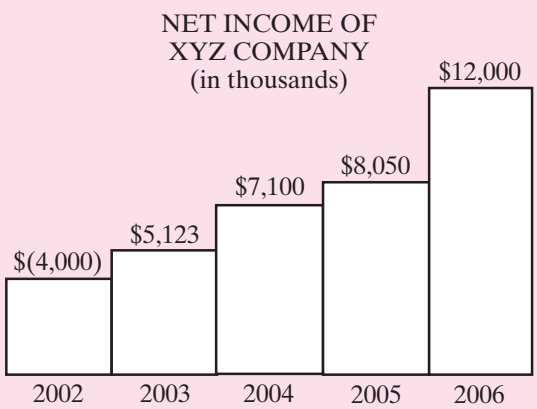
c. Is there a correlation between the weight and the cost of the oranges? If so, is it positive or negative?

d. If the price is determined by the number of oranges purchased, do the variables have a causal relationship? Explain your answer.

e. On the graph, draw a line of best fit that approximates the data in the table and write its equation.

f. Use the equation written in **d** to approximate the cost of 4 pounds of oranges.

16. Explain why the graph on the right is misleading. (*Hint:* In accounting, numbers enclosed by parentheses denote negative numbers.)



Exploration

a. Marny took the SAT in 2004 and scored a 1370. She was in the 94th percentile. Jordan took the SAT in 2000 and scored 1370. He was in the 95th percentile. Explain how this is possible.

b. Taylor's class rank stayed the same even though he had a cumulative grade point average of 3.4 one semester and 3.8 the next semester. Explain how this is possible.

CUMULATIVE REVIEW CHAPTERS 1–16

Part I

Answer all questions in this part. Each correct answer will receive 2 credits. No partial credit will be allowed.

1. When the domain is the set of integers, the solution set of the inequality $0 < 0.1x - 0.4 \leq 0.2$ is

(1) $\{\}$ (2) $\{4, 5\}$ (3) $\{4, 5, 6\}$ (4) $\{5, 6\}$

2. The product $(2a + 3)(2a - 3)$ can be written as

(1) $2a^2 - 9$ (3) $4a^2 + 9$

(2) $4a^2 - 9$ (4) $4a^2 - 12a + 9$

3. When 0.00034 is written in the form 3.4×10^n, the value of n is

(1) -3 (2) -4 (3) 3 (4) 4

4. When $\frac{x}{3} + \frac{1}{2} = \frac{x-2}{6}$, x equals

(1) -5 (2) -1 (3) $\frac{1}{2}$ (4) 1

5. The mean of the set of even integers from 2 to 100 is

(1) 49 (2) 50 (3) 51 (4) 52

6. The probability that 9 is the sum of the numbers that appear when two dice are rolled is

(1) $\frac{4}{6}$ (2) $\frac{2}{36}$ (3) $\frac{2}{6}$ (4) $\frac{4}{36}$

7. If the circumference of a circle is 12 centimeters, then the area of the circle is

(1) 36 square centimeters (3) $\frac{36}{\pi}$ square centimeters

(2) 144 square centimeters (4) $\frac{144}{\pi}$ square centimeters

8. Which of the following is not an equation of a function?

(1) $y = 3x + 2$ (3) $y^2 = x$

(2) $y = x^2 + 3x + 1$ (4) $y = |x|$

9. The value of ${}_{10}P_8$ is

(1) 80 (2) 90 (3) 1,814,400 (4) 3,628,800

10. Which of the following is an equation of a line parallel to the line whose equation is $y = -2x + 4$?

(1) $2x + y = 7$ (2) $y - 2x = 7$ (3) $2x - y = 7$ (4) $y = 2x + 7$

Part II

Answer all questions in this part. Each correct answer will receive 2 credits. Clearly indicate the necessary steps, including appropriate formula substitu-

tions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

11. In a bridge club, there are three more women than men. How many persons are members of the club if the probability that a member chosen at random, is a woman is $\frac{3}{5}$?

12. Find to the nearest degree the measure of the smallest angle in a right triangle whose sides measure 12, 35, and 37 inches.

Part III

Answer all questions in this part. Each correct answer will receive 3 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

13. The lengths of the sides of a triangle are in the ratio 3 : 5 : 6. The perimeter of the triangle is 49.0 meters. What is the length of each side of the triangle?

14. Huy worked on an assignment for four days. Each day he worked half as long as he worked the day before and spent a total of 3.75 hours on the assignment.

a. How long did Huy work on the assignment each day?

b. Find the mean number of hours that Huy worked each day.

Part IV

Answer all questions in this part. Each correct answer will receive 4 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

15. The perimeter of a garden is 16 feet. Let x represent the width of the garden.

a. Write an equation for the area of the land, y, in terms of x.

b. Sketch the graph of the equation that you wrote in **a**.

c. What is the maximum area of the land?

16. Each morning, Malcolm leaves for school at 8:00 o'clock. His brother Marvin leaves for the same school at 8:15. Malcolm walks at 2 miles an hour and Marvin rides his bicycle at 8 miles an hour. They follow the same route to school and arrive at the same time.

a. At what time do Malcolm and Marvin arrive at school?

b. How far is the school from their home?