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://http://cnx.org/content/m16300/1.7/>.

[2] This content is available online at <http://http://cnx.org/content/m16297/1.8/>.

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, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

2.3 Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs[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

Stem	**Leaf**
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:

[3] This content is available online at <http://http://cnx.org/content/m16849/1.11/>.

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.

Problem ***(Solution on p. 106.)***

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.

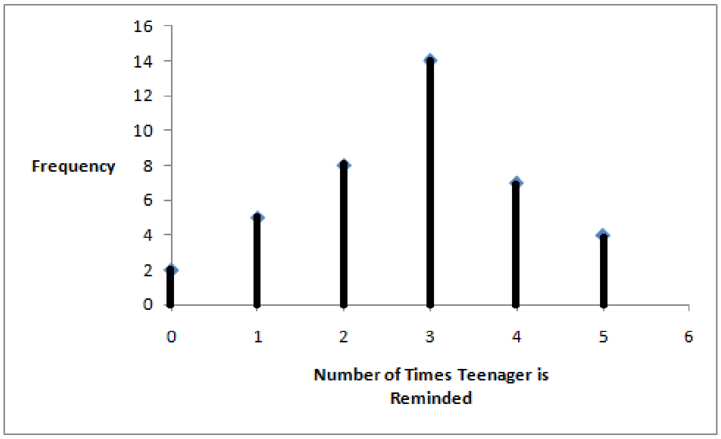
Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequencies** indicated by the **heights** of the vertical lines.

Example 2.3

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

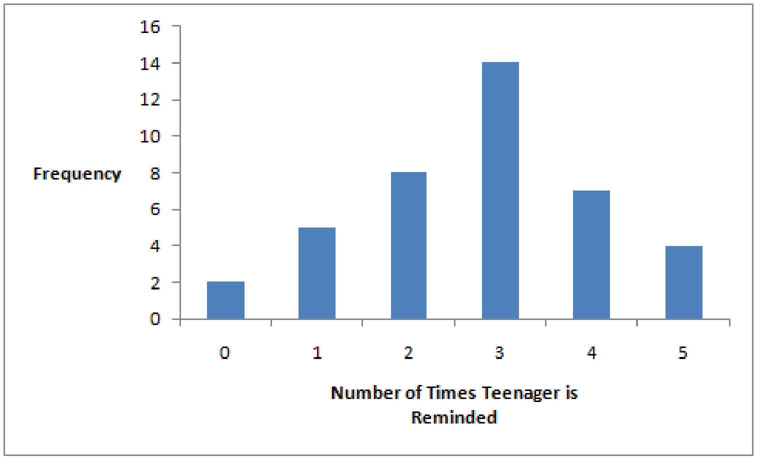
Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Table 2.2



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal. The bar graph shown in **Example 4** uses the data of **Example 3** and is similar to the line graph. Frequencies are represented by the **heights of the bars.**

Example 2.4



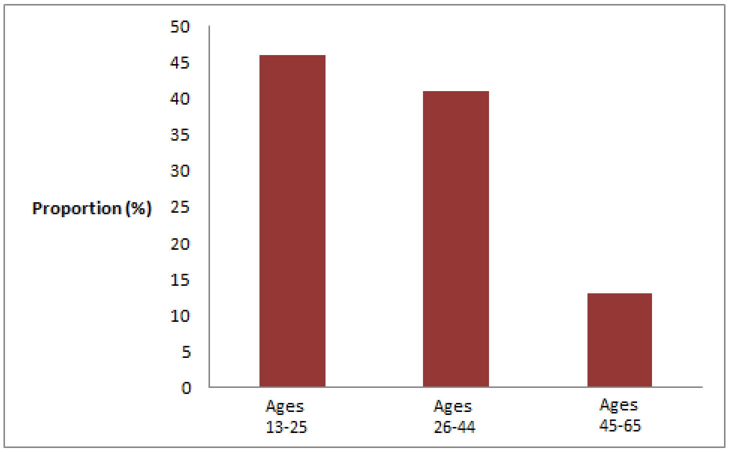
The **bar graph** shown in **Example 5** has age groups represented on the **x-axis** and proportions on the **y-axis**.

Example 2.5

By the end of March 2009, in the United States Facebook had over 56 million users. The table shows the age groups, the number of users in each age group and the proportion (%) of users in each age group. **Source:** ***http://www.insidefacebook.com/2009/03/25/number-of-us-facebook-users-over-35-nearly-doubles-in-last-60-days/***

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13 - 25	25,510,040	46%
26 - 44	23,123,900	41%
45 - 65	7,431,020	13%

Table 2.3

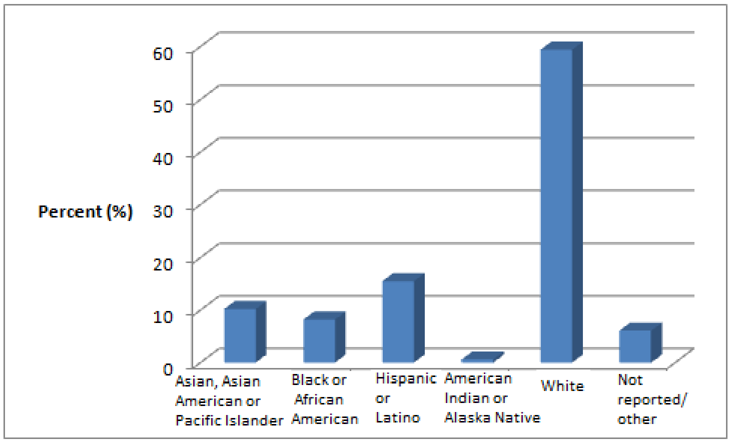


Example 2.6

The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2009, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**. (**Source: http://www.collegeboard.com**)

Race/Ethnicity	AP Examinee Population	Overall Student Population
Asian, Asian American or Pacific Islander	10.2%	5.4%
Black or African American	8.2%	14.5%
Hispanic or Latino	15.5%	15.9%
American Indian or Alaska Native	0.6%	1.2%
White	59.4%	61.6%
Not reported/other	6.1%	1.4%

Table 2.4



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

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

[5] This content is available online at <http://http://cnx.org/content/m16298/1.12/>.

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:

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

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

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

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 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 2.7
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.

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

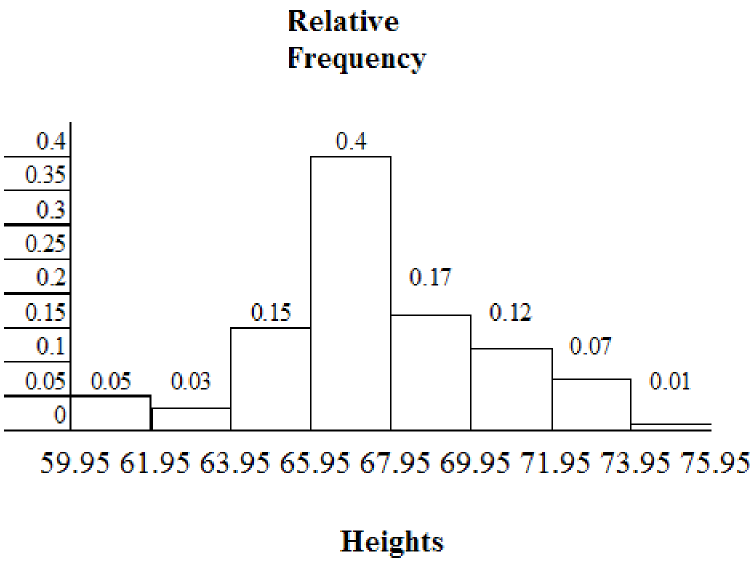
NOTE: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. For this example, using 1.76 as the width would also work.

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 2.8
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.

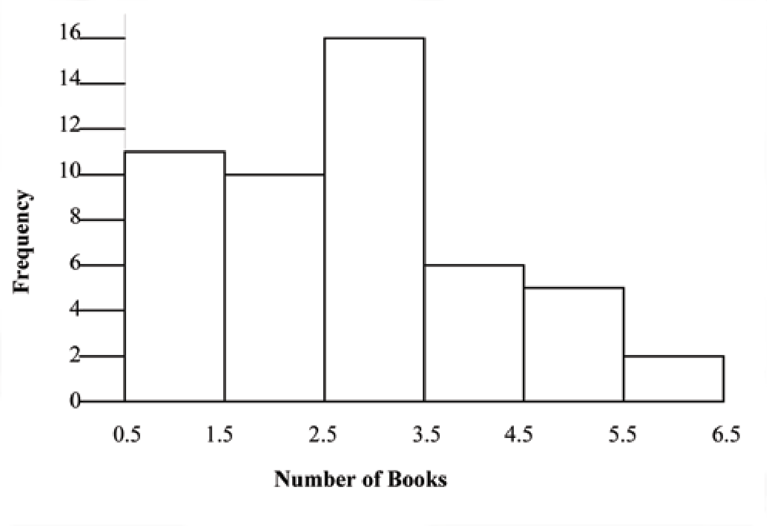
Problem ***(Solution on p. 106.)***
Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 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:

[6] This content is available online at <http://http://cnx.org/content/m16296/1.11/>.

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.

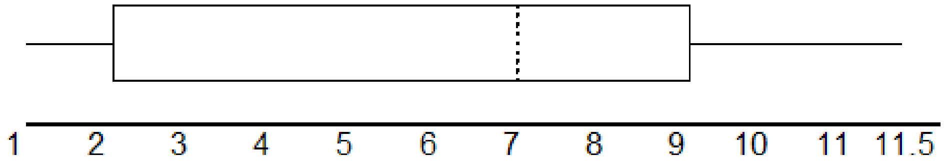
To construct a box plot, use a horizontal number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. **The middle fifty percent of the data fall inside the 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.

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]):

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



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.

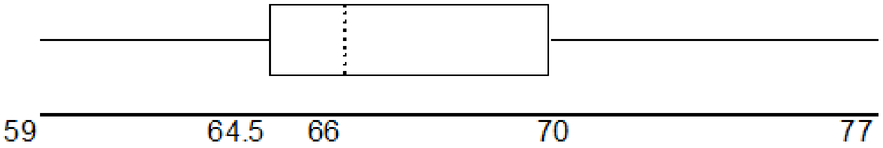
Example 2.9

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

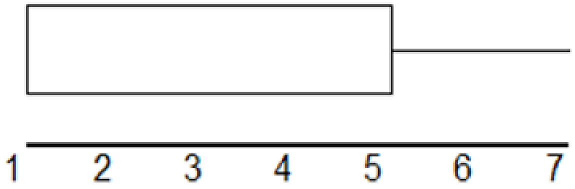
Construct a box plot with the following properties:

- 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.

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 2.10

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. 106.)***

- 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).

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.

[8] This content is available online at <http://http://cnx.org/content/m16314/1.15/>.

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 less 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 2.11
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 2.12
For the two data sets in the test scores example (p. 64), 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 2.13: 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 2.5

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 2.14
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

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. p% 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 15th 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 properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

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.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example 2.15
On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile would 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 2.16
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.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile would be considered good, as answering more questions correctly is desirable.

Example 2.17
At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units

- 70% of students are enrolled in 7 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.

Do the following Practice Problems for Interpreting Percentiles

Exercise 2.6.1 ***(Solution on p. 107.)***

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 2.6.2 ***(Solution on p. 108.)***

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Exercise 2.6.3 ***(Solution on p. 108.)***
On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise 2.6.4 ***(Solution on p. 108.)***
Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Exercise 2.6.5 ***(Solution on p. 108.)***
In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise 2.6.6 ***(Solution on p. 108.)***
In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Exercise 2.6.7 ***(Solution on p. 108.)***

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Exercise 2.6.8 ***(Solution on p. 108.)***

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

**With contributions from Roberta Bloom

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.

Example 2.18

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

[9] This content is available online at <http://http://cnx.org/content/m17102/1.9/>.

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 2.19
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 2.20: 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 2.21
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 Central Limit Theorem**.

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

2.7.2 Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

Table 2.6

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

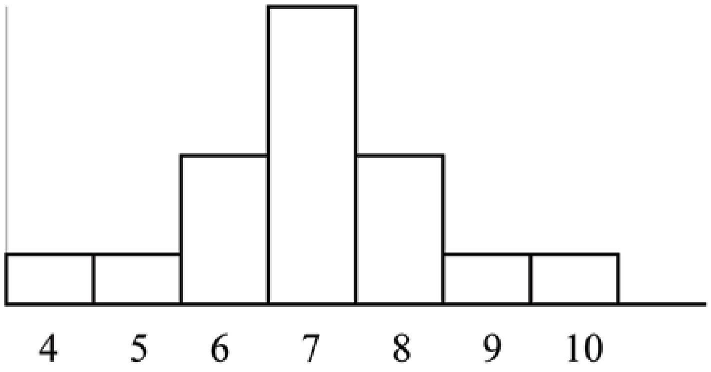
statistic of a sampling distribution is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean $\overline{x}$ is an example of a statistic which estimates the population mean μ.

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.

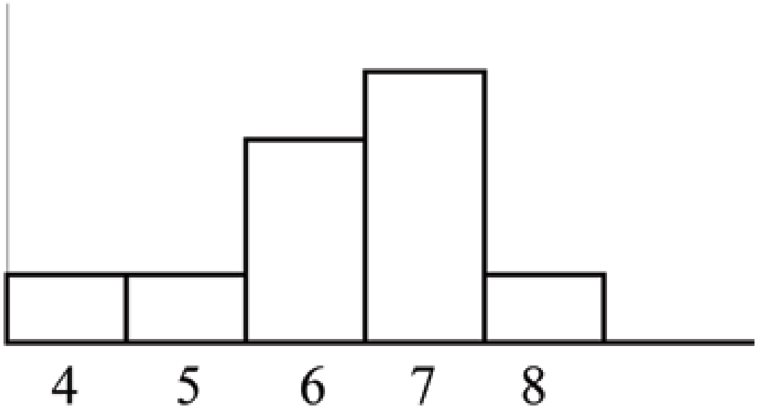


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean, the median, and the mode are often the same.**

The histogram for the data:

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

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



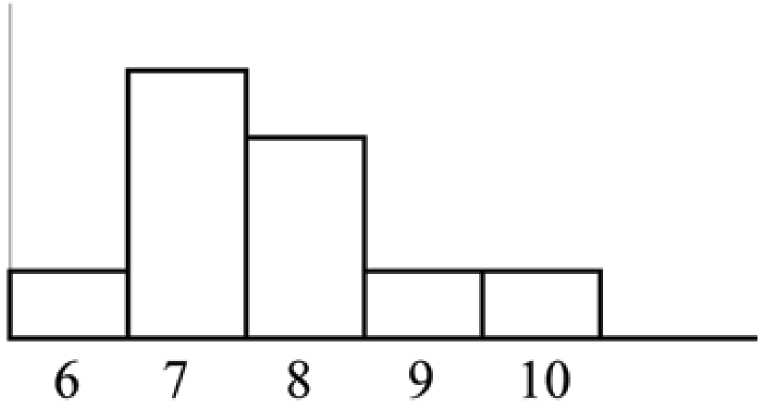
The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

[10] This content is available online at <http://http://cnx.org/content/m17104/1.7/>.

The histogram for the data:

6 ; 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]

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

The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

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

The standard deviation provides a measure of the overall variation in a data set
The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

[11] This content is available online at <http://http://cnx.org/content/m17103/1.12/>.

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

The standard deviation can be used to determine whether a data value is close to or far from the mean. Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

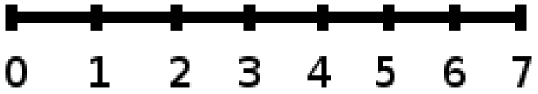
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.
- A wait time that is only one standard deviation from the average is considered close to the average.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

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 because
$5 + (1)(2) = 7$.

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



- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: 7=5+**(1)**(2)
- 1 is **two standard deviations less than the mean** of 5 because: 1=5+**($-$2)**(2)

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

- **sample:** $x = \overline{x} + (\#ofSTDEV)(s)$

- **Population:** $x = \mu + (\#ofSTDEV)\,(\sigma)$

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

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

Calculating the Standard Deviation
If x is a data value, 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 data is for a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \overline{x}$.

The procedure to calculate the standard deviation depends on whether the data is for the entire population or comes from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is 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 an **average of the squares of the deviations** (the $x- \ \overline{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the data is from a **population**, 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 is 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. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

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

Formulas for the Population Standard Deviation

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

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is 1. If a value appears three times in the data set or population, f is 3.

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

NOTE: **In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ or s from the summary statistics.** We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example 2.22

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

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

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

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 2.7

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.

Solution

For the TI-83,83+,84+, enter data into the list editor.
Put the data values in list L1 and the frequencies in list L2.
STAT CALC 1-VarStats L1, L2

$\overline{x}$=10.525
Use Sx because this is sample data (not a population): Sx=.715891

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: x = $\overline{x}$ + (#ofSTDEVs)(s)
- For a population: x = μ + (#ofSTDEVs)(σ)
- For this example, use x = $\overline{x}$ + (#ofSTDEVs)(s) because the data is from a sample

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 standard deviation calculation shown in the table
The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero**. (For this example, there are n=20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

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

Notice that instead of dividing by n=20, the calculation divided by n-1=20-1=19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one $(n - 1)$. Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population variance.

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

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is 0, 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.

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 2.23
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 2.8

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.

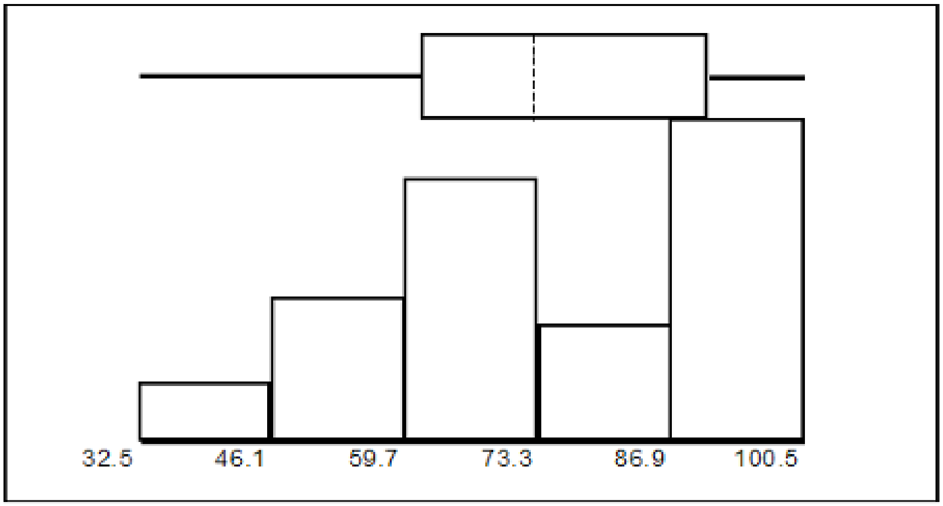


Figure 2.1

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.

Comparing Values from Different Data Sets

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

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

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

Sample	$x = \overline{x} + z\,s$	$z = \frac{x-\overline{x}}{s}$
Population	$x = \mu + z\,\sigma$	$z = \frac{x-\mu}{\sigma}$

Table 2.9

Example 2.24
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 2.10

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

$\#ofSTDEVs = \frac{value-mean}{standard\ deviation}$; $z = \frac{x-\mu}{\sigma}$

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

For Ali, z = #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.

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

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

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

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

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

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

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 = μ + (#ofSTDEVs)(σ)

[12] This content is available online at <http://http://cnx.org/content/m16310/1.9/>.