2| DESCRIPTIVE STATISTICS

Chapter 2 Table of Contents

Adapted from: Claude Laflamme, Business Statistics -- BSTA 200 -- Humber College -- Version 2016 Revision A. OpenStax CNX. Jan 10, 2020 http://cnx.org/contents/0bec2053-5fc5-49c9-a97a-c33b7e0a095c@12.342.

Introduction

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)

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.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective

way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

2.1| Frequency Distributions and Graphs

A **frequency distribution** lists categories (or intervals) and the number of times each occurs (frequency). A **relative frequency distribution** lists categories (or intervals) and the proportion of times each occurs.

$$Relative\ Frequency = \frac{class\ frequency}{total\ frequency} = \frac{f}{n}$$

Frequency Distribution for Qualitative Data

Example 2-1 Letter Grades

The letter grades in a statistics class are given as follows (raw data):

C	C	A	A	C	A	A	A	B	C	A	F	C	D	D	B	A	C	B	B
A	A	D	C	A	C	C	B	B	B	D	D	D	C	B	C	C	C	F	C

Table 2-1 Letter Grades for 40 statistics students

Construct a frequency distribution table. Construct a relative frequency distribution table.

a. How many students scored a B or better?

b. What percent of students scored a D?

c. What proportion of students scored a C or worse?

Solution 2-1

The following is a frequency distribution table that helps summarize the data in Table 2-1 above.

Grade	Number of Students, f
A	10
B	8
C	14
D	6
F	2
Total	**n = 40**

Table 2-2 Frequency distribution of the grades

Grade	Frequency	Relative Frequency
A	10	10/40 = 25%
B	8	8/40 = 20%
C	14	14/40 = 35%
D	6	6/40 = 15%
F	2	2/40 = 5%
Total	**n = 40**	**100%**

Table 2-3 Relative Frequency Distribution of Letter Grades

a. 18 students scored a B or higher. (A or B)

b. 15% scored a D.

c. 55% scored a C or worse.

Graphical Representation of Qualitative Data

Bar Chart - consists of rectangular bars showing the frequencies of each category of a distribution. The **bar chart** for the frequency distribution in Table 2-2 is seen below:

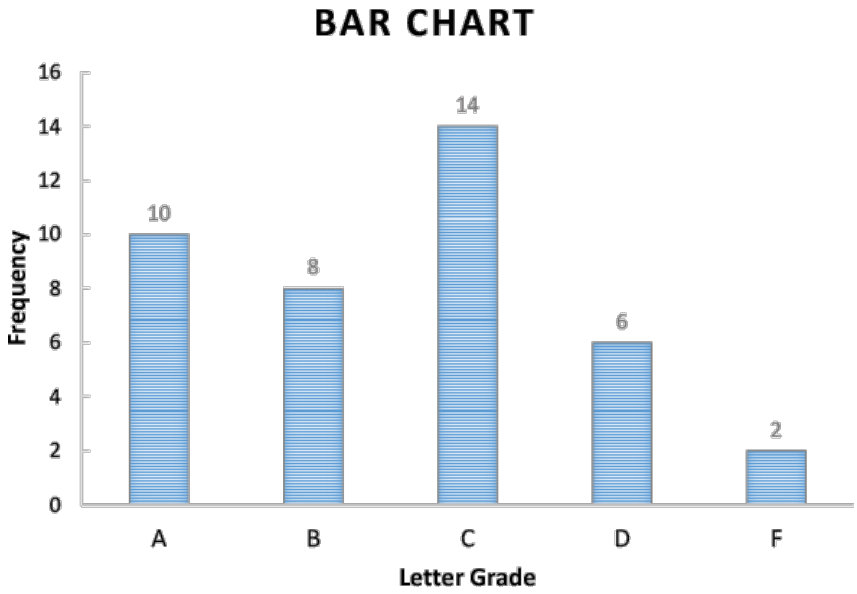


Figure 2.2 Bar Chart for Letter Grades

Pie Chart - a circle divided into sections reflecting the percentage of frequencies in each category of the distribution. The Pie Chart for Table 2-3 is seen below.

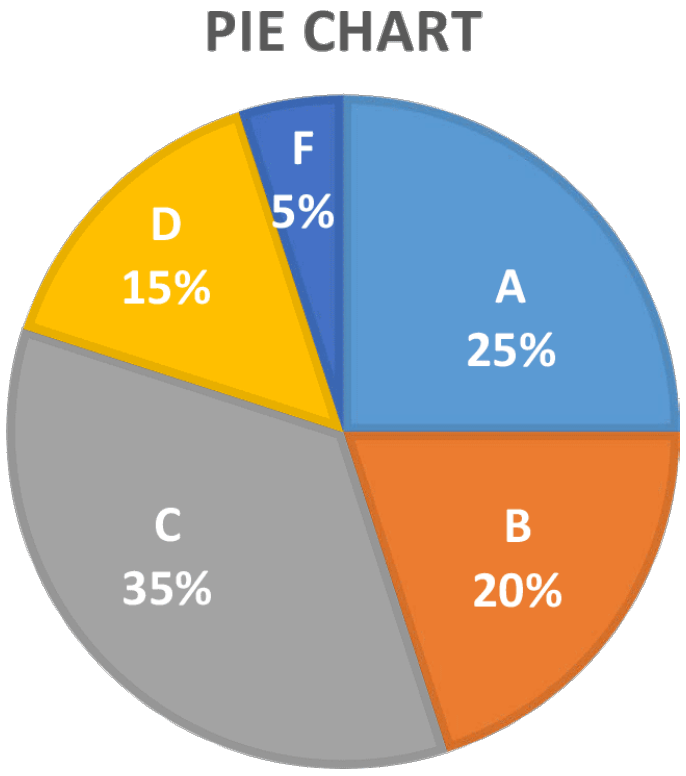


Figure 2.3 Pie Chart for Letter Grades

Frequency Distribution for Quantitative Data

Example 2-2 Age of Patients

The following are ages of the 25 patients on the 6th floor of a hospital. Construct a frequency distribution and a relative frequency distribution for these data: **12, 21, 51, 53, 26, 21, 21, 38, 55, 52, 38, 37, 40, 13, 38, 55, 31, 54, 48, 49, 54, 42, 56, 26, 54**.

Age Class	Frequency	Relative Frequency
10 to under 20	2	2/25 = 8%
20 to under 30	5	5/25 = 20%
30 to under 40	5	5/25 = 20%
40 to under 50	4	4/25 = 16%
50 to under 60	9	9/25 = 36%
Total	**n = 25**	**100%**

Table 2-4

What are the lower and upper class limits? What is the class width?

Solution 2-2

The numbers **10, 20, 30, 40, 50** are **lower limits**.

The numbers **20, 30, 40, 50, 60** are **upper limits**.

The difference between two consecutive lower (or upper) limits is called **Class Width**.

Class Width = 20 - 10 = 10 or 30 – 20 = 10 …

Note that it is more convenient to use classes with intervals widths that are multiples of 5.

NOTE
The range that captures all the data values is partitioned into five non-overlapping intervals or classes. The endpoints of each class are called class limits, lower and upper respectively, or class boundaries.

Class Midpoint

Class midpoints represent the value in the middle of the class intervals.

$$Class\ Midpoint = \frac{Lower\ Limit + Upper\ Limit}{2}$$

Age Class	Frequency	Class Midpoint
10 to under 20	2	(10 + 20)/2 = 15
20 to under 30	5	(20 + 30)/2 = 25
30 to under 40	5	(30 + 40)/2 = 35
40 to under 50	4	(40 + 50)/2 = 45
50 to under 60	9	(50 + 60)/2 = 55
Total	**n = 25**	

Steps to constructing classes for a frequency distribution

Step 1: Determine the appropriate number of classes

We can use the $2^k > n$ rule to determine the number of classes, where *k* equals the number of classes and *n* equals the number of data points.

k	3	4	5	6	7	8	9	10
2^k	8	16	32	64	128	256	512	1024

Table 2-6

From previous example, n = 25 so we use 5 classes since $2^5 = 32 > 25$. For n = 30, we see that $2^5 = 32 > 30$, so we still use 5 classes.

Table 2-5 For n = 95, $2^7 = 128 > 95$, so we use 7 classes.

Step 2: Determine the class width.

$$class\ width = \frac{highest\ value - lowest\ value}{number\ of\ classes}$$

Round up the result to the next higher integer to ensure that all of the data will be included in the frequency table. If the range of the data is large, a multiple of 5 is more convenient to use. From the given data, highest value = 56, lowest value =12, number of classes = 5. So the class width is (56 – 12)/5 = 44/5 = 8.8. Rounding this to the next integer gives 9. However, a class width of 10 is used because a multiple of 5 more convenient.

Step 3: Set the lower limit of the first class to the minimum value in the data set or a more convenient value smaller than the minimum. Then add the class width to the first lower limit to obtain the lower limit of the second class, then again for the third, and so on. The upper limits are now obvious.

Example 2-3 Years of Service

The following data represents the years of service of a sample of management employees:

35, 28, 32, 33.5, 31, 33, 34, 32, 30, 23.5, 16, 23, 19, 35, 21, 26, 31.5, 11.5, 28, 37.5, 28, 22, 19, 23, 29, 29, 28, 19, 13, 15, 12, 30, 15, 15, 26, 20, 15, 19.5, 22, 32

Use the steps described above to construct a frequency distribution for the data.

Solution 2-3

Step 1: Finding the number of classes.

From the given data, n = 40. We use the $2^k > n$ rule. Since $2^5 = 32 < 40$ and $2^6 = 64 > 40$, we use 6 classes.

Step 2: Finding the class width.

Highest value = 37.5

Lowest Value = 11.5

Number of classes = 6

Class width = $\frac{37.5-11.5}{6} = 4.33 \approx 5$ (rounded up to a convenient value)

Step 3: Construct the table.

Minimum value = 11.5 $\rightarrow$ Convenient Lower Limit of first class = 10.

Since the class width is 5, we have the following frequency distribution table.

Years	Frequency
10 to under 15	4
15 to under 20	8
20 to under 25	7
25 to under 30	8
30 to under 35	10
35 to under 40	3
Total	**40**

Table 2-7

Graphical Representation of Quantitative Data

HISTOGRAM: A histogram is a graph that displays interval data by using adjacent vertical bars to represent the frequencies. There are no spaces between the bars of a histogram unless a class frequency is zero. The frequency histogram for the Age data in Table 2-4 is shown below.

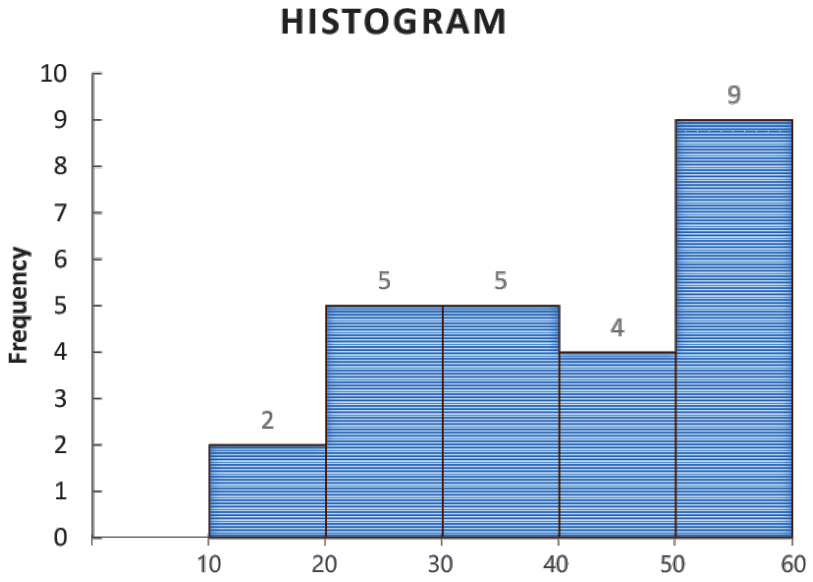


Figure 2.4

FREQUENCY POLYGON: a line plot of class frequencies against their respective class **midpoints**. The frequency polygon for the Age data in Table 2-4 is shown below.

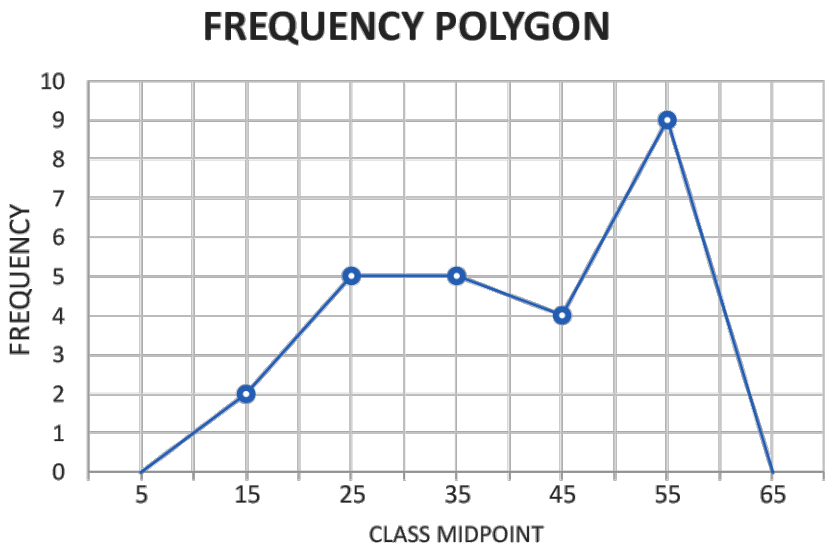


Figure 2.5

Cumulative Frequency Distribution

A **cumulative frequency distribution** shows the number (or percent) of observations that are less than or equal to each class interval. The cumulative frequency for each interval is the sum of the frequencies for that class and all previous classes. The following table shows the cumulative frequencies and relative cumulative frequencies for the Age data in Table 2-4.

Class Interval	Frequency	Cumulative Frequency	Cumulative Relative Frequency
10 to under 20	2	2	2/25 = 8%
20 to under 30	5	5 + 2 = 7	7/25 = 28%
30 to under 40	5	5 + 7 = 12	12/25 = 48%
40 to under 50	4	4 + 12 = 16	16/25 = 64%
50 to under 60	9	9 + 16 = 25	25/25 = 100%
Total	n = 25		

Table 2-8

The table shows that 28% of the patients are under 30 years old while 64% are under 50 years old.

Cumulative Frequency Graph (Ogive)

An **Ogive** (pronounced *oh-jive*) is a graphical representation of a cumulative frequency distribution. On an ogive, cumulative frequencies are plotted against class upper limits. The ogive for the preceding **table** is shown below.

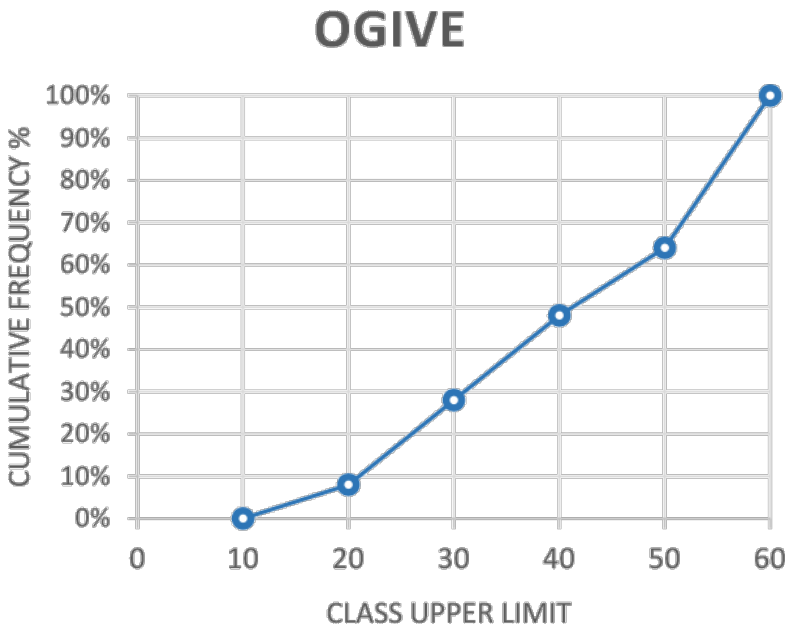


Figure 2.6

It can be seen on the ogive that about 40% of the patients are under 35 years old.

Try It 2-1: Lost Games

Canadian teams in the NHL performed poorly during the 2015-2016 season. No Canadian team made the playoffs that season. The bar graph below shows the number of regulation losses each team had.

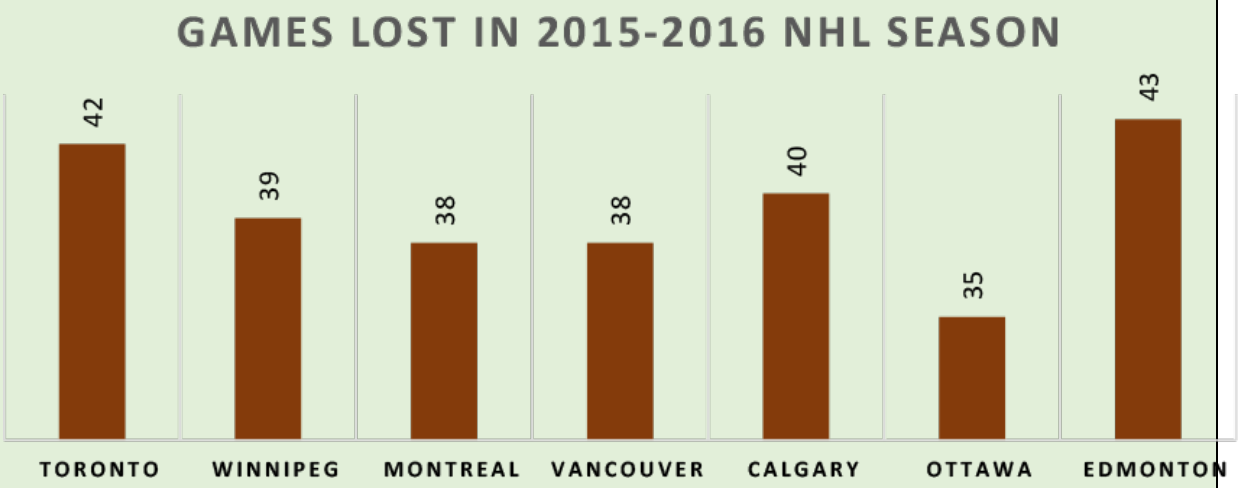


Figure 2.7

a) Which team lost the highest number of games in regulation?

b) Which team lost the least number of games in regulation?

c) How many games were lost in total by the teams?

d) What is the relative frequency of the games lost by Calgary?

e) Which of the Canadian teams performed best during the 2015-2016 season?

Chapter 2 Try It Solutions

Try It 2-2: Music Type

Some psychologists believe that the type of music you enjoy listening to tells a lot about your personality. The following pie chart shows the music preference of students at a postsecondary institution.

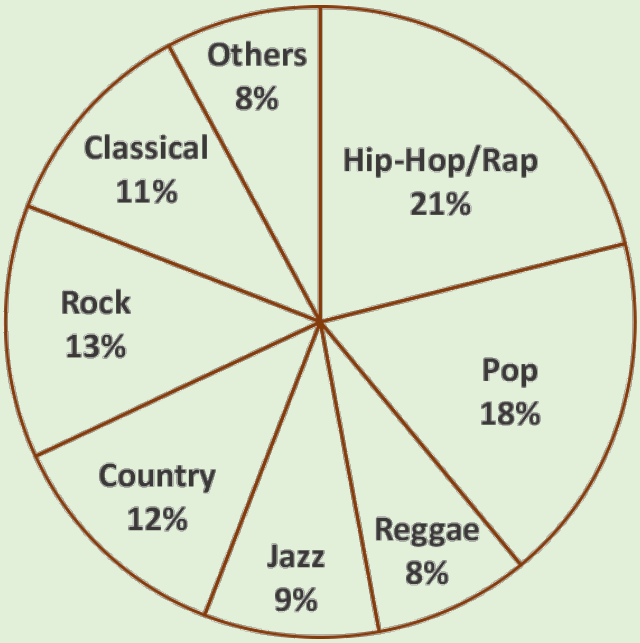


Figure 2.8

What percent of students that preferred Pop or Jazz?

a) If 32 students preferred Reggae, how many students were sampled?

b) Suppose 600 students were sampled, how many preferred Country?

c) Suppose 90 students preferred Pop, how many preferred Rock?

d) What is the size of the angle (degrees) representing the Classical sector?

Chapter 2 Try It Solutions

Try It 2-3: Exercise Hours

It is generally believed that physical exercises help prevent excess weight gain, help prevent diseases, and so on. To understand the state of physical activities of his students, a fitness instructor collected data on the number of hours per week her students exercise. The data are summarized in the histogram below.

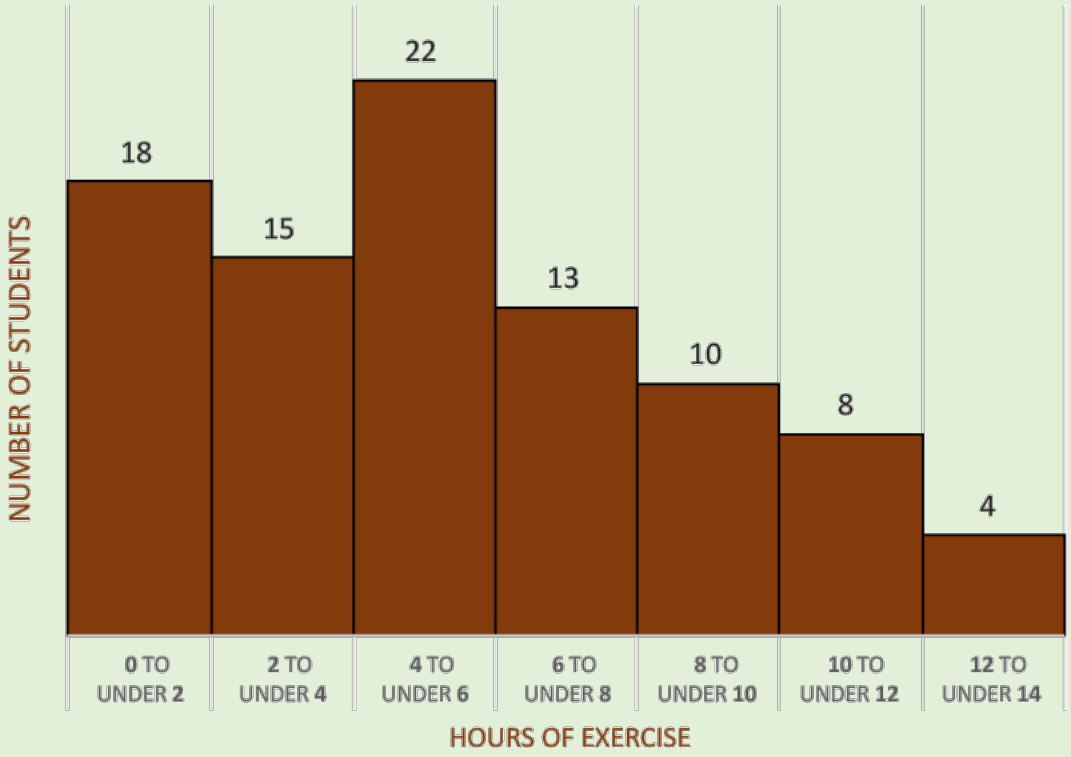


Figure 2.9

a) The fitness instructor collected data from how many students?

b) How many class intervals were used?

c) Based on the 2^k rule, are the number of classes appropriate? Explain.

d) What is the width of each class interval?

e) How many students exercise for less than 4 hours per week?

f) What percent of the students exercise between 4 and 10 hours per week?

g) What is the class midpoint for those that exercise from 6 to under 8 hours?

Chapter 2 Try It Solutions

2.2| Measures of the Location of the Data

Percentiles

Percentiles divide ordered data into hundredths. That is, 100 equal parts. There are 99 percentiles, which we denote by $P_1, P_2, P_3, \ldots P_{99}$. To score in the 90th percentile of an exam does not mean that you received 90% on a test. It means that 90% of test scores are less than your score and 10% of the test scores are 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.

A Formula for Finding the pth Percentile

Suppose we would like to find the pth percentile. We must first find the **location** of the pth percentile.

Location of a Percentile

$$\boldsymbol{L_p = (n+1)\frac{p}{100}}$$

where L_p is the location of the pth percentile when the data is ordered from **smallest to largest**, and n is the total number of observations.

To find the pth percentile

- Order the data from smallest to largest.
- Calculate L_p.
- If L_p is an integer, then the pth percentile is the data value in the L_p position in the ordered set of data.
- If L_p is not an integer, then we must calculate the pth percentile as outlined in the examples below.

Example 2-4

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a) Find the 70th percentile.

b) Find the 83rd percentile.

Solution 2-4

a) $p = 70,\ n = 29$. The percentile locator formula is $L_p = (n+1)\frac{p}{100}$. The locator for the 70th percentile is

$$L_{70} = (29+1)\left(\frac{70}{100}\right) = 21$$

Since 21 is an integer, and the data value in the 21st position in the ordered data set is 64, the 70th percentile is 64 years. Or $P_{70} = 64$.

b) $p = 83,\ n = 29$. The locator for the 83rd percentile is

$$L_{70} = (29+1)\left(\frac{83}{100}\right) = 24.9$$

Since 24.9 is NOT an integer, the 83rd percentile is 0.9 (90%) of the distance between 24th position and the 25th position.

The age in the 24th position is 71 and the age in the 25th position is 72.

Find the **difference** between these two values: 72 - 71 = 1.

Multiply the difference by the decimal part of the location: 0.9(1) = 0.9.

Add this result to the smaller value (in the 24th position): 71 + 0.9 = 71.9.

The 83rd percentile is 71.9 years. That is, P83 = 71.9.

Try It 2-4

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a) Calculate the 20th percentile.

b) Calculate the 55th percentile.

Chapter 2 Try It Solutions

Quartiles

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, Q_2, and the 50th percentile.

To calculate quartiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into four equal parts.

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.

$$Median = Q_2 = \frac{6.8 + 7.2}{2} = 7$$

The median or **second quartile** is seven. Half of the values are smaller than seven and half of the values are larger than seven.

To find the **first quartile**, we use our location of a percentile formula. The first quartile $Q_1 = P_{25}$, so we find L_{25}.

$$L_{25} = (14 + 1)\left(\frac{25}{100}\right) = 3.25$$

The first quartile falls between the 3rd and 4th sorted values. The 3rd value is 2 and the 4th value is also 2. Since these are the same value, we do not need to perform any calculations.

The first quartile $Q_1 = 2$.

The **third quartile** $Q_3 = P_{75}$, so we find L_{75}.

$$L_{75} = (14 + 1)\left(\frac{75}{100}\right) = 11.25$$

The third quartile is therefore between the 11th and 12th sorted values. The 11th value is 9 and the 12th value is 10. The distance between the 11th and 12th values is $10 - 9 = 1$.

0.25 of the distance is $0.25(1) = 0.25$.

The third quartile $Q_3 = 9.25$.

The **interquartile range** $\boldsymbol{IQR}$ 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. From the preceding example, $Q_1 = 2$ and $Q_3 = 9.25$.

Therefore,

$$IQR = Q_3 - Q_1 = 9.25 - 2 = 7.25$$

The range of the middle 50% of the values is 7.25.

Example 2-5

For the following 13 real estate prices, find the quartiles and calculate the IQR.

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 2-5

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

There are 13 data points.

Since there are 13 observations, the median is the 7th (middle) value in the sorted data.

$$Median = Q_2 = 488{,}800$$

For the first quartile, the locator is $L_{25} = (13 + 1)\left(\frac{25}{100}\right) = 3.5$. The first quartile is 0.5 (midway) of the distance between the 3rd and 4th sorted values.

The distance between the 3rd and 4th values is 387, 000 − 230, 500 = 156, 500.

0.5 of the distance gives 0.5(156, 500) = 78, 250.

So, the value midway the 3rd and 4th values is 230, 500 + 78, 250 = 308, 750.

The first quartile $Q_1 = 308{,}750$.

For the third quartile, the percentile locator is $L_{75} = (13 + 1)\left(\frac{75}{100}\right) = 10.5$. The distance between the 10th and 11th values is 659,000 - 639,000 = 20,000.

0.5 of the distance gives $0.5(20{,}000) = 10{,}000$.

The third quartile $Q_3 = 639{,}000 + 10{,}000 = 649{,}000$.

The interquartile range $IQR = Q_3 - Q_1 = 649{,}000 - 308{,}750 = 340{,}250$.

Try It 2-5

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

Chapter 2 Try It Solutions

VIDEO
Watch a video on calculating Percentiles and Quartiles here.

Outliers

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.

Potential outliers are values below $Q_1 - 1.5(IQR)$ or above $Q_3 + 1.5(IQR)$.

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 2-6

Calculate the IQR and determine if there are potential outliers for the following 11 scores. 20; 6; 15; 10; 45; 18; 25; 22; 19; 37; 13

Solution 2-6

Order the data from smallest to largest.

6; 10; 13; 15; 18; 19; 20; 22; 25; 37; 45

$$L_{25} = (11 + 1)\left(\frac{25}{100}\right) = 3rd$$

$$Q_1 = 13$$

$$L_{75} = (11 + 1)\left(\frac{75}{100}\right) = 9th$$

$$Q_3 = 25$$

$$IQR = Q_3 - Q_1 = 25 - 13 = 12$$

$$1.5(IQR) = 1.5(12) = 18$$

$$Q_1 - 1.5(IQR) = 13 - 18 = -5$$

$$Q_3 + 1.5(IQR) = 25 + 18 = 43$$

No score is less than -5. However, 45 is more than 43. Therefore, 45 is a potential **outlier**.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15^{th} percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example 2-7

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution 2-7

- Twenty-five percent of students finished the exam under 35 minutes.
- Seventy-five percent of students finished the exam in over 35 minutes.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 2-8

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 2-8

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

Try It 2-6

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.

Chapter 2 Try It Solutions

Example 2-9

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 2-9

- 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.

Boxplot and the 5-Number Summary

A **boxplot** is a graphical display, based on quartiles, that helps picture the distribution of a data set. The following values are needed to construct a box plot. They are also referred to as the **5-number summary**.

1. Minimum value (lowest value)
2. First quartile (Q_1)
3. Median (Q_2)
4. Third quartile (Q_3)
5. Maximum value (highest value)

A boxplot divides data into quartiles (see figure below). It consists of a rectangle in the middle, bounded on the left (bottom) by the first quartile (Q_1), and on the right (top) by the third quartile (Q_3). A vertical line is also drawn within the box representing the median. Two horizontal lines, called whiskers, extend from the bottom and top of the box. The bottom whisker extends from Q_1 to the smallest non-outlier in the data set, and the top whisker extends from Q_3 to the largest non- outlier. Outliers (if they exist) are plotted separately as points (or asterisks) on the chart.

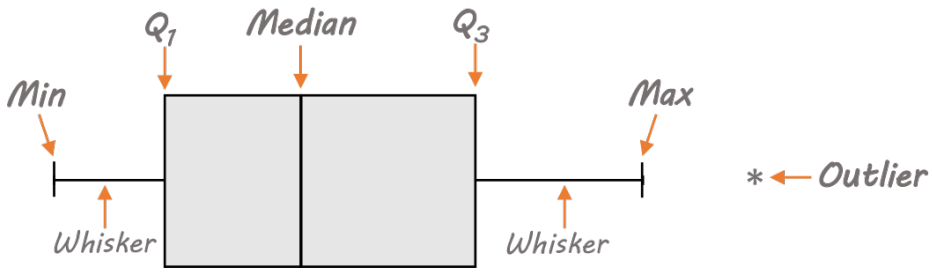


Figure 2.10 Boxplot

Example 2-10

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 number of exercise minutes from the 15 anonymous students are shown below.

50; 10; 45; 90; 0; 100; 20; 30; 30; 40; 0; 60; 120; 70; 30

Determine the 5-number summary and construct a boxplot for the data.

Solution 2-10

$$Min = 0$$

$$L_{25} = (15 + 1)\left(\frac{25}{100}\right) = 4th$$

$$Q_1 = 20$$

$$Median = 40$$

$$L_{75} = (15 + 1)\left(\frac{75}{100}\right) = 12th$$

$$Q_3 = 70$$

$$Max = 120$$

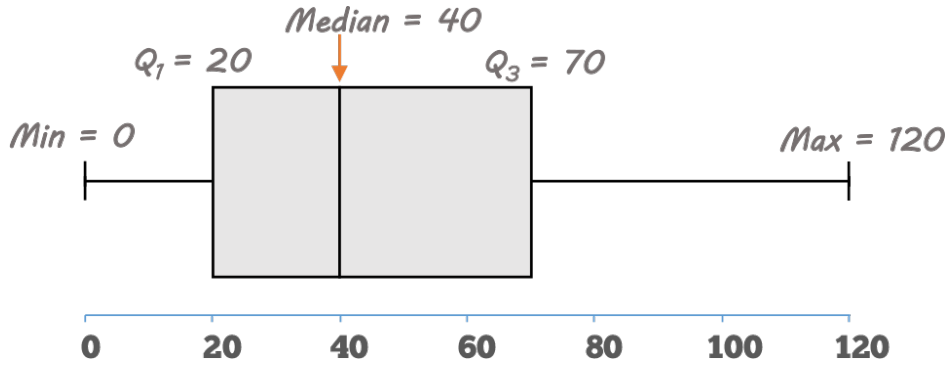


Figure 2.11 Boxplot for Exercise Minutes Data

Since 75% of the students exercise for 70 minutes or less daily, and since the *IQR* is 50 minutes (70 – 20 = 5), we know that half of the students surveyed exercise between 20 minutes and 70 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

2.3| 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. Technically this is the arithmetic mean. We will discuss the geometric mean later. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts meaning an equal number of observations on each side. The weight of 25 people are below this weight and 25 people are heavier than this weight. 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. Formally, the arithmetic mean is called the first moment of the distribution by mathematicians. However, in practice among non- statisticians, "average" is commonly accepted for "arithmetic mean."

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.

Median

You can quickly find the location of the median by using the expression:

$$Location\ of\ median = \frac{n+1}{2}$$

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. We average the 50th and 51st values to find the median.

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.

Try It 2-7

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.

Chapter 2 Try It Solutions

Try It 2-8

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?

Chapter 2 Try It Solutions

Mode

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.

Example 2-11

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Solution 2-11

The most frequent score is 72, which occurs five times. Mode = 72.

Example 2-12

Five real estate exam scores are 430, 430, 480, 480, 495. What is the mode?

Solution 2-12

The data set is bimodal because the scores 430 and 480 each occur twice. Mode = 430, 480.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE
The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Calculating the Arithmetic 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:

$$Mean = \frac{data\ sum}{number\ of\ values}$$

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:

$$Midpoint = \frac{Lower\ Limit + Upper\ Limit}{2}$$

We can now modify the mean definition to be

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$$

where f = frequency of the interval and m = midpoint of the interval.

Example 2-13

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 2-9

Solution 2-13

Find the midpoints for all intervals

Grade interval	Midpoint	Number of students
50–56.5	53.25	1
56.5–62.5	59.5	0
62.5–68.5	65.5	4
68.5–74.5	71.5	4
74.5–80.5	77.5	2
80.5–86.5	83.5	3
86.5–92.5	89.5	4
92.5–98.5	95.5	1

Table 2-10

Calculate the sum of the product of each interval frequency and midpoint

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

$$\sum f = 1 + 0 + 4 + 4 + 2 + 3 + 4 + 1 = 19$$

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

Try It 2-9

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 2-11

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

Chapter 2 Try It Solutions

2.4| Sigma Notation and Calculating the Arithmetic Mean

Formula for Population Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Formula for Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This unit is here to remind you of material that you once studied and said at the time "I am sure that I will never need this!"

Here are the formulas for a population mean and the sample mean. The Greek letter μ is the symbol for the population mean and $\bar{x}$ is the symbol for the sample mean. Both formulas have a mathematical symbol that tells us how to make the calculations. It is called Sigma notation because the symbol is the Greek capital letter sigma: ∑. Like all mathematical symbols it tells us what to do: just as the plus sign tells us to add and the x tells us to multiply. These are called mathematical operators. The ∑ symbol tells us to add a specific list of numbers.

Let's say we have a sample of animals from the local animal shelter and we are interested in their average age. If we list each value, or observation, in a column, you can give each one an index number. The first number will be number 1 and the second number 2 and so on.

Animal	Age
1	9
2	1
3	8.5
4	10.5
5	10
6	8.5
7	12
8	8
9	1
10	9.5

Table 2-12

Each observation represents a particular animal in the sample. Purr is animal number one and is a 9 year old cat, Toto is animal number 2 and is a 1 year old puppy and so on.

To calculate the mean we are told by the formula to add up all these numbers, ages in this case, and then divide the sum by 10, the total number of animals in the sample.

Animal number one, the cat Purr, is designated as X_1, animal number 2, Toto, is designated as X_2 and so on through Dundee who is animal number 10 and is designated as X_{10}.

The i in the formula tells us which of the observations to add together. In this case it is X_1 through X_{10} which is all of them. We know which ones to add by the indexing notation, the i = 1 and the n or capital N for the population. For this example the indexing notation would be i = 1 and

because it is a sample we use a small n on the top of the Σ which would be 10.

The standard deviation requires the same mathematical operator and so it would be helpful to recall this knowledge from your past.

The sum of the ages is found to be 78 and dividing by 10 gives us the sample mean age as 7.8 years.

2.5| Skewness and the Mean, Median, and Mode

Consider the following data set.

4; 5; 6; 6; 6; 7; 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.

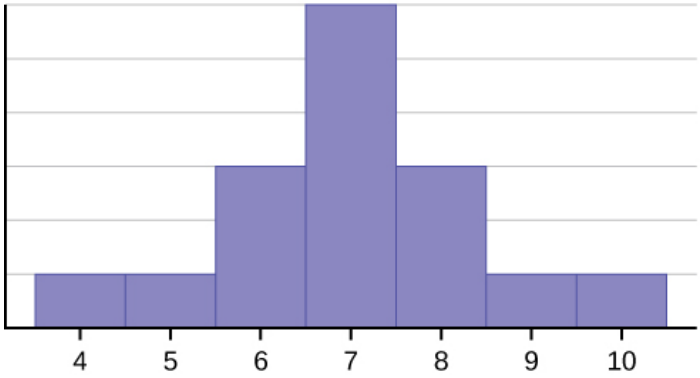


Figure 2.12

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 shown in Figure 2.12 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. We can formally measure the skewness of a

distribution just as we can mathematically measure the center weight of the data or its general "spreadness". The mathematical formula for skewness is:

$$a_3 = \frac{\sum(x_i - \bar{x})^3}{ns^3}$$

The greater the deviation from zero indicates a greater degree of skewness. If the skewness is negative then the distribution is skewed left as in Figure 2.13. A positive measure of skewness indicates right skewness such as in Figure 2.14.

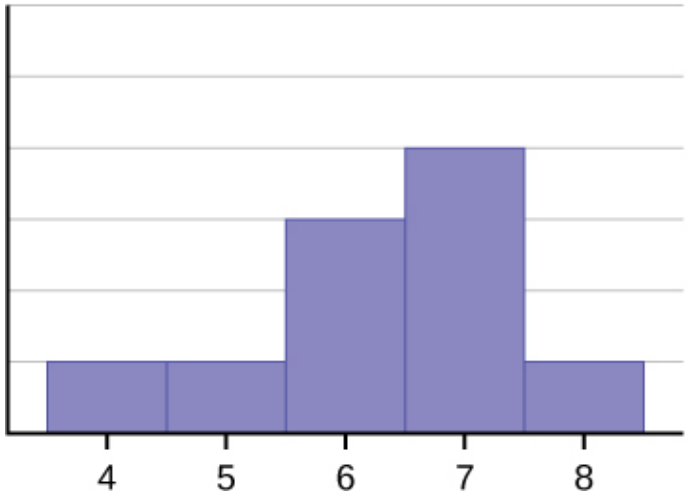


Figure 2.13

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10 shown in Figure 2.13, is also not symmetrical. It is **skewed to the right**.

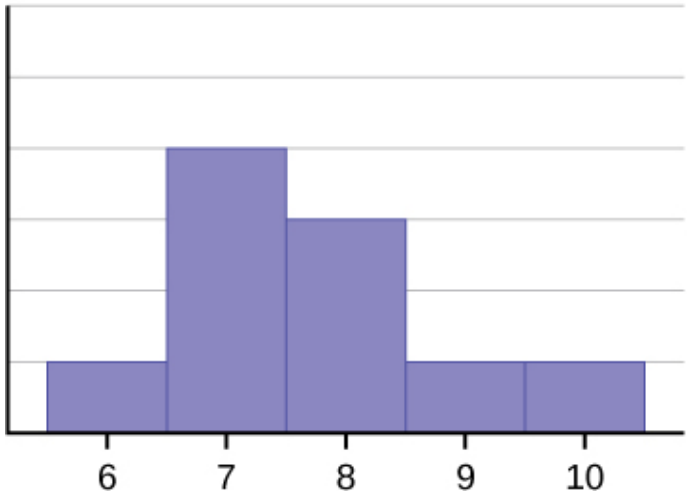


Figure 2.14

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

As with the mean, median and mode, and as we will see shortly, the variance, there are mathematical formulas that give us precise measures of these characteristics of the distribution of the data. Again, looking at the formula for skewness we see that this is a relationship between the mean of the data and the individual observations cubed.

$$a_3 = \frac{\sum(x_i - \bar{x})^3}{ns^3}$$

Where s is the sample standard deviation of the data, X_i, and $\bar{x}$ is the arithmetic mean and n is the sample size.

Formally the arithmetic mean is known as the first moment of the distribution. The second moment we will see is the variance, and skewness is the third moment. The variance measures the squared differences of the data from the mean and skewness measures the cubed differences of the data from the mean. While a variance can never be a negative number, the measure of skewness can and this is how we determine if the data are skewed right of left. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the

skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean and standard deviation are dimensional quantities (this is why we will take the square root of the variance) that is, have the same units as the measured quantities X_i, the skewness is conventionally defined in such a way as to make it nondimensional. It is a pure number that characterizes only the shape of the distribution. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive X and a negative value signifies a distribution whose tail extends out towards more negative X. A zero measure of skewness will indicate a symmetrical distribution.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

2.6| 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. There are several ways to measure variation. Three commonly used measures of variation are the **range**, the **variance** and the **standard deviation.**

The Range

The **range** is the simplest measure of variation. The range measures the difference between the highest and lowest values in a data set.

$$Range = Highest\ Value - Lowest\ Value$$

Example 2-14

Find the range of the following workout times (mins): 13.3, 21.1, 12.1, 25.5, 28.9, 26, 16.8, 13.6, 27.6, 15.4, 18.4

Solution 2-14

$$Highest\ Value = 28.9$$

$$Lowest\ Value = 12.1$$

$$\boldsymbol{Range = 28.9 - 12.1 = 16.8}$$

While the range gives us some idea of how spread out the data is, it tells us nothing about how data is dispersed within the range, whether it is compact or spread out or clustered around certain values.

The **standard deviation** is a more common measure of variation that takes all of the data points into consideration.

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 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. The standard deviation is always non-negative.

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.

Calculating the Standard Deviation

If *x* is a number, then the difference "*x* – 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 $\boldsymbol{\sigma^2}$ represents the **population variance**; the population standard deviation σ is the square root of the population variance. The symbol $\boldsymbol{s^2}$ 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.

$$standard\ deviation = \sqrt{variance}$$

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 Sample Variance

The two formulas shown below will give the same results and either formula may be used.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

Formulas for Sample Standard deviation

The standard deviation is found by taking the square root of the variance. Either of the formulas below will give identical results.

$$s^2 = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$s^2 = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

Formula for the Population Variance

The steps in calculating the population variance are nearly identical to the steps involved in calculating the sample variance. Notice one important difference. When calculating the sample variance, we divide by $n-1$ (sample size minus 1) and when calculating the population variance, we simply divide by N (sample size).

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

Formula for the Population Standard Deviation

To find the population standard deviation, we take the square root of the population variance.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Two important observations concerning the variance and standard deviation: the deviations are measured from the mean and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data, what is the "normal" or most usual value of the observation. Later we will be trying to measure the "unusualness" of an observation or a sample mean and thus we need a measure from the mean. The second observation is that the deviations are squared. This does two things, first it makes the deviations all positive and second it changes the units of measurement to that of the mean and the original observations. If the data are weights then the mean is measured in pounds, but the variance is measured in pounds-squared. One reason to use the standard deviation is to return to the original units of

measurement by taking the square root of the variance. Further, when the deviations are squared it explodes their value. For example, a deviation of 10 from the mean when squared is 100, but a deviation of 100 from the mean is 10,000. What this does is place great weight on outliers when calculating the variance.

Example 2-15

In a fifth grade class, a teacher has noticed that the heights of the girls in her class seems highly variable. To explore this, she took the height measurement of each of the 8 girls in her class. The following data is the height in centimetres for a SAMPLE of n = 8 fifth grade girls:

135; 147; 139; 126; 141; 153; 142; 149

Calculate the sample variance and sample standard deviation.

Solution 2-15

$$\bar{x} = \frac{135 + 147 + 139 + 126 + 141 + 153 + 142 + 149}{8} = 141.5$$

The average height is 141.5 cm.

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.

x	$x - \bar{x}$	$(x - \bar{x})^2$
135	135-141.5 = -6.5	$(-6.5)^2$ = 42.25
147	147-141.5 = 5.5	$(5.5)^2$ = 30.25
139	139 - 141.5 = -2.5	$(-2.5)^2$ = 6.25
126	126 - 141.5 = -15.5	$(-15.5)^2$ = 240.25
141	141 - 141.5 = -0.5	$(-0.5)^2$ = 0.25
153	153 - 141.5 = 11.5	$(11.5)^2$ = 132.25
142	142 - 141.5 = 0.5	$(0.5)^2$ = 0.25
149	149 - 141.5 = 7.5	$(7.5)^2$ = 56.25
	Total	508

Table 2-13

The sample variance, s^2, is equal to the sum of the last column (508) divided by the total number of data values minus one (8 - 1):

$$s^2 = \frac{508}{8-1} = 72.57142857$$

The sample variance is 72.57 cm^2, rounded to 2 decimal places.

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

$$s = \sqrt{72.57142857} = 8.5189$$

The sample standard deviation, rounded to 2 decimal places, is 8.52 cm.

VIDEO
Here is a video on how to use the BAII Plus calculator to obtain the mean, standard deviation and variance.

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 135 is farther from the mean than is the data value 141 which is indicated by the deviations -6.5 and -0.5. 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. **If you add the deviations, the sum is always zero**. (In the example above, there are *n* = 8 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. By squaring the deviations we are placing an extreme penalty on observations that are far from the mean; these observations get greater weight in the calculations of the variance. We will see later on that the variance (standard deviation) plays the critical role in determining our conclusions in inferential statistics. We can begin now by using the standard deviation as a measure of "unusualness." "How did you do on the test?" "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

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* = 8, the calculation divided by *n* – 1 = 8 – 1 = 7 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.** This estimate requires us to use an estimate of the population mean rather than the actual population mean. Based on the theoretical mathematics that lies behind these calculations, dividing by (n – 1) gives a better estimate of the population variance.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, the 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.

Try It 2-10 Prof Tan Le's Class

Use the following data (final exam scores) from Prof. Tan Le's summer statistics class:

42, 56, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to 2 decimal places.

b. Calculate the following to one decimal place:

i. The sample mean

ii. The median

iii. The mode

iv. The range

v. The sample variance

vi. The sample standard deviation

vii. The first quartile

viii. The third quartile

ix. IQR

Chapter 2 Try It Solutions

The Empirical Rule

For a symmetric unimodal (**bell-shaped or normal**) distribution, the Empirical Rule states that:

Approximately 68% of the data is within one standard deviation of the mean, $\bar{x} \pm s$ or $\bar{x} \pm \sigma$.

Approximately 95% of the data is within two standard deviations of the mean, $\bar{x} \pm 2s$ or $\bar{x} \pm 2\sigma$.

Approximately 99.7% of the data is within three standard deviations of the mean, $\bar{x} \pm 3s$ or $\bar{x} \pm 3\sigma$.

Example 2-16

Suppose a distribution is bell-shaped with a mean of **25** and a standard deviation **4**.

a) About 68% of the data values lie between _____ and _____.
b) About 95% of the data values lie between _____ and _____.
c) About 99.7% of the data values lie between _____ and _____.

Solution 2-16

a) About 68% lie between 25 – 4 and 25 + 4 → **(between 21 and 29)**

b) About 95% lie between 25 – 2(4) and 25 + 2(4) → **(between 17 and 33)**

c) About 99.7% lie between 25 – 3(4) and 25 + 3(4) → **(between 13 and 37)**

Video
The empirical rule is also known as the 68-95-99.7 rule. See video.

Example 2-17

The scores on an exam are a normally distributed with a mean of 15 and a standard deviation of 3.

a) Approximately what percent of the scores are Between 9 and 21?

b) Approximately what percent of the scores are less than 6?

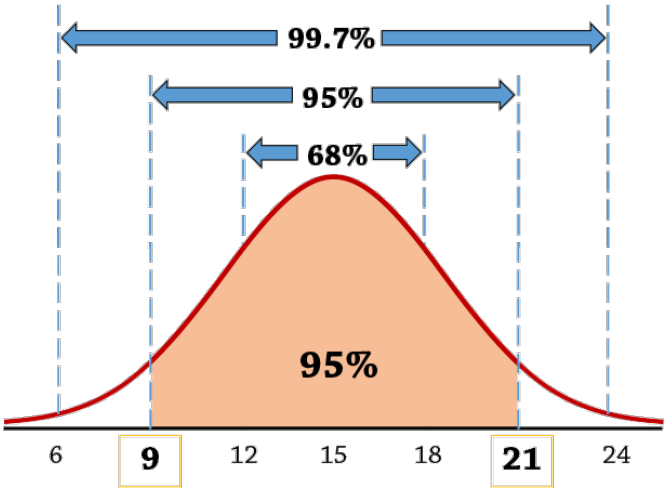
c) Approximately what percent of scores are below 21?

d) Approximately what percent of the scores are below 21?

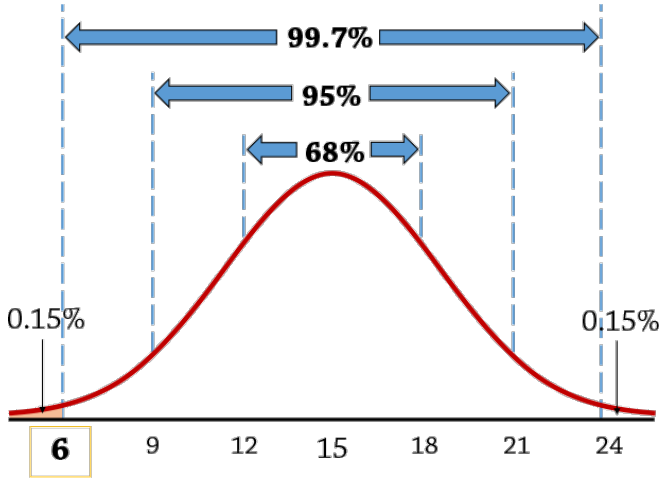
e) Approximately what percent of scores are between 12 and 24?

Solution 2-17

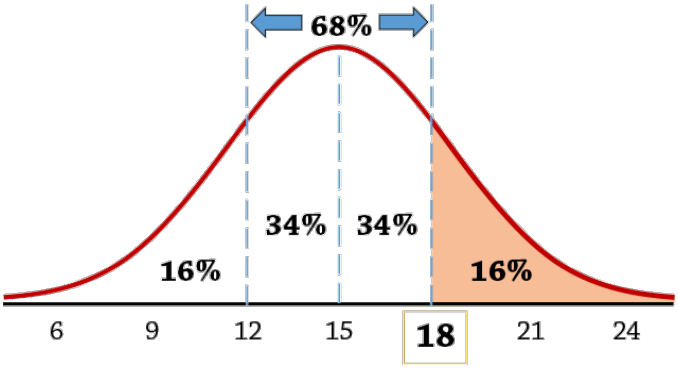
a) Approximately 95%.



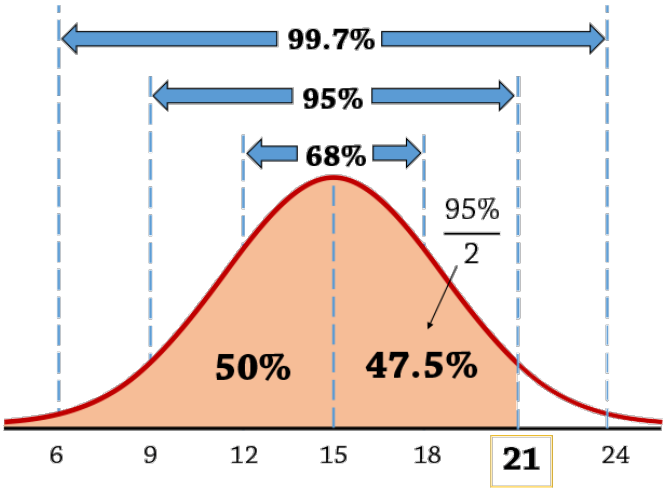
b) Approximately 0.15%.



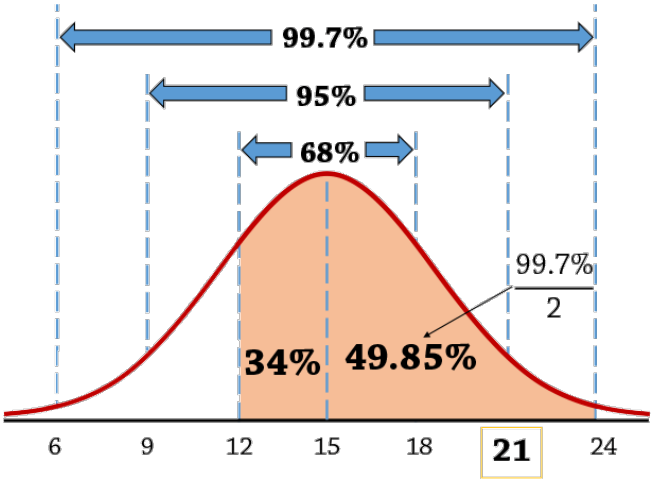
c) Approximately 16%.



d) Approximately 97.5%.



e) Approximately 83.5%.



Coefficient of Variation

Another useful way to compare distributions besides simple comparisons of means or standard deviations is to adjust for differences in the scale of the data being measured. Quite simply, a large variation in data with a large mean is different than the same variation in data with a small mean. To adjust for the scale of the underlying data the Coefficient of Variation (CV) has been developed. Mathematically:

Sample Coefficient of Variation

$$CV = \frac{s}{\bar{x}}$$

Population Coefficient of Variation

$$CV = \frac{\sigma}{\mu}$$

We can see that this measures the variability of the underlying data as a percentage of the mean value; the center weight of the data set. This measure is useful in comparing risk where an adjustment is warranted because of differences in scale of two data sets. In effect, the scale is changed to common scale, percentage differences and allows direct comparison of the two or more magnitudes of variation of different data sets.

Example 2-18

Two students, Pete and Ahmed, from different colleges, wanted to find out whose school has a higher GPA variability. They both agree that too much variability in GPAs is bad and want to know whose school is performing better in that respect. Here is a summary of data from their schools.

	Standard Deviation of GPA	**Mean GPA**
Pete's school	9.0	78
Ahmed's school	8.6	70

Table 2-14

Solution 2-18

We compute and compare the coefficient of variation for each school.

Pete's School: $CV = \frac{9.0}{78} = 0.115 \; or \; 11.5\%$

Ahmed's School: $CV = \frac{8.6}{70} = 0.123 \; or \; 12.3\%$

Although the standard deviation is higher for Pete's school, Ahmed's school has more GPA variability relative to the mean.

Chapter 2 Try It Solutions

Try It 2-1

a) Edmonton

b) Ottawa

c) 42+39+38+38+40+35+43 = 275

d) 40/275 = 14.55%

e) Ottawa

Try It 2-2

a) 18+9 = 27

b) 32/0.08 = 400

c) 0.12*600 = 72

d) 90/0.18*0.13 = 65

e) 0.11*360 = 39.6°

Try It 2-3

a) 18+15+22+13+10+8+4 = 90

b) 7

c) Yes. 27 = 128 > 90

d) 2

e) 18+15 = 33

f) (22+13+10)/90 = 50%

g) (6 + 8)/2 = 7 hours

Try It 2-4

a) $p = 20, \quad n = 29$. The locator for the 20th percentile is

$$L_{20} = (29 + 1)\left(\frac{20}{100}\right) = 6$$

The age in the sixth position is 27.

The 20th percentile is 27 years. Or, $P_{20} = 27$.

b) $p = 55, \quad n = 29$. The locator for the 55th percentile is

$$L_{55} = (29 + 1)\left(\frac{55}{100}\right) = 16.5$$

The age in the 16th position is 52 and the age in the 17th position is 55.

The 55th percentile is 0.5 (50%) of the distance between 52 and 55.

The distance between the 16th and 17th position is $55 - 52 = 3$.
0.5 of the distance between the 16th and 17th position
is 0.5(3) = 1.50.5(3) = 1.5.
The 55th percentile is therefore $52+1.5 = 53.5$.
The 55th percentile is **53.5 years**.

Try It 2-5

Class A

Data sorted from smallest to largest:

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

There are 20 data points.

First Quartile

Locator for the first quartile: $L_{25} = (20 + 1)\left(\frac{25}{100}\right) = 5.25$.

The first quartile is 0.25 of the distance between the 5th and 6th data points.

$$Q_1 = P_{25} = 69 + 0.25(76 - 69) = 70.75$$

Median

Since there are 20 data points, the median is between the 10th and 11th data points. The 10th is 80 and the 11th is 81. Therefore, the median is:

$$Q_2 = \frac{80 + 81}{2} = 80.5$$

Third Quartile

Locator for the third quartile: $L_{25} = (20 + 1)\left(\frac{75}{100}\right) = 15.75$.

The third quartile is 0.75 of the distance between the 15th and 16th data points.

$$Q_3 = P_{75} = 90 + 0.75(91 - 90) = 90.75$$

The Interquartile Range

$$IQR = Q_3 - Q_1 = 90.75 - 70.75 = 20$$

Class B

Data sorted from smallest to largest, n = 20.

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

First Quartile

Locator for the first quartile: $L_{25} = (20 + 1)\left(\frac{25}{100}\right) = 5.25$.

The first quartile is 0.25 of the distance between the 5th and 6th data points.

$$Q_1 = P_{25} = 72 + 0.25(73 - 72) = 72.25$$

Median

Both the 10th and 11th values are 80.

$$Q_2 = \frac{80 + 80}{2} = 80$$

Third Quartile

Locator for the third quartile: $L_{25} = (20 + 1)\left(\frac{75}{100}\right) = 15.75$.

The third quartile is 0.75 of the distance between the 15th and 16th data points.

$$Q_3 = P_{75} = 92 + 0.75(95 - 92) = 94.25$$

The Interquartile Range

$$IQR = Q_3 - Q_1 = 94.25 - 72.25 = 22$$

The data for Class B has a larger IQR, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class B are more spread out and not clustered about the median.

Try It 2-6

Eighty percent of students earned less than 49 points. Twenty percent of students earned more than 49 points. A higher percentile is good because getting more points on an assignment is desirable.

Try It 2-7

The calculation for the mean is:

$$\bar{x} = \frac{3 + 4 + 8 + 8 + \cdots + 44 + 47}{40} = 23.6$$

To find the median 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

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

Try It 2-8

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

$$Median = 30{,}000$$

(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 2-9

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers.
The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.
Mean = (1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9) = 409.75/38 = 10.78

Try It 2-10

a. See Table 2-15.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
42	1	0.04	0.04
56	1	0.04	0.08
63	1	0.04	0.12
67	1	0.04	0.16
68	2	0.08	0.24
69	2	0.08	0.32
72	1	0.04	0.36
73	1	0.04	0.4
74	1	0.04	0.44
78	1	0.04	0.48
80	1	0.04	0.52
83	1	0.04	0.56
88	3	0.12	0.68
90	1	0.04	0.72
92	1	0.04	0.76
94	4	0.16	0.92
96	1	0.04	0.96
100	1	0.04	1
Total	**25**	**1**	**---**

Table 2-15

b.

i. The sample mean = 79.2

ii. The median = 79

iii. The mode = 94

iv. The range = 58

v. The sample variance = 209.6

vi. The sample standard deviation = 14.5

vii. The first quartile = 68.5

viii. The third quartile = 93

ix. IQR = 93 – 68.5 = 24.5

KEY TERMS

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

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.

Mean (arithmetic) 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

$$\bar{x} = \frac{Sum\ of\ all\ values\ in\ the\ sample}{Number\ of\ values\ in\ the\ sample}$$

and the mean for a population (denoted by μ) is

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

Mean (geometric) a measure of central tendency that provides a measure of average geometric growth over multiple time periods.

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

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.

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 is the 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

2.2| 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 50^{th} 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 25^{th} percentile, the second quartile (Q_2 or median) is 50^{th} percentile, and the third quartile (Q_3) is the 75^{th} 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.

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

2.3| 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.

2.5| 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 **left (or negative) skewed** distribution has a shape like Figure 2.13. A **right (or positive) skewed** distribution has a shape like Figure 2.14. A **symmetrical** distribution looks like Figure 2.12.

2.6| Measures of the Spread of the Data

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \frac{\sqrt{\sum(x-\bar{x})^2}}{n-1}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \frac{\sqrt{\sum(x-\mu)^2}}{N}$.

REFERENCES

2.1| Frequency Distributions and Graphs

NHL 2015-2016 Standings (League). Available online at https://www.nhl.com/standings (accessed September 2, 2016).

2.2| Measures of the Location of the Data

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from San Jose Mercury News.

Data from Time Magazine; survey by Yankelovich Partners, Inc.

2.3| Measures of the Center of the Data

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/ r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

2.6| Measures of the Spread of the Data

King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).