Chapter 2

Descriptive Statistics

2.1 Descriptive Statistics[1]

2.1.1 Student Learning Objectives

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

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- 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.

2.1.2 Introduction

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 to calculate, and even more importantly, to interpret these measurements and graphs.

2.2 Displaying Data[2]

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The 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 in order to get a picture of the data. Then, more formal tools may be applied.

[1] This content is available online at <http://cnx.org/content/m16300/1.7/>.

[2] This content is available online at <http://cnx.org/content/m16297/1.7/>.

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

2.3 Stem and Leaf Graphs (Stemplots)[3]

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis.It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of **one digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.1
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem-and-Leaf Diagram

3	3
4	299
5	355
6	1378899
7	2348
8	03888
9	0244446
10	0

Table 2.1

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stemplot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example 2.2
Create a stem plot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

The data are the distance (in kilometers) from a home to the nearest supermarket.

[3] This content is available online at <http://cnx.org/content/m16849/1.7/>.

Problem ***(Solution on p. 92.)***

1. Are there any outliers?
2. Do the data seem to have any concentration of values?

HINT: The leaves are to the right of the decimal.

NOTE: This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the Texas Instruments (TI) website[4] .

2.4 Histograms[5]

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either "frequency" or "relative frequency". The graph will have the same shape with either label. **Frequency** is commonly used when the data set is small and **relative frequency** is used when the data set is large or when we want to compare several distributions. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on Sampling and Data (Section 1.1), we defined frequency as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:

$$RF = \frac{f}{n} \tag{2.1}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received an A, then,

$f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$

Seven and a half percent of the students received an A.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 - 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value

[4] http://education.ti.com/educationportal/sites/US/sectionHome/support.html

[5] This content is available online at <http://cnx.org/content/m16298/1.11/>.

is 1.5, a convenient starting point is 1.495 (1.5 - 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 - .0005 = 0.9995). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 (2 - 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example 2.3
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5

70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71

72; 72; 72; 72.5; 72.5; 73; 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 - 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74. 74+ 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.

$$\frac{74.05 - 59.95}{8} = 1.76 \tag{2.2}$$

NOTE: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. For this example, using 1.76 as the width would also work.

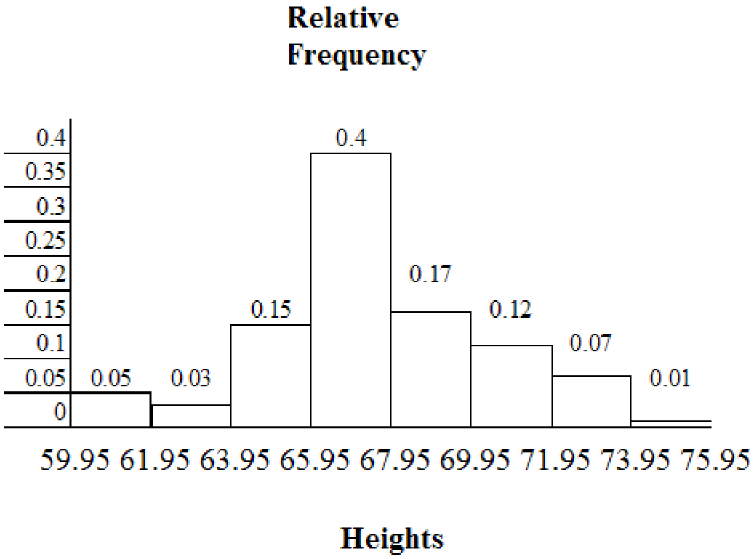
The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95

- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95 - 61.95. The heights that are 63.5 are in the interval 61.95 - 63.95. The heights that are 64 through 64.5 are in the interval 63.95 - 65.95. The heights 66 through 67.5 are in the interval 65.95 - 67.95. The heights 68 through 69.5 are in the interval 67.95 - 69.95. The heights 70 through 71 are in the interval 69.95 - 71.95. The heights 72 through 73.5 are in the interval 71.95 - 73.95. The height 74 is in the interval 73.95 - 75.95.

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Example 2.4
The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1

2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3

4; 4; 4; 4; 4; 4

5; 5; 5; 5; 5

6; 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Problem ***(Solution on p. 92.)***

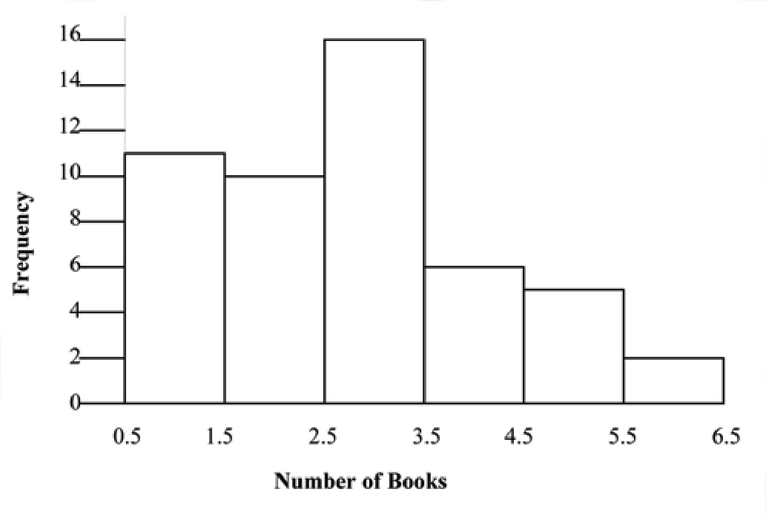
Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _______ to _______, the 5 in the middle of the interval from _______ to _______, and the _______ in the middle of the interval from _______ to _______ .

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{bars} = 1 \tag{2.3}$$

where 1 is the width of a bar. Therefore, $bars = 6$.

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



2.4.1 Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

2.5 Box Plots[6]

Box plots or **box-whisker plots** give a good graphical image of the concentration of the data. They also show how far from most of the data the extreme values are. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The median, the first quartile, and the third quartile will be discussed here, and then again in the section on measuring data in this chapter. We use these values to compare how close other data values are to them.

The **median**, a number, is a way of measuring the "center" of the data. You can think of the median as the "middle value," although 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

The median is between the 7th value, 6.8, and the 8th value 7.2. To find the median, add the two values together and divide by 2.

$$\frac{6.8+7.2}{2}=7 \tag{2.4}$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

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 is the middle value of the lower half of the data and the third quartile is the middle value of the upper half of the data. To get the idea, consider the same data set shown above:

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

The median or **second quartile** is 7. The lower half of the data is 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is 2.

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

The number 2, which is part of the data, is the **first quartile**. One-fourth of the values are the same or less than 2 and three-fourths of the values are more than 2.

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 9.

7.2; 8; 8.3; **9**; 10; 10; 11.5

The number 9, which is part of the data, is the **third quartile**. Three-fourths of the values are less than 9 and one-fourth of the values are more than 9.

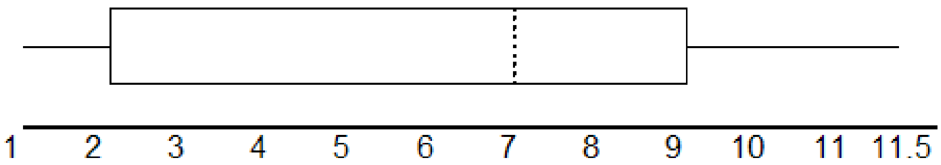
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The box plot gives a good quick picture of the data.

[6] This content is available online at <http://cnx.org/content/m16296/1.8/>.

Consider the following data:

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

The first quartile is 2, the median is 7, and the third quartile is 9. The smallest value is 1 and the largest value is 11.5. The box plot is constructed as follows (see calculator instructions in the back of this book or on the TI web site[7]):



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.

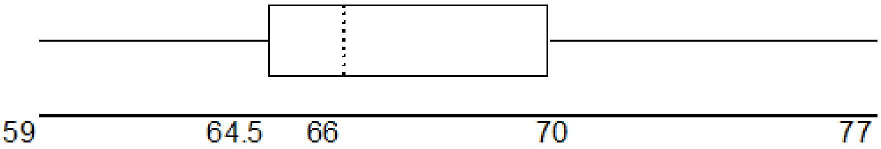
Example 2.5

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:

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



a. Each quarter has 25% of the data.

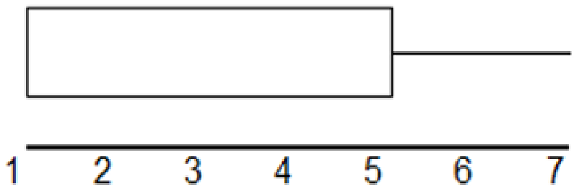
b. The spreads of the four quarters are 64.5 - 59 = 5.5 (first quarter), 66 - 64.5 = 1.5 (second quarter), 70 - 66 = 4 (3rd quarter), and 77 - 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.

c. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 =$ 5.5.

d. The interval 59 through 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.

[7] http://education.ti.com/educationportal/sites/US/sectionHome/support.html

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 1, the median and the third quartile were both 5, and the largest value was 7, the box plot would look as follows:



Example 2.6
Test scores for a college statistics class held during the day are:

99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

Test scores for a college statistics class held during the evening are:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

Problem ***(Solution on p. 92.)***

- What are the smallest and largest data values for each data set?
- What is the median, the first quartile, and the third quartile for each data set?
- Create a boxplot for each set of data.
- Which boxplot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?
- For each data set, what percent of the data is between the smallest value and the first quartile? (Answer: 25%) the first quartile and the median? (Answer: 25%) the median and the third quartile? the third quartile and the largest value? What percent of the data is between the first quartile and the largest value? (Answer: 75%)

The first data set (the top box plot) has the widest spread for the middle 50% of the data. $IQR = Q3 - Q1$ is $82.5 - 56 = 26.5$ for the first data set and $89 - 78 = 11$ for the second data set. So, the first set of data has its middle 50% of scores more spread out.

25% of the data is between M and $Q3$ and 25% is between $Q3$ and $Xmax$.

2.6 Measures of the Location of the Data[8]

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3, is the same as the 75th percentile (75th %ile). The median, M, is called both the second quartile and the 50th percentile (50th %ile).

[8] This content is available online at <http://cnx.org/content/m16314/1.10/>.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that 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 your score was higher than 90% of the people who took the test and lower than the scores of the remaining 10% of the people who took the test. Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

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).

$$IQR = Q_3 - Q_1 \tag{2.5}$$

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

Example 2.7
For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

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
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

$M = 488,800$

$Q_1 = \frac{230500+387000}{2} = 308750$

$Q_3 = \frac{639000+659000}{2} = 649000$

$IQR = 649000 - 308750 = 340250$

$(1.5)\,(IQR) = (1.5)\,(340250) = 510375$

$Q_1 - (1.5)\,(IQR) = 308750 - 510375 = -201625$

$Q_3 + (1.5)\,(IQR) = 649000 + 510375 = 1159375$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2.8
For the two data sets in the test scores example (p. 57), find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.
c. The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Example 2.9: Finding Quartiles and Percentiles Using a Table
Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

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 2.2

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

Find the median: Look again at the "cumulative relative frequency " column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 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 %ile is between the 25th (7) and 26th (7) values. **The median is 7.**

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 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8** . Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3, is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example 2.10
Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile. What is another name for the first quartile?
4. Construct a box plot of the data.

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?

3. Find the mean and standard deviation.
4. Find the mode.
5. Construct 2 different histograms. For each, starting value = _____ ending value = ____.
6. Find the median, first quartile, and third quartile.
7. Construct a box plot.
8. Construct a table of the data to find the following:
   - The 10th percentile
   - The 70th percentile
   - The percent of students who own less than 4 sweaters

2.7 Measures of the Center of the Data[9]

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 (previously discussed under box plots in this chapter). 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.

The mean can also 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"): $\overline{x}$.

The Greek letter μ (pronounced "mew") represents the population mean. If you take a truly random sample, the sample mean is a good estimate of the population mean.

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

$$\overline{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{2.6}$$

$$\overline{x} = \frac{3\times 1+2\times 2+1\times 3+5\times 4}{11} = 2.7 \tag{2.7}$$

In the second example, 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 2 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 median itself 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 median itself.

[9] This content is available online at <http://cnx.org/content/m17102/1.8/>.

Example 2.11
AIDS data indicating the number of months an AIDS patient 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
The calculation for the mean is:

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

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

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

The median is 24.

Example 2.12
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
$\overline{x} = \frac{5000000+49\times30000}{50} = 129400$

$M = 30000$

(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.

Another measure of the center is the mode. The **mode** is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example 2.13: Statistics exam scores for 20 students are as follows
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

Problem
Find the mode.

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

Example 2.14
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises an average 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.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

2.7.1 The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample gets closer and closer to μ. This is discussed in more detail in the section **The Central Limit Theorem** of this course.

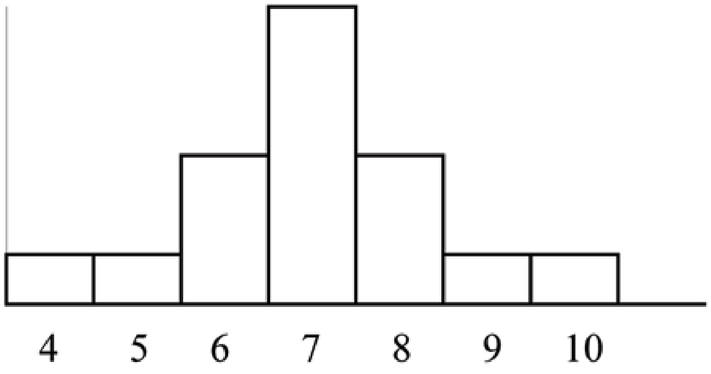
NOTE: The formula for the mean is located in the Summary of Formulas (Section 2.10) section course.

2.8 Skewness and the Mean, Median, and Mode[10]

Consider the following data set:

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

This data produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.



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

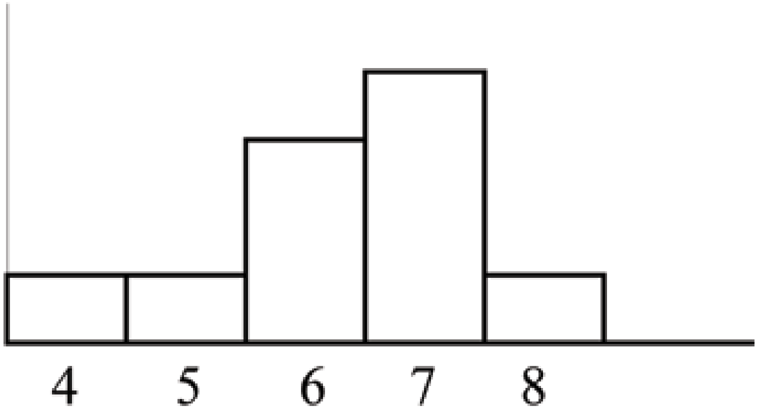
[10] This content is available online at <http://cnx.org/content/m17104/1.5/>.

line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean, the median, and the mode are the same.**

The histogram for the data:

4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

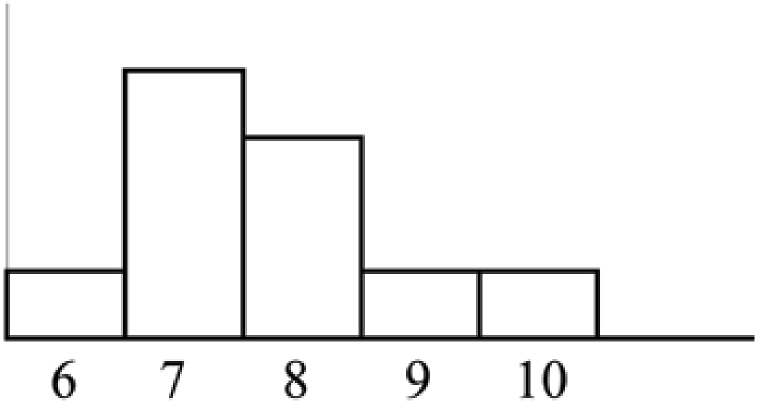


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. **Notice that the mean is the largest statistic, 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 less than the mode. If the distribution of data is skewed to the right, the mode is less than the median, which is less than the mean.

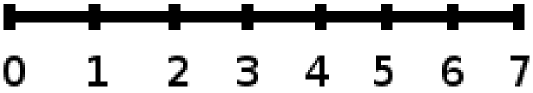
Skewness and symmetry become important when we discuss probability distributions in later chapters.

2.9 Measures of the Spread of the Data[11]

The most common measure of spread is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean. For example, if the mean of a set of data containing 7 is 5 and the **standard deviation** is 2, then the value 7 is one (1) standard deviation from its mean because 5 + (1)(2) = 7.

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is **one** standard deviation to the **right** of 5. If 1 were also part of the data set, then 1 is **two** standard deviations to the **left** of 5 because 5 +(-2)(2) = 1.

1=5+(-2)(2) ; 7=5+(1)(2)



Formula: value = $\overline{x}$ + (#ofSTDEVs)(s)

Generally, a value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations.

If x is a value and $\overline{x}$ is the sample mean, then $x - \overline{x}$ is called a deviation. In a data set, there are as many deviations as there are data values. Deviations are used to calculate the sample standard deviation.

Calculation of the Sample Standard Deviation
To calculate the standard deviation, calculate the variance first. The **variance** is the average of the squares of the deviations. The standard deviation is the square root of the variance. You can think of the standard deviation as a special average of the deviations (the $x - \overline{x}$ values). The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma) represents the population standard deviation. We use s^2 to represent the sample variance and σ^2 to represent the population variance. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

> NOTE: In practice, use either a calculator or computer software to calculate the standard deviation. However, please study the following step-by-step example.

Example 2.15
In a fifth grade class, the teacher was interested in the average age and the standard deviation of the ages of her students. What follows are the ages of her students 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

$$\overline{x} = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525 \tag{2.8}$$

[11] This content is available online at <http://cnx.org/content/m17103/1.8/>.

The average age is 10.53 years, rounded to 2 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**	*Deviations*^{2}	**(Freq.)**(*Deviations*^{2})
x	f	$(x-\overline{x})$	$(x-\overline{x})^2$	$(f)\,(x-\overline{x})^2$
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	$10-10.525=-0.525$	$(-0.525)^2=0.275625$	$4\times .275625=1.1025$
10.5	4	$10.5-10.525=-0.025$	$(-0.025)^2=0.000625$	$4\times .000625=.0025$
11	6	$11-10.525=0.475$	$(0.475)^2=0.225625$	$6\times .225625=1.35375$
11.5	3	$11.5-10.525=0.975$	$(0.975)^2=0.950625$	$3\times .950625=2.851875$

Table 2.3

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:

$s=\sqrt{0.5125}=.0715891$ 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.

Problem 1

Verify the mean and standard deviation calculated above on your calculator or computer. Find the median and mode.

Solution

- Median = 10.5
- Mode = 11

Problem 2

Find the value that is 1 standard deviation above the mean. Find $(\overline{x}+1s)$.

Solution

$(\overline{x}+1s)=10.53+(1)\,(0.72)=11.25$

Problem 3

Find the value that is two standard deviations below the mean. Find $(\overline{x}-2s)$.

Solution

$(\overline{x}-2s)=10.53-(2)\,(0.72)=9.09$

Problem 4
Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

- $(\overline{x} - 1.5s) = 10.53 - (1.5)\,(0.72) = 9.45$
- $(\overline{x} + 1.5s) = 10.53 + (1.5)\,(0.72) = 11.61$

Explanation of the table: The deviations show how spread out the data are about the mean. The value 11.5 is farther from the mean than 11. The deviations 0.975 and 0.475 indicate that. **If you add the deviations, the sum is always zero**. (For this example, there are 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. The variance, then, is the average squared deviation. It is small if the values are close to the mean and large if the values are far from the mean.

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.

For the sample variance, we divide by the total number of data values 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.** By dividing by $(n-1)$, we get a better estimate of the population variance.

Your concentration should be on what the standard deviation does, not on the arithmetic. 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 sample standard deviation, s , is either zero or larger than zero. When $s = 0$, there is no spread. When s is a lot larger than zero, the data values are very spread out about the mean. Outliers can make s 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**.

NOTE: The formula for the standard deviation is at the end of the chapter.

Example 2.16
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; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:

i. The sample mean
ii. The sample standard deviation
iii. The median
iv. The first quartile

v. The third quartile
vi. IQR

c. 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

a.

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
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 2.4

b. i. The sample mean = 73.5
ii. The sample standard deviation = 17.9
iii. The median = 73
iv. The first quartile = 61
v. The third quartile = 90
vi. 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; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: 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.

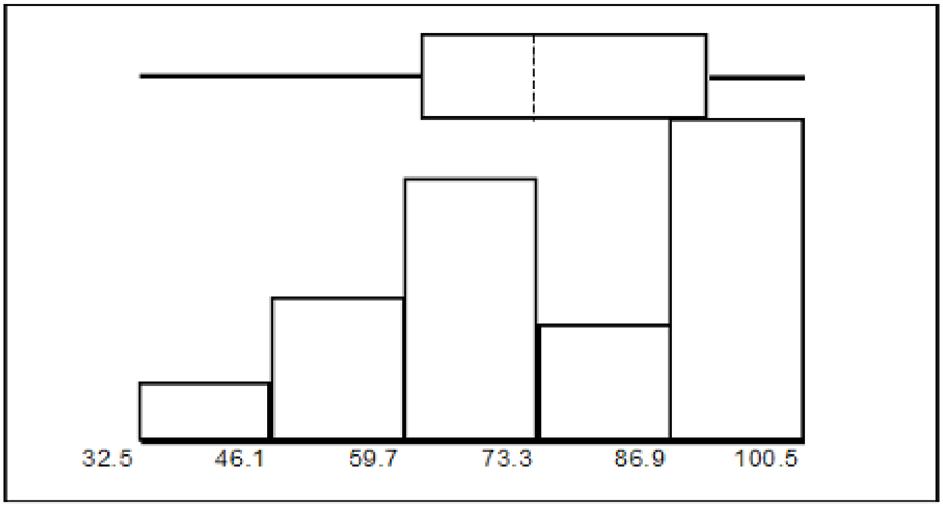


Figure 2.1

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.

Example 2.17
Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. 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 2.5

Solution
Use the formula **value = mean + (#ofSTDEVs)(stdev)** and solve for #ofSTDEVs for each student (stdev = standard deviation):

$\#ofSTDEVs = \frac{value - mean}{stdev}$:

For John, $\#ofSTDEVs = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $\#ofSTDEVs = \frac{77-80}{10} = -0.3$

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

2.10 Summary of Formulas[12]

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- $\overline{x}$ = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x * f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x * f$ = The sum of values multiplied by their respective frequencies
- $(x - \overline{x})$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean)
- $(x - \overline{x})^2$ or $(x - \mu)^2$ = Deviations squared
- $f\ (x - \overline{x})^2$ or $f\ (x - \mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $\overline{x} = \frac{\sum x}{n}$ or $\overline{x} = \frac{\sum f \cdot x}{n}$
- $\mu = \frac{\sum x}{N}$ or $\mu = \frac{\sum f \cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x-\overline{x})^2}{n-1}}$
- $\sigma = \sqrt{\frac{\Sigma(x-\overline{\mu})^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot (x-\overline{\mu})^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- $x = \overline{x}$+ (#ofSTDEVs)(s)
- $x = \mu$ + (#ofSTDEVs)(σ)

[12] This content is available online at <http://cnx.org/content/m16310/1.8/>.

2.11 Practice 1: Center of the Data[13]

2.11.1 Student Learning Outcomes

- The student will calculate and interpret the center, spread, and location of the data.
- The student will construct and interpret histograms an box plots.

2.11.2 Given

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

2.11.3 Complete the Table

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Table 2.6

2.11.4 Discussion Questions

Exercise 2.11.1 ***(Solution on p. 93.)***

What does the frequency column sum to? Why?

Exercise 2.11.2 ***(Solution on p. 93.)***

What does the relative frequency column sum to? Why?

Exercise 2.11.3

What is the difference between relative frequency and frequency for each data value?

Exercise 2.11.4

What is the difference between cumulative relative frequency and relative frequency for each data value?

2.11.5 Enter the Data

Enter your data into your calculator or computer.

[13] This content is available online at <http://cnx.org/content/m16312/1.12/>.

2.11.6 Construct a Histogram

Determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram below. Label the horizontal and vertical axes with words. Include numerical scaling.

2.11.7 Data Statistics

Calculate the following values:

Exercise 2.11.5 ***(Solution on p. 94.)***
Sample mean = $\overline{x}$ =

Exercise 2.11.6 ***(Solution on p. 94.)***
Sample standard deviation = s_x =

Exercise 2.11.7 ***(Solution on p. 94.)***
Sample size = n =

2.11.8 Calculations

Use the table in section 2.11.3 to calculate the following values:

Exercise 2.11.8 ***(Solution on p. 94.)***
Median =

Exercise 2.11.9 ***(Solution on p. 94.)***
Mode =

Exercise 2.11.10 ***(Solution on p. 94.)***
First quartile =

Exercise 2.11.11 ***(Solution on p. 94.)***
Second quartile = median = 50th percentile =

Exercise 2.11.12 ***(Solution on p. 94.)***
Third quartile =

Exercise 2.11.13 ***(Solution on p. 94.)***
Interquartile range (IQR) = _____ - _____ = _____

Exercise 2.11.14 ***(Solution on p. 94.)***
10th percentile =

Exercise 2.11.15 ***(Solution on p. 94.)***
70th percentile =

Exercise 2.11.16 ***(Solution on p. 94.)***

Find the value that is 3 standard deviations:

a. Above the mean
b. Below the mean

2.11.9 Box Plot

Construct a box plot below. Use a ruler to measure and scale accurately.

2.11.10 Interpretation

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

2.12 Practice 2: Spread of the Data[14]

2.12.1 Student Learning Objectives

- The student will calculate measures of the center of the data.
- The student will calculate the spread of the data.

2.12.2 Given

The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976-77 through 2004-2005. (*Source: Graphically Speaking by Bill King, LTCC Institutional Research, December 2005*).

Use these values to answer the following questions:

- μ = 1000 FTES
- Median - 1014 FTES
- σ = 474 FTES
- First quartile = 528.5 FTES
- Third quartile = 1447.5 FTES
- n = 29 years

2.12.3 Calculate the Values

Exercise 2.12.1 ***(Solution on p. 94.)***

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Exercise 2.12.2 ***(Solution on p. 94.)***

75% of all years have a FTES:

a. At or below:
b. At or above:

Exercise 2.12.3 ***(Solution on p. 94.)***

The population standard deviation =

Exercise 2.12.4 ***(Solution on p. 94.)***

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Exercise 2.12.5 ***(Solution on p. 94.)***

What is the IQR? What does the IQR represent?

Exercise 2.12.6 ***(Solution on p. 94.)***

How many standard deviations away from the mean is the median?

[14] This content is available online at <http://cnx.org/content/m17105/1.10/>.

2.13 Homework[15]

Exercise 2.13.1 ***(Solution on p. 94.)***

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

Table 2.7

a. Find the sample mean $\overline{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students saw fewer than three movies?
j. Find the 40th percentile.
k. Find the 90th percentile.

Exercise 2.13.2

The median age for U.S. blacks currently is 30.1 years; for U.S. whites it is 36.6 years. (Source: U.S. Census)

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.

b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise 2.13.3 ***(Solution on p. 95.)***

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

[15] This content is available online at <http://cnx.org/content/m16801/1.12/>.

X	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	5		
3	8		
4	12		
5	12		
7	1		

Table 2.8

a. Find the sample mean $\overline{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.

Exercise 2.13.4
600 adult Americans were asked by telephone poll, What do you think constitutes a middle-class income? The results are below. Also, include left endpoint, but not the right endpoint. (*Source: Time magazine; survey by Yankelovich Partners, Inc.*)

NOTE: "Not sure" answers were omitted from the results.

Salary ($)	Relative Frequency
$<$ 20,000	0.02
20,000 - 25,000	0.09
25,000 - 30,000	0.19
30,000 - 40,000	0.26
40,000 - 50,000	0.18
50,000 - 75,000	0.17
75,000 - 99,999	0.02
100,000+	0.01

Table 2.9

a. What percent of the survey answered "not sure" ?
b. What percent think that middle-class is from $25,000 - $50,000 ?
c. Construct a histogram of the data

1. Should all bars have the same width, based on the data? Why or why not?
2. How should the <20,000 and the 100,000+ intervals be handled? Why?

d. Find the 40th and 80th percentiles

Exercise 2.13.5 ***(Solution on p. 95.)***

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year (Source: San Jose Mercury News).

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _______ to _______.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population were the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:

i. the population mean, μ.
ii. the population standard deviation, σ.
iii. the weight that is 2 standard deviations below the mean.
iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the average weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

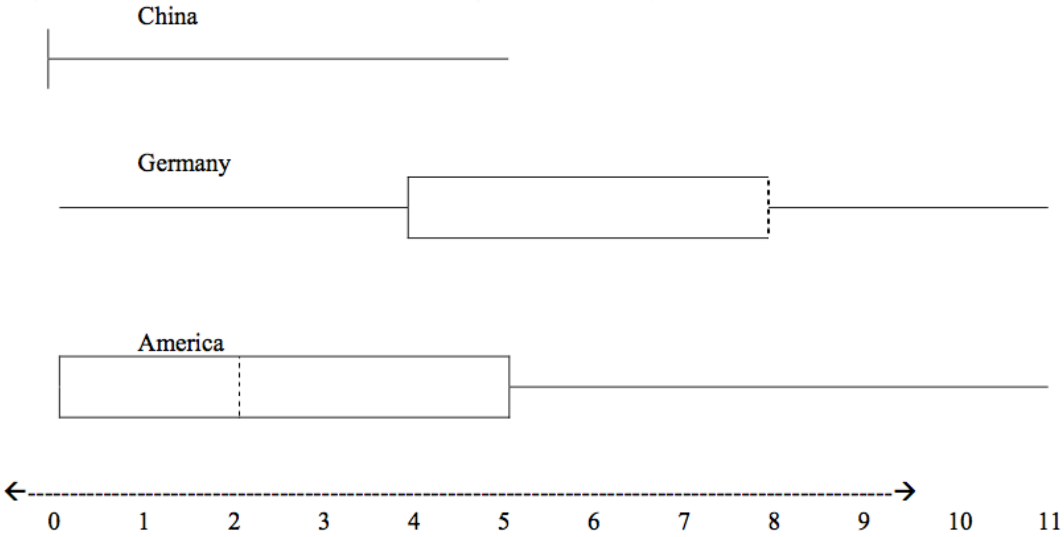
Exercise 2.13.6

An elementary school class ran 1 mile in an average of 11 minutes with a standard deviation of 3 minutes. Rachel, a student in the class, ran 1 mile in 8 minutes. A junior high school class ran 1 mile in an average of 9 minutes, with a standard deviation of 2 minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran 1 mile in an average of 7 minutes with a standard deviation of 4 minutes. Nedda, a student in the class, ran 1 mile in 8 minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Exercise 2.13.7

In a survey of 20 year olds in China, Germany and America, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Explain how it is possible that more Americans than Germans surveyed have been to over eight foreign countries.
c. Compare the three box plots. What do they imply about the foreign travel of twenty year old residents of the three countries when compared to each other?

Exercise 2.13.8

Twelve teachers attended a seminar on mathematical problem solving. Their attitudes were measured before and after the seminar. A positive number change attitude indicates that a teacher's attitude toward math became more positive. The twelve change scores are as follows:

3; 8; -1; 2; 0; 5; -3; 1; -1; 6; 5; -2

a. What is the average change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

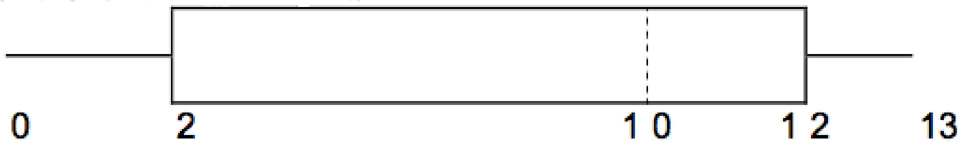
Exercise 2.13.9 ***(Solution on p. 95.)***

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best G.P.A. when compared to his school? Explain how you determined your answer.

Student	**G.P.A.**	**School Ave. G.P.A.**	**School Standard Deviation**
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Table 2.10

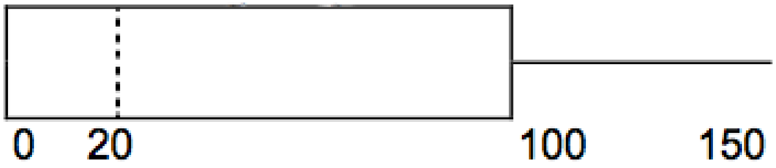
Exercise 2.13.10
Given the following box plot:



a. Which quarter has the smallest spread of data? What is that spread?
b. Which quarter has the largest spread of data? What is that spread?
c. Find the Inter Quartile Range (IQR).
d. Are there more data in the interval 5 - 10 or in the interval 10 - 13? How do you know this?
e. Which interval has the fewest data in it? How do you know this?

I. 0-2
II. 2-4
III. 10-12
IV. 12-13

Exercise 2.13.11
Given the following box plot:



a. Think of an example (in words) where the data might fit into the above box plot. In 2-5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to fourth quartiles are far apart?

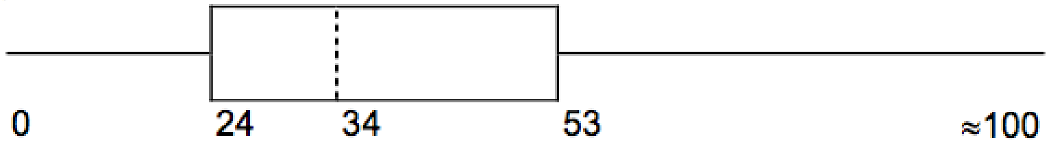
Exercise 2.13.12
Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows. (*Source: West magazine*)

Age Group	Percent of Community
0-17	18.9
18-24	8.0
25-34	22.8
35-44	15.0
45-54	13.1
55-64	11.9
65+	10.3

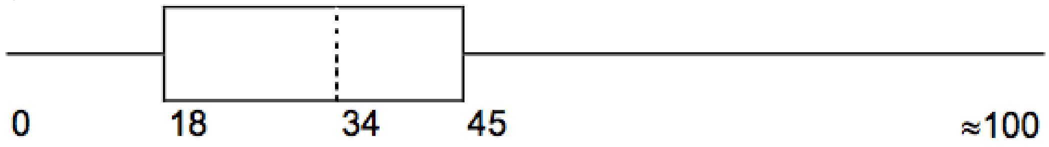
Table 2.11

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not?
b. What percent of the community is under age 35?
c. Which box plot most resembles the information above?

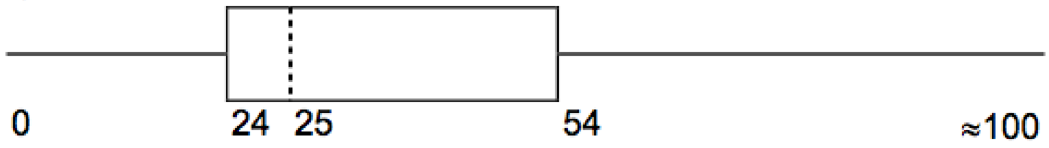
i.



ii.



iii.



Exercise 2.13.13

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, each asked adult consumers the number of fiction paperbacks they had purchased the previous month. The results are below.

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.12

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.13

Publisher C

# of books	Freq.	Rel. Freq.
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

Table 2.14

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise 2.13.14

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all on-board transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Below is a summary of the bills for each group.

Singles

Amount($)	Frequency	Rel. Frequency
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

Table 2.15

Couples

Amount($)	Frequency	Rel. Frequency
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

Table 2.16

a. Fill in the relative frequency for each group.

b. Construct a histogram for the Singles group. Scale the x-axis by $50. widths. Use relative frequency on the y-axis.

c. Construct a histogram for the Couples group. Scale the x-axis by $50. Use relative frequency on the y-axis.

d. Compare the two graphs:

- **i.** List two similarities between the graphs.
- **ii.** List two differences between the graphs.
- **iii.** Overall, are the graphs more similar or different?

e. Construct a new graph for the Couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.

f. Compare the graph for the Singles with the new graph for the Couples:

- **i.** List two similarities between the graphs.
- **ii.** Overall, are the graphs more similar or different?

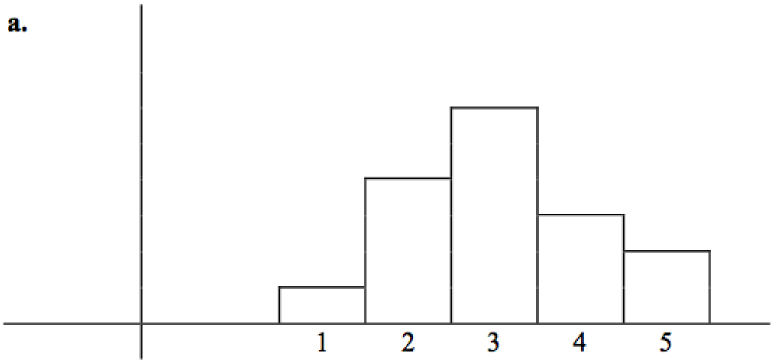
i. By scaling the Couples graph differently, how did it change the way you compared it to the Singles?

j. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person in a couple? Explain why in one or two complete sentences.

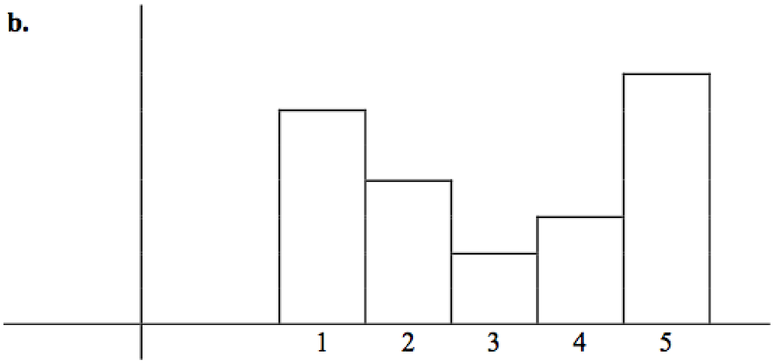
Exercise 2.13.15 ***(Solution on p. 95.)***

Refer to the following histograms and box plot. Determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

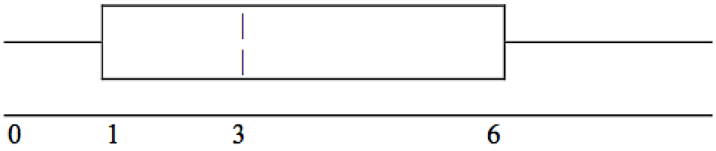
a.



b.



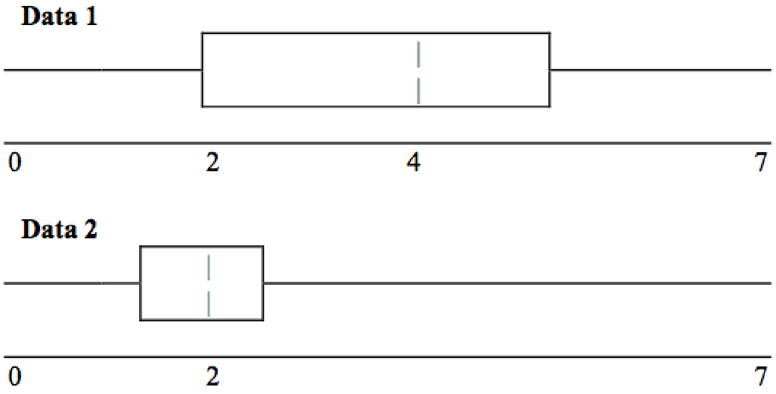
c.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for (b) is larger than the standard deviation for (a).
d. We cannot determine if any of the third quartiles for the three graphs is different.

Exercise 2.13.16

Refer to the following box plots.



a. In complete sentences, explain why each statement is false.

- **i.** **Data 1** has more data values above 2 than **Data 2** has above 2.
- **ii.** The data sets cannot have the same mode.
- **iii.** For **Data 1**, there are more data values below 4 than there are above 4.

b. For which group, Data 1 or Data 2, is the value of "7" more likely to be an outlier? Explain why in complete sentences

Exercise 2.13.17 ***(Solution on p. 96.)***

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _______ days to _______ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that 3 - 5 days seem to be popular lengths of engineering conferences.

Exercise 2.13.18

A survey of enrollment at 35 community colleges across the United States yielded the following figures (*source: Microsoft Bookshelf*):

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.

- **c.** If you were to build a new community college, which piece of information would be more valuable: the mode or the average size?
- **d.** Calculate the sample average.
- **e.** Calculate the sample standard deviation.
- **f.** A school with an enrollment of 8000 would be how many standard deviations away from the mean?

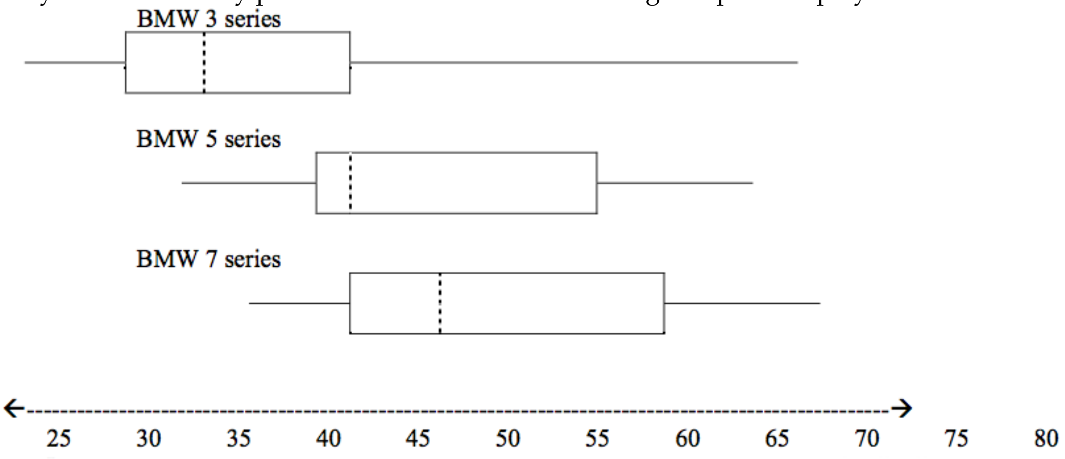
Exercise 2.13.19 ***(Solution on p. 96.)***

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years. (*Source: Bureau of the Census*)

- **a.** What does it mean for the median age to rise?
- **b.** Give two reasons why the median age could rise.
- **c.** For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

Exercise 2.13.20

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.

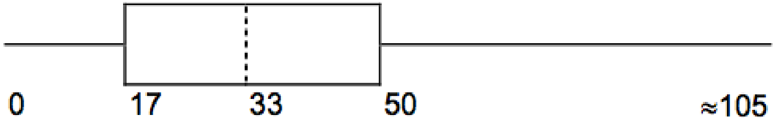


- **a.** In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
- **b.** Which group is most likely to have an outlier? Explain how you determined that.
- **c.** Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
- **d.** Look at the BMW 5 series. Which quarter has the smallest spread of data? What is that spread?
- **e.** Look at the BMW 5 series. Which quarter has the largest spread of data? What is that spread?
- **f.** Look at the BMW 5 series. Find the Inter Quartile Range (IQR).
- **g.** Look at the BMW 5 series. Are there more data in the interval 31-38 or in the interval 45-55? How do you know this?
- **h.** Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
  - **i.** 31-35
  - **ii.** 38-41

iii. 41-64

Exercise 2.13.21 *(Solution on p. 96.)*

The following box plot shows the U.S. population for 1990, the latest available year. (Source: Bureau of the Census, 1990 Census)



a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?

b. 12.6% are age 65 and over. Approximately what percent of the population are of working age adults (above age 17 to age 65)?

Exercise 2.13.22

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

	Javier	**Ercilla**
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

Table 2.17

a. How can you determine which survey was correct ?

b. Explain what the difference in the results of the surveys implies about the data.

c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

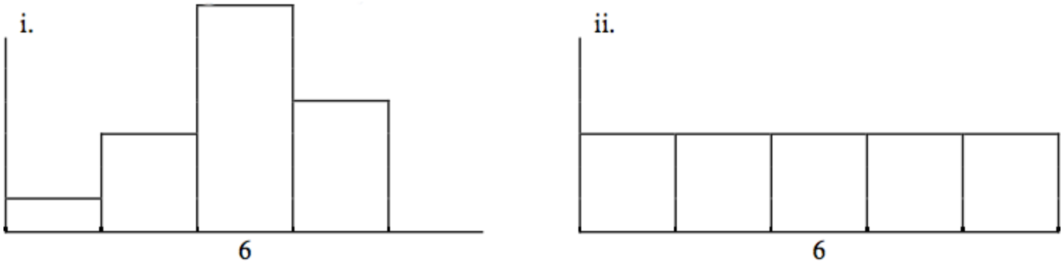


Figure 2.2

d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

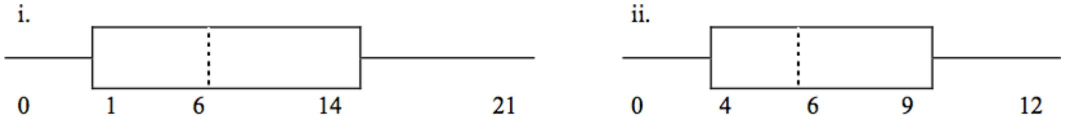


Figure 2.3

Exercise 2.13.23 ***(Solution on p. 96.)***

Student grades on a chemistry exam were:

77, 78, 76, 81, 86, 51, 79, 82, 84, 99

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

2.13.1 Try these multiple choice questions.

The next three questions refer to the following information. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency
7	1
14	3
15	1
18	1
19	4
20	3
22	1
23	1
26	1
40	2
42	2
	Total = 20

Table 2.18

Exercise 2.13.24 ***(Solution on p. 96.)***

What is the IQR?

A. 8

B. 11
C. 15
D. 35

Exercise 2.13.25 *(Solution on p. 96.)*
What is the mode?

A. 19
B. 19.5
C. 14 and 20
D. 22.65

Exercise 2.13.26 *(Solution on p. 96.)*
Is this a sample or the entire population?

A. sample
B. entire population
C. neither

The next two questions refer to the following table. X = the number of days per week that 100 clients use a particular exercise facility.

X	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Table 2.19

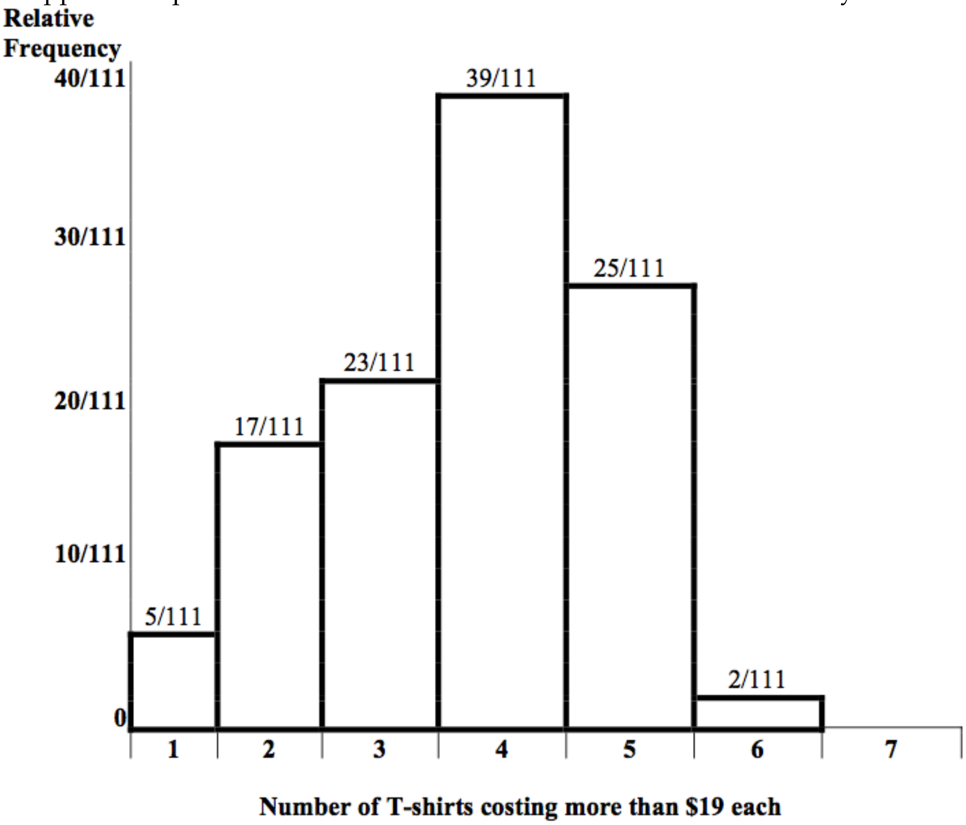
Exercise 2.13.27 *(Solution on p. 96.)*
The 80th percentile is:

A. 5
B. 80
C. 3
D. 4

Exercise 2.13.28 *(Solution on p. 96.)*
The number that is 1.5 standard deviations BELOW the mean is approximately:

A. 0.7
B. 4.8
C. -2.8
D. Cannot be determined

The next two questions refer to the following histogram. Suppose one hundred eleven people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



Exercise 2.13.29 ***(Solution on p. 96.)***

The percent of people that own at most three (3) T-shirts costing more than $19 each is approximately:

A. 21
B. 59
C. 41
D. Cannot be determined

Exercise 2.13.30 ***(Solution on p. 96.)***

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

A. cluster
B. simple random
C. stratified
D. convenience

2.14 Lab: Descriptive Statistics[16]

Class Time:

Names:

2.14.1 Student Learning Objectives

- 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.

2.14.2 Collect the Data

Record the number of pairs of shoes you own:

1. Randomly survey 30 classmates. Record their values.

Survey Results

______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______
______	______	______	______	______

Table 2.20

2. Construct a histogram. Make 5-6 intervals. Sketch the graph using a ruler and pencil. Scale the axes.

[16] This content is available online at <http://cnx.org/content/m16299/1.12/>.

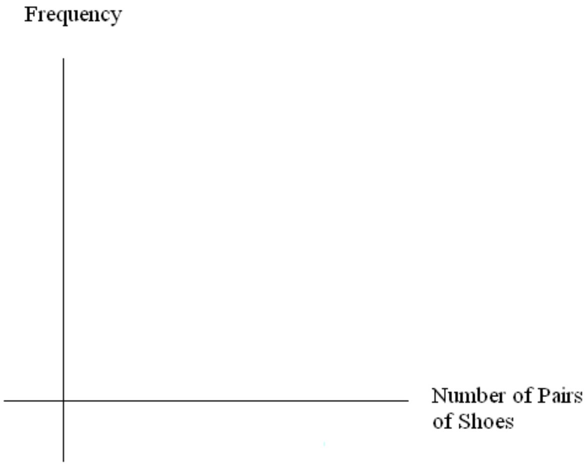


Figure 2.4

3. Calculate the following:
   - $\overline{x}$ =
   - s =
4. Are the data discrete or continuous? How do you know?
5. Describe the shape of the histogram. Use complete sentences.
6. Are there any potential outliers? Which value(s) is (are) it (they)? Use a formula to check the end values to determine if they are potential outliers.

2.14.3 Analyze the Data

1. Determine the following:
   - Minimum value =
   - Median =
   - Maximum value =
   - First quartile =
   - Third quartile =
   - IQR =
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:

Solutions to Exercises in Chapter 2

Solution to Example 2.2 (p. 51)
The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 miles.

Stem	**Leaf**
1	1 5
2	3 5 7
3	3 3 3 5 8
4	0 2 5 5 7 8
5	5 6 6
6	5 7
7	
8	
9	
10	
11	
12	3

Table 2.21

Solution to Example 2.4 (p. 54)

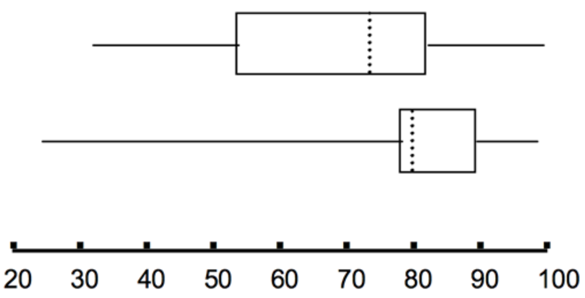
- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Solution to Example 2.6 (p. 57)
First Data Set

- $Xmin = 32$
- $Q1 = 56$
- $M = 74.5$
- $Q3 = 82.5$
- $Xmax = 99$

Second Data Set

- $Xmin = 25.5$
- $Q1 = 78$
- $M = 81$
- $Q3 = 89$
- $Xmax = 98$



Solution to Example 2.8 (p. 58)
For the IQRs, see the answer to the test scores example (First Data Set, p. 92 Second Data Set, p. 92 p. 475). The first data set has the larger IQR, so the scores between $Q3$ and $Q1$ (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot (IQR) = \left(\frac{3}{2}\right) \cdot (26.5) = 39.75$
- $Xmax - Q3 = 99 - 82.5 = 16.5$
- $Q1 - Xmin = 56 - 32 = 24$

$\left(\frac{3}{2}\right) \cdot (IQR) = $ 39.75 is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (IQR) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $Xmax - Q3 = 98 - 89 = 9$
- $Q1 - Xmin = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (IQR) = $ 16.5 is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and Data Chapter (Section 1.9)). Get the percentiles from that chart.

First Data Set

- $\textit{30th \%ile (between the 6th and 7th values)} = \frac{(56 + 59)}{2} = 57.5$
- $\textit{80th \%ile (between the 16th and 17th values)} = \frac{(84 + 84.5)}{2} = 84.25$

Second Data Set

- $\textit{30th \%ile (7th value)} = 78$
- $\textit{80th \%ile (18th value)} = 90$

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Solution to Example 2.10 (p. 59)

1. $\frac{(8+9)}{2} = 8.5$
2. 9
3. 6
4. First Quartile = 25th %ile

Solutions to Practice 1: Center of the Data

Solution to Exercise 2.11.1 (p. 71)
65

Solution to Exercise 2.11.2 (p. 71)
1
Solution to Exercise 2.11.5 (p. 72)
4.75
Solution to Exercise 2.11.6 (p. 72)
1.39
Solution to Exercise 2.11.7 (p. 72)
65
Solution to Exercise 2.11.8 (p. 72)
4
Solution to Exercise 2.11.9 (p. 72)
4
Solution to Exercise 2.11.10 (p. 72)
4
Solution to Exercise 2.11.11 (p. 72)
4
Solution to Exercise 2.11.12 (p. 72)
6
Solution to Exercise 2.11.13 (p. 72)
$6 - 4 = 2$
Solution to Exercise 2.11.14 (p. 72)
3
Solution to Exercise 2.11.15 (p. 72)
6
Solution to Exercise 2.11.16 (p. 73)

a. 8.93
b. 0.58

Solutions to Practice 2: Spread of the Data

Solution to Exercise 2.12.1 (p. 74)
6
Solution to Exercise 2.12.2 (p. 74)

a. 1447.5
b. 528.5

Solution to Exercise 2.12.3 (p. 74)
474 FTES
Solution to Exercise 2.12.4 (p. 74)
50%
Solution to Exercise 2.12.5 (p. 74)
919
Solution to Exercise 2.12.6 (p. 74)
0.03

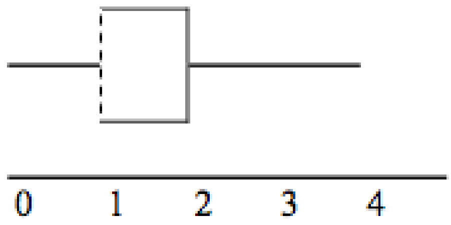
Solutions to Homework

Solution to Exercise 2.13.1 (p. 75)

a. 1.48
b. 1.12

e. 1
f. 1
g. 2

h.

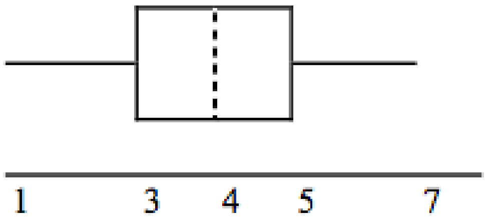


i. 80%
j. 1
k. 3

Solution to Exercise 2.13.3 (p. 75)

a. 3.78
b. 1.29
e. 3
f. 4
g. 5

h.

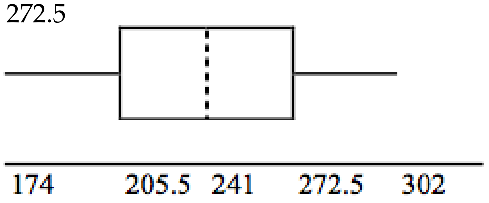


i. 32.5%
j. 4
k. 5

Solution to Exercise 2.13.5 (p. 77)

b. 241
c. 205.5
d. 272.5

e.



f. 205.5, 272.5
g. sample
h. population
i. **i.** 236.34
  ii. 37.50
  iii. 161.34
  iv. 0.84 std. dev. below the mean
j. Young

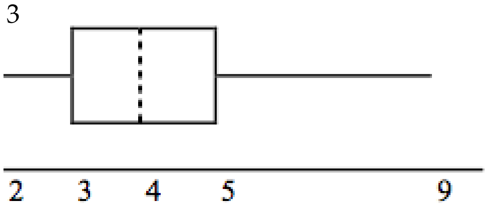
Solution to Exercise 2.13.9 (p. 78)
Kamala

Solution to Exercise 2.13.15 (p. 83)

a. True
b. True
c. True
d. False

Solution to Exercise 2.13.17 (p. 84)

b. 4,3,5
c. 4
d. 3

e. 

f. 3,5
g. 3.94
h. 1.28
i. 3
j. mode

Solution to Exercise 2.13.19 (p. 85)

c. Maybe

Solution to Exercise 2.13.21 (p. 86)

a. more children
b. 62.4%

Solution to Exercise 2.13.23 (p. 87)

b. 51,99

Solution to Exercise 2.13.24 (p. 87)
A
Solution to Exercise 2.13.25 (p. 88)
A
Solution to Exercise 2.13.26 (p. 88)
B
Solution to Exercise 2.13.27 (p. 88)
D
Solution to Exercise 2.13.28 (p. 88)
A
Solution to Exercise 2.13.29 (p. 89)
C
Solution to Exercise 2.13.30 (p. 89)
D