

2 | DESCRIPTIVE STATISTICS



Figure 2.1 When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

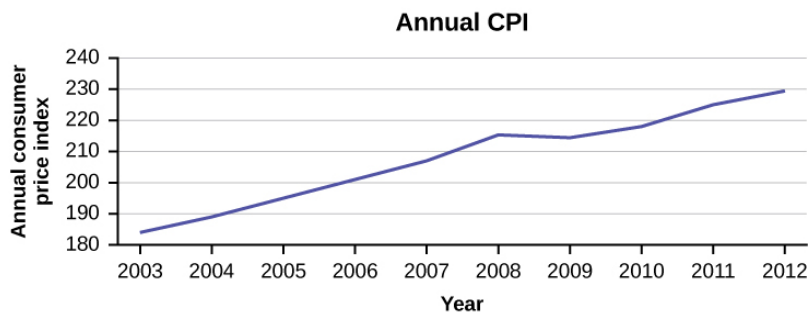
Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

Solution 2.12**Figure 2.10****Try It**

2.12 The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO₂ emissions for the United States.

CO ₂ Emissions			
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Table 2.20**Uses of a Time Series Graph**

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

2.3 | Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**.

Quartiles are special percentiles. The first quartile, Q_1 , is the same as the 25th percentile, and the third quartile, Q_3 , is the same as the 75th percentile. The median, M , is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1

Ordered from smallest to largest:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1 , is the middle value of the lower half of the data, and the third quartile, Q_3 , is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.

1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, Q_3 , is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$IQR = Q_3 - Q_1$$

The *IQR* can help to determine potential **outliers**. A value is suspected to be a potential outlier if it is less than $(1.5)(IQR)$ below the first quartile or more than $(1.5)(IQR)$ above the third quartile. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.13

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution 2.13

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230,500 + 387,000}{2} = 308,750$$

$$Q_3 = \frac{639,000 + 659,000}{2} = 649,000$$

$$IQR = 649,000 - 308,750 = 340,250$$

$$(1.5)(IQR) = (1.5)(340,250) = 510,375$$

$$Q_1 - (1.5)(IQR) = 308,750 - 510,375 = -201,625$$

$$Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375$$

No house price is less than $-201,625$. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Try It

2.13 For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

\$33,000; \$64,500; \$28,000; \$54,000; \$72,000; \$68,500; \$69,000; \$42,000; \$54,000; \$120,000; \$40,500

Example 2.14

For the two data sets in the **test scores example**, find the following:

- The interquartile range. Compare the two interquartile ranges.
- Any outliers in either set.

Solution 2.14

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Table 2.21

- The *IQR* for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$
The *IQR* for the night group is $Q_3 - Q_1 = 89 - 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

- Day class outliers are found using the *IQR* times 1.5 rule. So,
 $Q_1 - IQR(1.5) = 56 - 26.5(1.5) = 16.25$

$$Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$$

Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

$$Q_1 - IQR(1.5) = 78 - 11(1.5) = 61.5$$

$$Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5$$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Try It

2.14 Find the interquartile range for the following two data sets and compare them.

Test Scores for Class A

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Test Scores for Class B

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Example 2.15

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Table 2.22

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives,

sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight.** Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3 , is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Try It

2.15 Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Table 2.23

Example 2.16

Using **Table 2.22**:

- Find the 80th percentile.
- Find the 90th percentile.
- Find the first quartile. What is another name for the first quartile?

Solution 2.16

Using the data from the frequency table, we have:

- The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th and 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
- The 90th percentile will be the 45th data value (location is $0.90(50) = 45$) and the 45th data value is nine.
- Q_1 is also the 25th percentile. The 25th percentile location calculation: $P_{25} = 0.25(50) = 12.5 \approx 13$ the 13th data value. Thus, the 25th percentile is six.

Try It

2.16 Refer to the **Table 2.23**. Find the third quartile. What is another name for the third quartile?

Example 2.20

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution 2.20

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It

2.20 On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Example 2.21

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Solution 2.21

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Try It

2.21 During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Example 2.22

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

Min = 0

$Q_1 = 20$

Med = 40
 $Q_3 = 60$
 Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes ($60 - 20 = 40$), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120.$$

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

Min = 0
 $Q_1 = 20$
 $Q_3 = 60$
 Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

2.4 | Box Plots

Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box**. The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

NOTE

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

NOTE

See the calculator instructions on the **TI web site** (<http://education.ti.com/educationportal/sites/US/sectionHome/support.html>) or in the appendix.

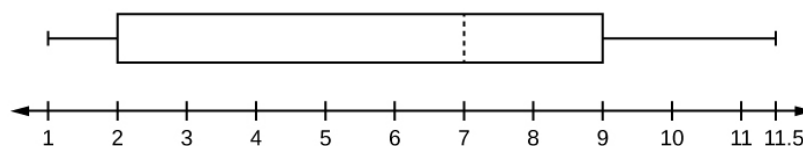


Figure 2.11

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE

It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example 2.23

The following data are the heights of 40 students in a statistics class.

59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77

Construct a box plot with the following properties; the calculator instructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- Q1: First quartile = 64.5
- Q2: Second quartile or median = 66
- Q3: Third quartile = 70

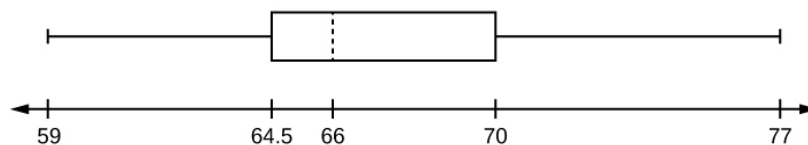


Figure 2.12

- a. Each quarter has approximately 25% of the data.
- b. The spreads of the four quarters are $64.5 - 59 = 5.5$ (first quarter), $66 - 64.5 = 1.5$ (second quarter), $70 - 66 = 4$ (third quarter), and $77 - 70 = 7$ (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- c. Range = maximum value – the minimum value = $77 - 59 = 18$
- d. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
- e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
- f. The middle 50% (middle half) of the data has a range of 5.5 inches.



Using the TI-83, 83+, 84, 84+ Calculator

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1 : First quartile = 64.5.

Q_2 : Second quartile or median = 66.

Q_3 : Third quartile = 70.

To construct the box plot:

Press 4:Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE, and use the arrow keys to examine the box plot.

Try It Σ

2.23 The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



Figure 2.13

In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five,

Try It Σ



2.24 The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Example 2.25

Graph a box-and-whisker plot for the data values shown.

10; 10; 10; 15; 35; 75; 90; 95; 100; 175; 420; 490; 515; 515; 790

The five numbers used to create a box-and-whisker plot are:

Min: 10

Q_1 : 15

Med: 95

Q_3 : 490

Max: 790

The following graph shows the box-and-whisker plot.

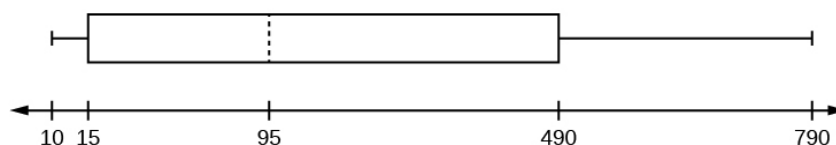


Figure 2.15

Try It Σ

2.25 Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0; 5; 5; 15; 30; 30; 45; 50; 50; 60; 75; 110; 140; 240; 330

2.5 | Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common

practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average” is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced “x bar”): \bar{x} .

The Greek letter μ (pronounced “mew”) represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4}{11} = 2.7$$

$$\bar{x} = \frac{3(1) + 2(2) + 1(3) + 5(4)}{11} = 2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then

$\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and

the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.26

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

Calculate the mean and the median.

Solution 2.26

The calculation for the mean is:

$$\bar{x} = \frac{[3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + \dots + 35 + 37 + 40 + (44)(2) + 47]}{40} = 23.6$$

To find the median, M , first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24 + 24}{2} = 24$$



Using the TI-83, 83+, 84, 84+ Calculator

To find the mean and the median:

Clear list L1. Press STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$$\bar{x} = 23.6, M = 24$$

Try It Σ

2.26 The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Example 2.27

Suppose that in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the "center": the mean or the median?

Solution 2.27

$$\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129,400$$

$$M = 30,000$$

(There are 49 people who earn \$30,000 and one person who earns \$5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Try It Σ

2.27 In a sample of 60 households, one house is worth \$2,500,000. Half of the rest are worth \$280,000, and all the others are worth \$315,000. Which is the better measure of the "center": the mean or the median?

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

Table 2.24

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean \bar{x} is an example of a statistic which estimates the population mean μ .

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: $mean = \frac{\text{data sum}}{\text{number of data values}}$. We simply need to modify the definition to fit within the restrictions

of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{\text{lower boundary} + \text{upper boundary}}{2}$. We can now modify the mean definition to be

$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example 2.30

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Table 2.25

Solution 2.30

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Table 2.26

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$$

$$\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$$

Try It Σ

2.30 Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Table 2.27

What is the best estimate for the mean number of hours spent playing video games?

2.6 | Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

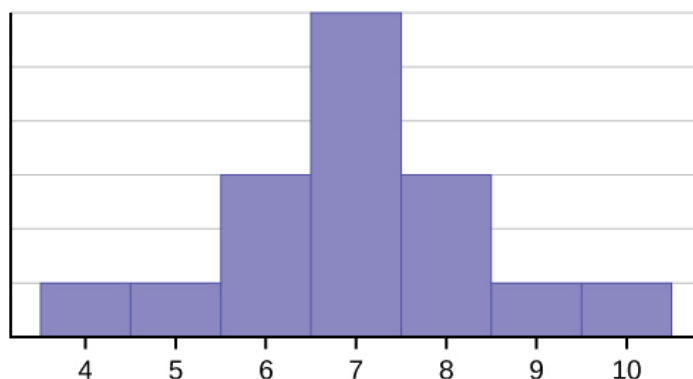


Figure 2.16

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

2.7 | Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. the average wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes; at supermarket B the standard deviation for the wait time is four minutes.

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket A. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket A, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.

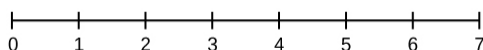


Figure 2.24

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer

- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- **sample:** $x = \bar{x} + (\# \text{ of } STDEV)(s)$
- **Population:** $x = \mu + (\# \text{ of } STDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol \bar{x} is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference " $x - \text{mean}$ " is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N , the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $n - 1$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}}$
- For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$
- For the population standard deviation, the denominator is N , the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

NOTE

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 2.32

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s .

Data	Freq.	Deviations	<i>Deviations</i> ²	(Freq.)(<i>Deviations</i> ²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times 0.950625 = 2.851875$
				The total is 9.7375

Table 2.29

The sample variance, s^2 , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one ($20 - 1$):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891, \text{ which is rounded to two decimal places, } s = 0.72.$$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
- For a sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$

- For a population: $x = \mu + (\text{\#ofSTDEVs})(\sigma)$
 - For this example, use $x = \bar{x} + (\text{\#ofSTDEVs})(s)$ because the data is from a sample
- a. Verify the mean and standard deviation on your calculator or computer.
 - b. Find the value that is one standard deviation above the mean. Find $(\bar{x} + 1s)$.
 - c. Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.
 - d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution 2.32

- a.
 - Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
 - Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
 - Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
 - Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
 - $\bar{x} = 10.525$
 - Use Sx because this is sample data (not a population): $Sx=0.715891$
- b. $(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
- c. $(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$
- d.
 - $(\bar{x} - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
 - $(\bar{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Try It

2.32 On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is -1.525 for the data value nine. **If you add the deviations, the sum is always zero.** (For [Example 2.32](#), there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample.

For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

NOTE



Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data.** Display your data in a histogram or a box plot.

Example 2.33

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
 - The sample mean
 - The sample standard deviation
 - The median
 - The first quartile
 - The third quartile
 - IQR
- Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution 2.33

- See **Table 2.30**
- The sample mean = 73.5
 - The sample standard deviation = 17.9
 - The median = 73
 - The first quartile = 61
 - The third quartile = 90
 - $IQR = 90 - 61 = 29$

- c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5) \div 5$, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5$ = the ending value; No data values fall on an interval boundary.

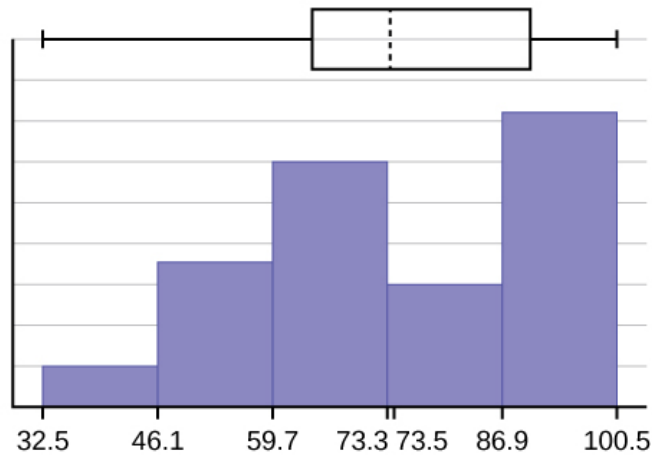


Figure 2.25

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores ($IQR = 29$) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Table 2.30

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Table 2.30

Try It



2.33 The following data show the different types of pet food stores in the area carry.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of

the measures of center by finding the mean of the grouped data with the formula: $\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f}$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example 2.34

Find the standard deviation for the data in **Table 2.31**.

Class	Frequency, f	Midpoint, m	m^2	\bar{x}^2	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5


Table 2.31

For this data set, we have the mean, $\bar{x} = 7.58$ and the standard deviation, $s_x = 3.5$. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since $7.58 - 3.5 - 3.5$

$= 0.58$. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{\sum f(m - \bar{x})^2}{n - 1}}$ where s_x

$=$ sample standard deviation, \bar{x} = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Try It Σ

 **2.34** Find the standard deviation for the data from the previous example

Class	Frequency, f
0–2	1
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

Table 2.32

First, press the **STAT** key and select **1:Edit**



```

EDIT  CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
  
```

Figure 2.26

Input the midpoint values into **L1** and the frequencies into **L2**

L1	L2	L3	2
1	1	-----	
4	6		
7	10		
10	7		
13	0		
16	2		
-----	-----		
L2(7) =			

Figure 2.27

Select **STAT**, **CALC**, and **1: 1-Var Stats**

EDIT	CALC	TESTS
1	1-Var Stats	
2	2-Var Stats	
3	Med-Med	
4	LinReg(ax+b)	
5	QuadReg	
6	CubicReg	
7	QuartReg	

Figure 2.28

Select **2nd** then **1** then , **2nd** then **2** **Enter**

1-Var Stats
$\bar{x}=7.576923077$
$\Sigma x=197$
$\Sigma x^2=1799$
$Sx=3.500549407$
$\sigma x=3.432571103$
$\downarrow n=26$

Figure 2.29

You will see displayed both a population standard deviation, σ_x , and the sample standard deviation, s_x .

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: $\text{value} = \text{mean} + (\text{\#ofSTDEVs})(\text{standard deviation})$; solve for #ofSTDEVs.

- $\# of STDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z . In symbols, the formulas become:

Sample	$x = \bar{x} + zS$	$z = \frac{x - \bar{x}}{S}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Table 2.33

Example 2.35

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Table 2.34

Solution 2.35

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \# \text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

$$\text{For John, } z = \# \text{ of STDEVs} = \frac{2.85 - 3.0}{0.7} = -0.21$$

$$\text{For Ali, } z = \# \text{ of STDEVs} = \frac{77 - 80}{10} = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of -0.21 is higher than Ali's z-score of -0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Try It

2.35 Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Table 2.35

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

2.8 | Descriptive Statistics

2.1 Descriptive Statistics

Class Time:
Names:

Student Learning Outcomes

- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data

Record the number of pairs of shoes you own.

1. Randomly survey 30 classmates about the number of pairs of shoes they own. Record their values.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 2.36 Survey Results

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil and scale the axes.



Figure 2.30

3. Calculate the following values.
- a. \bar{x} = _____
 - b. s = _____
4. Are the data discrete or continuous? How do you know?

5. In complete sentences, describe the shape of the histogram.
6. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following values.
 - a. $\text{Min} = \underline{\hspace{2cm}}$
 - b. $M = \underline{\hspace{2cm}}$
 - c. $\text{Max} = \underline{\hspace{2cm}}$
 - d. $Q_1 = \underline{\hspace{2cm}}$
 - e. $Q_3 = \underline{\hspace{2cm}}$
 - f. $IQR = \underline{\hspace{2cm}}$
2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:
 - a. above the mean.
 - b. below the mean.

KEY TERMS

Box plot a graph that gives a quick picture of the middle 50% of the data

First Quartile the value that is the median of the of the lower half of the ordered data set

Frequency the number of times a value of the data occurs

Frequency Polygon looks like a line graph but uses intervals to display ranges of large amounts of data

Frequency Table a data representation in which grouped data is displayed along with the corresponding frequencies

Histogram a graphical representation in x - y form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Interquartile Range or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Interval also called a class interval; an interval represents a range of data and is used when displaying large data sets

Mean a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by \bar{x}) is

$$\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}, \text{ and the mean for a population (denoted by } \mu) \text{ is}$$

$$\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}.$$

Median a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint the mean of an interval in a frequency table

Mode the value that appears most frequently in a set of data

Outlier an observation that does not fit the rest of the data

Paired Data Set two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Percentile a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Relative Frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Skewed used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Standard Deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variance mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a

deviation can be represented as $x - \bar{x}$ where x is a value of the data and \bar{x} is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

CHAPTER REVIEW

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on y-axis with the frequency being graphed on the x-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

2.3 Measures of the Location of the Data

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile, the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3 , and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

2.4 Box Plots

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

2.5 Measures of the Center of the Data

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

2.6 Skewness and the Mean, Median, and Mode

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like **Figure 2.17**. A **left (or negative) skewed** distribution has a shape like **Figure 2.18**. A **symmetrical** distribution looks like **Figure 2.16**.

5. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in **Table 2.37**.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Table 2.37

6. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in **Table 2.38**.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

Table 2.38

7. Several children were asked how many TV shows they watch each day. The results of the survey are shown in **Table 2.39**.

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Table 2.39

56. Describe the relationship between the mode and the median of this distribution.

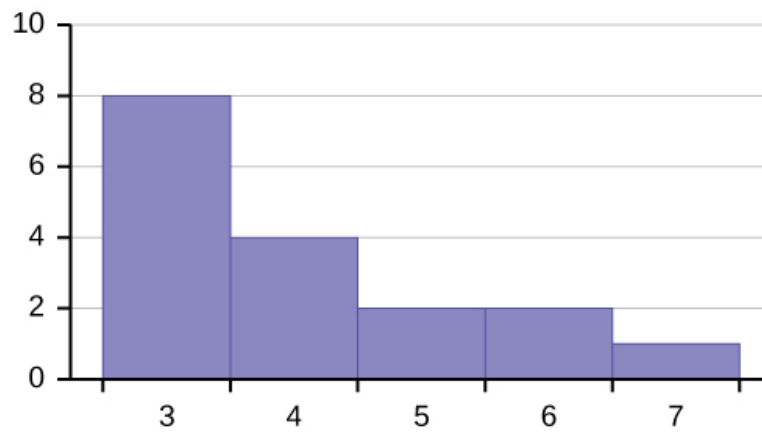


Figure 2.33

57. Describe the relationship between the mean and the median of this distribution.

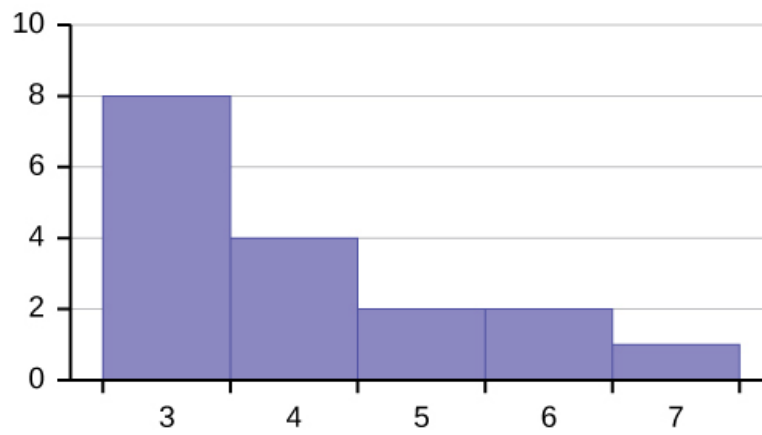


Figure 2.34

58. Describe the shape of this distribution.

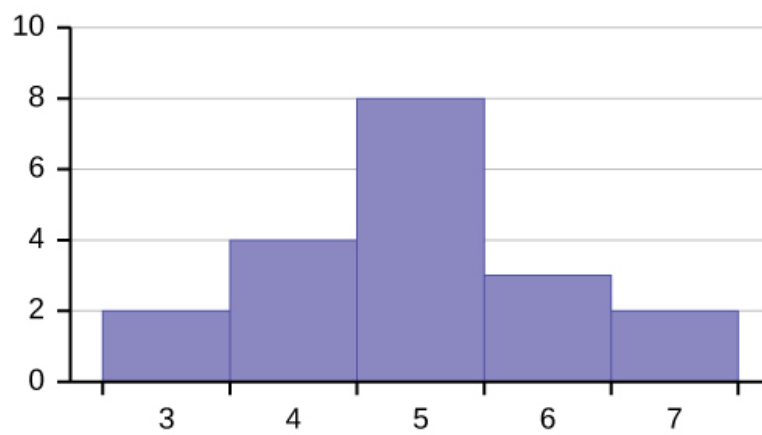


Figure 2.35

59. Describe the relationship between the mode and the median of this distribution.

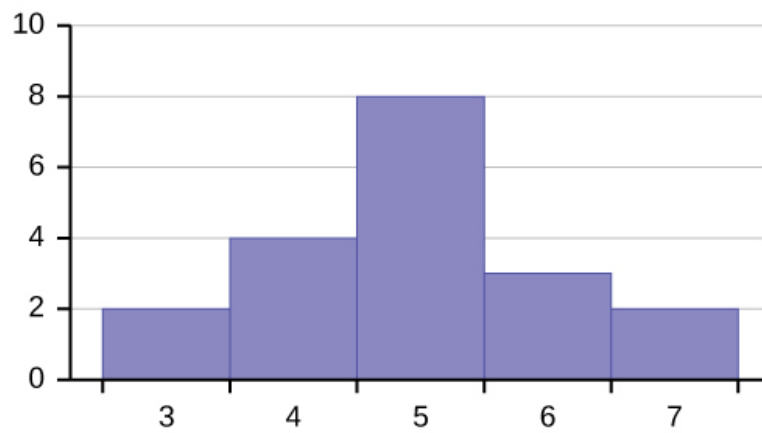


Figure 2.36

60. Are the mean and the median the exact same in this distribution? Why or why not?

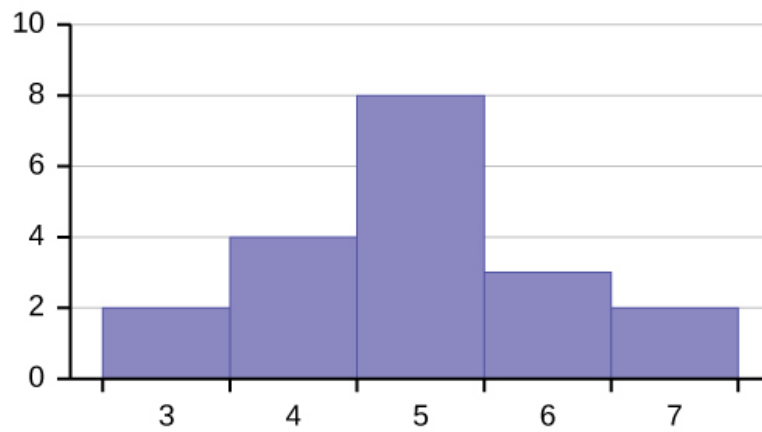


Figure 2.37

61. Describe the shape of this distribution.

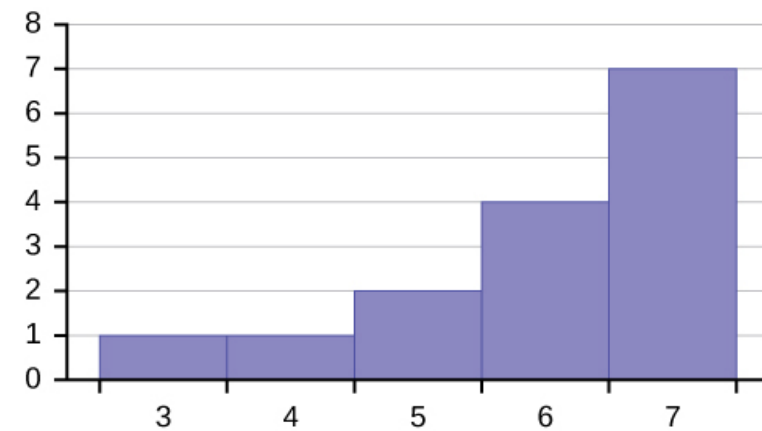


Figure 2.38