(continued)

AL Team	*Home Runs*	*NL Team*	*Home Runs*
Chicago	162	Cincinnati	209
Cleveland	209	Colorado	223
Detroit	212	Florida	128
Kansas City	151	Houston	168
Minnesota	105	Los Angeles	187
New York	193	Milwaukee	165
Oakland	235	Montreal	163
Seattle	244	New York	179
Tampa Bay	145	Philadelphia	161
Texas	230	Pittsburgh	171
Toronto	212	San Diego	153
		San Francisco	188
		St. Louis	194

(a) Calculate the mean number of home runs for the American League teams, and then calculate the mean number of home runs for the National League teams.

(b) Calculate the median number of home runs for the American League teams, and then calculate the median number of home runs for the National League teams.

(c) Based on your results from (a) and (b), do you feel that the number of home runs hit in the American League and in the National League are substantially different?

MINI PROJECT

Again, use the data collected in the mini-project in Section 1.1. For each of the following, calculate (a) the mean, (b) the median, (c) the first quartile, (d) the third quartile, (e) the midrange, and (f) the mode.

- Age
- Male height
- Female height
- Number of units enrolled in this term
- Number of hours worked per week (outside of class)

SECTION 2.2
Measures of Dispersion

Although measures of central tendency are important when we are trying to describe a set of data, they are not enough. Two sets of data may be centered in the same location, but still be totally different sets of data. Another important characteristic to consider is how the data are dispersed or spread out. In some sets the values are closely grouped together, while in others they are far apart from each other.

Here are the five test scores of two statistics students.

Jack	85	70	55	41	99
Jill	72	68	70	65	75

The mean score for each student is 70, but the two sets of scores are different. Jill's scores are close together; no score is more than 5 points away from her mean score. Jack's scores are more spread out. Two of his scores are 29 points away from his mean score. We can expect Jill to score closer to 70 on the next test than Jack.

Range

The first measure of dispersion that we will discuss is the **range**. To find the range for a set of data, we subtract the lowest value from the highest value.

Range = highest − lowest

EXAMPLE 2.20

Here are the rushing totals (in yards) for Barry Sanders in the years 1989 through 1998. Find the range for these ten seasons.

1470	1304	1548	1352	1115
1883	1500	1553	2053	1491

The highest value is 2053 yards, and the lowest is 1115 yards.

$$\begin{aligned} \text{Range} &= 2053 - 1115 \\ &= 938 \end{aligned}$$

The range for this set of data is 938 yards. ■

The range tells us how far apart the lowest and highest values are. One problem with the range is that it is sensitive to outliers. If a set of data has an outlier, the range uses it in its calculation. Another problem with the range is that two sets of data can be spread out in totally different fashions, but have the same range.

Here are the scores of two golfers from last month's matches.

Jack	71	72	73	71	73	82	70	72	68
Greg	75	73	77	78	78	81	74	71	85

The range for both golfers is 14 strokes, but the scores are dispersed in a much different fashion. Jack's scores are closely grouped together between 68 and 73, with an outlier at 82. Greg's scores are evenly dispersed from 71 to 85.

The range should be used only as a first step in investigating the dispersion for a set of data. Although it can give us an idea about how spread out the values are, it cannot paint the whole picture for us.

Interquartile Range

Another measure of dispersion is the interquartile range, the distance between the first and third quartiles.

Interquartile range = $Q_3 - Q_1$

Instead of telling us how far apart the two extreme values are, it tells us the range in which we can find the middle 50% of the values. It is not sensitive to outliers.

EXAMPLE 2.21

Here are the SAT math scores for 19 randomly selected students. Find the interquartile range.

480	370	540	660	650	710	470
490	630	390	430	320	470	400
430	570	450	470	530		

Recall that the first step is to put the values in ascending order, and find the median. Then to find the first quartile we find the median of the first group. The third quartile is the median of the second group.

First Group	320	370	390	400	430	430	450	470	470
				Median:	470				
Second Group	480	490	530	540	570	630	650	650	710

Recall from the last section that the first quartile is 430, and the third quartile is 570.

570 − 430 = 140

The interquartile range for these scores is 140 points. ■

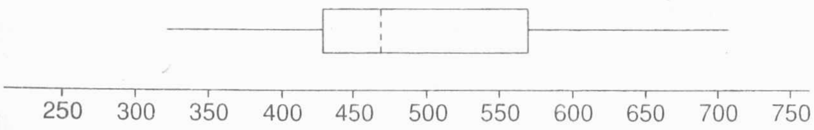
Boxplot

One more way to represent a set of data graphically is with a **boxplot**. A boxplot needs five values for its construction: the lowest value, the first quartile, the median, the third quartile, and the highest value. These five values are often referred to as the **five-number summary** for a set of data. Above a horizontal axis, we draw a box from the first quartile to the third quartile. We put a dashed line in the box at the median. Finally, extend out line segments from the box to the lowest and highest values.

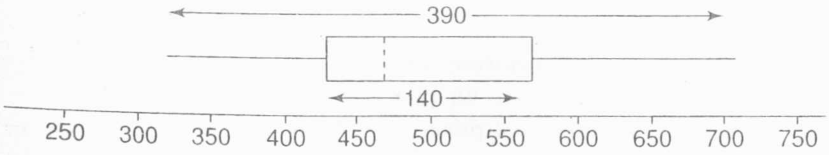
For the previous 19 test scores, here is the five-number summary.

Lowest	Q_1	*Median*	Q_3	*Highest*
320	430	470	570	710

Here is a boxplot for this set of data.



The range is the distance from the far left to the far right, and the interquartile range is the width of the box.



EXAMPLE 2.22

Each year, *Forbes* magazine prints a list of the 400 richest people/families in the world. From their 1999 issue, here are the ages of the 54 richest individuals. All of these individuals have a net worth of at least $5 billion. Construct a boxplot for these data.

43	68	46	43	34	59	55	42
73	71	73	75	71	84	43	36
68	54	77	76	62	59	62	69
31	60	82	88	68	59	81	51
75	79	51	62	41	50	74	82
61	56	66	75	52	59	44	58
51	58	72	62	72	50		

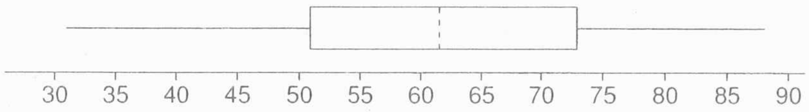
We will begin with the stem-and-leaf display.

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

Here is the five-number summary; verify it with the work from the previous section.

Lowest	Q_1	*Median*	Q_3	*Highest*
31	51	61.5	73	88

The range is 57 years, and the interquartile range is 22 years. Here is the boxplot. ■



Although the range and interquartile range can give us a good idea of a set of data's dispersion, they do have their drawbacks. They do not use all of the data values in their calculation. They measure dispersion from one end to the other, measuring the width between two values. Another way to look at dispersion is by examining how far the values are from the center of the set. For instance, an important characteristic of a new home is its distance from the center of town. We will continue now by examining three measures that determine dispersion in this fashion.

Mean Deviation

To find the **mean deviation** for a set of data, we begin by finding the distance from each value to the mean. We then find the mean of these distances. This is the mean deviation. Roughly interpreted, it tells us how far from the center of the data the values are on average. A measure of distance is nonnegative, so to calculate each distance we take the absolute value of the difference between the value and the mean. Here is the formula.

$$\text{Mean deviation} = \frac{\Sigma|x - \bar{x}|}{n}$$

ection.

:plot. ■

0

f a set of
data val-
measur-
xamining
t charac-
ie now by

ance from
the mean
data the
each dis-
the mean.

This formula uses sample notation, but the procedure is exactly the same for the mean deviation of a population. Simply replace $\bar{x}$ by μ, and n by N. Here are the steps for this calculation.

1. Find the mean.
2. Subtract the mean from each value.
3. Take the absolute value of each difference.
4. Total these distances.
5. Divide the total by the number of values in the set.

EXAMPLE 2.23

A taxi dispatcher is interested in the number of fares for his drivers on Fridays. He randomly selects seven drivers, and then randomly selects one Friday for each of the drivers. Here are the number of fares for each. Find the mean deviation for these totals.

32 27 30 41 29 38 34

For this calculation it is best to use a column approach.

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
32	−1	1
27	−6	6
30	−3	3
41	8	8
29	−4	4
38	5	5
34	1	1
$\bar{x} = 33$		28

$$\text{Mean deviation} = \frac{28}{7}$$
$$= 4$$

The mean deviation is 4 fares. We can say that on average, the values are 4 fares away from the mean. ■

We should note one more thing from the previous example. Look at the middle column, labeled $x - \bar{x}$. The sum of that column is 0. This shows us that the sum of the distances to the right of the mean (positive values) is equal to the sum of the distances to the left of the mean (negative values). This is evidence that the mean truly is in the center of the data.

In practice, that middle column is not necessary. Simply find the distance between each value and the mean, and place that in the right column.

EXAMPLE 2.24

Here is a list of the New York Stock Exchange's daily volumes for one week of trading, in millions of shares. Find the mean deviation for these values.

669 754 752 771 835

Since there is no suggestion that we are interested in anything except these five values, we will treat this set of data as a population.

x	$x - \mu$	$\lvert x - \mu \rvert$
669	−87.2	87.2
754	−2.2	2.2
752	−4.2	4.2
771	14.8	14.8
835	78.8	78.8
$\mu = 756.2$		187.2

$$\text{Mean deviation} = \frac{187.2}{5}$$
$$= 37.44$$

The mean deviation is 37.44 million shares. On the average, the values are 37.44 million shares away from the mean. ■

Variance

Another measure of dispersion that measures from the inside out is variance. Variance is similar to mean deviation, with two exceptions. The first difference between these two measures is that we square the difference between each value and the mean, rather than taking the absolute value. The other difference is that there are two formulas, depending on whether we are finding the variance of a sample or the variance of a population. Here are the two formulas.

Sample Variance	**Population Variance**
$s^2 = \dfrac{\Sigma(x - \bar{x})^2}{n - 1}$	$\sigma^2 = \dfrac{\Sigma(x - \mu)^2}{N}$

We use the symbol s^2 to represent sample variance, and the symbol σ^2 to represent population variance. σ is the lowercase Greek letter sigma, which is the Greek letter s.

Note that there is a significant difference in the two formulas. The denominator in the sample variance formula calls for us to subtract 1 from the sample size, whereas the denominator in the population variance uses the population size, without subtracting 1. Why the difference? Although this is beyond the scope of this course, subtracting 1 from the sample size makes the sample variance an *unbiased* estimator of the population variance.

It is crucial that you are able to identify whether a set of data is a sample or a population. Using the sample formula by mistake will produce a variance that is too large. Using the population formula by mistake will produce a variance that is too small.

EXAMPLE 2.25

A taxi dispatcher is interested in the number of fares for his drivers on Fridays. He randomly selects seven drivers, and then randomly selects one Friday for each of the drivers. Here are the number of fares for each. Find the variance for these totals.

32	27	30	41	29	38	34

Since the dispatcher is interested in all Fridays, these data are a sample. For this calculation, just as with mean deviation, it is best to use a column approach.

x	$x - \bar{x}$	$(x - \bar{x})^2$
32	−1	1
27	−6	36
30	−3	9
41	8	64
29	−4	16
38	5	25
34	1	1
$\bar{x} = 33$		152

$$s^2 = \frac{152}{7 - 1}$$

$$= \frac{152}{6}$$

$$= 25.33$$

The variance is 25.33. ■

A drawback for variance is that it lacks the interpretation that mean deviation has. The idea to keep in mind is that the bigger the variance is, the more spread out the values are. An advantage of variance is that its formula does not involve absolute values, which are difficult to manipulate algebraically.

EXAMPLE 2.26

Here is a list of the New York Stock Exchange's daily volumes for one week of trading, in millions of shares. Find the variance for these values.

669 754 752 771 835

Since there is no suggestion that we are interested in anything except these five values, we will treat this set of data as a population.

x	$x - \mu$	$(x - \mu)^2$
669	−87.2	7,603.84
754	−2.2	4.84
752	−4.2	17.64
771	14.8	219.04
835	78.8	6,209.44
$\mu = 756.2$		14,054.8

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

$$= \frac{14{,}054.8}{5}$$

$$= 2{,}810.96$$

The variance is 2,810.96. ■

Standard Deviation

The measure of dispersion that we will use most often in this course is the standard deviation. The **standard deviation** of a set of data is the square root of the variance.

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

We use the letter s to represent sample standard deviation, and σ to represent population standard deviation.

Sample Standard Deviation

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

Population Standard Deviation

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

EXAMPLE 2.27

A taxi dispatcher is interested in the number of fares for his drivers on Fridays. He randomly selects seven drivers, and then randomly selects one Friday for each of the drivers. Here are the number of fares for each. Find the standard deviation for these totals.

32 27 30 41 29 38 34

In a previous example, we calculated the sample variance for these values to be 25.33.

$$s = \sqrt{25.33}$$
$$= 5.03$$

The sample standard deviation is 5.03 fares. ■

EXAMPLE 2.28

Here is a list of the New York Stock Exchange's daily volumes for one week of trading, in millions of shares. Find the variance for these values.

669 754 752 771 835

In a previous example, we calculated the population variance for these values to be 2810.96.

$$\sigma = \sqrt{2810.96}$$
$$= 53.02$$

The standard deviation is 53.02 million shares. ■

EXAMPLE 2.29

A student is interested in how old women are when they first get married. To estimate the mean age, she goes into 11 randomly selected chat rooms, and asks randomly selected women how old they were at their first marriage until she gets a response in each room. Here are the 11 ages. Find the standard deviation of these ages.

21 20 19 16 22 21 21 19 18 24 18

andard
ıriance.

t popu-

drivers
ıdomly
ıber of

əs to be

ıes for
ı for

ıes to be

t get mar-
elected
ey were
ı. Here

24 18

Since the student is interested in the mean age for all women, these 11 values represent a sample.

x	$x - \bar{x}$	$(x - \bar{x})^2$
21	1.1	1.21
20	0.1	0.01
19	–0.9	0.81
16	–3.9	15.21
22	2.1	4.41
21	1.1	1.21
21	1.1	1.21
19	–0.9	0.81
18	–1.9	3.61
24	4.1	16.81
18	–1.9	3.61
$\bar{x}$ = 19.9		48.91

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

$$= \sqrt{\frac{48.91}{10}}$$

$$= 2.21$$

The standard deviation is 2.21 years. ■

EXAMPLE 2.30

On July 15, 1999, Bruce Springsteen and the E Street Band began a series of 15 concerts in Bruce's home state, New Jersey. Fans who attended several of the concerts raved about how different the shows were each night. Here is a list of the number of songs played at each of the 15 concerts. Find the standard deviation.

26	26	24	23	23	23	25	23
22	22	23	22	23	25	24	

Since there is no suggestion that we are interested in anything beyond these 15 concerts, we will treat this as a population.

x	$x - \mu$	$(x - \mu)^2$
26	2.4	5.76
26	2.4	5.76
24	0.4	0.16
23	–0.6	0.36
23	–0.6	0.36
23	–0.6	0.36
25	1.4	1.96
23	–0.6	0.36
22	–1.6	2.56
22	–1.6	2.56
23	–0.6	0.36
22	–1.6	2.56
23	–0.6	0.36
25	1.4	1.96
24	0.4	0.16
μ = 23.6		25.6

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

$$= \sqrt{\frac{25.6}{15}}$$

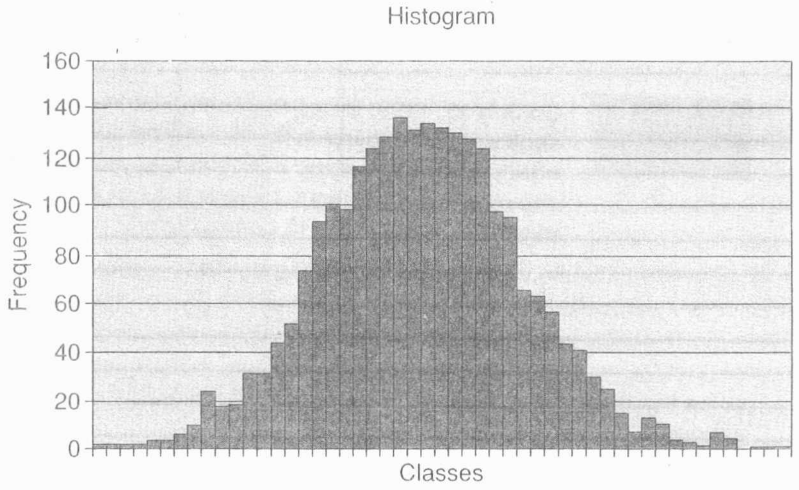
$$= 1.31$$

The standard deviation is 1.31 songs. ■

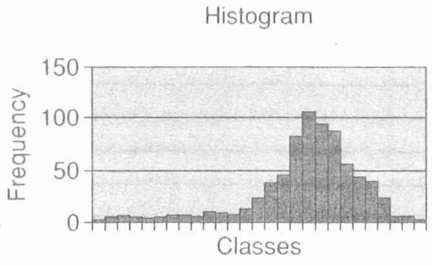
As mentioned before, we should try to take advantage of technology (computer, calculator) when calculating standard deviation. When using the calculator, the process is the reverse of what we have shown by hand. We first calculate the standard deviation, and then if the variance is needed we square the standard deviation.

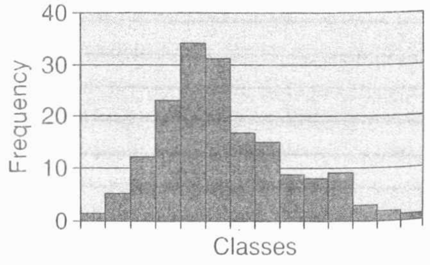
Skewness

Standard deviation is a measure of dispersion that is used often in inferential statistics. We introduce three of its uses here. The first has to do with **skewness.** A set of data is said to be skewed if it is not symmetrical. Here is a histogram from a set of data that is roughly symmetrical.



The peak of the previous histogram is located in the center. A set of data is skewed if we notice that the histogram is stretched to the left or stretched to the right, such as in the following two histograms.

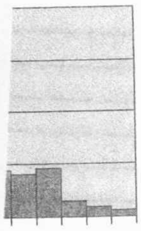




ımputer,
ıtor, the
ıtandard
on.

ial statis-
et of data
data that

ı is skewed if
ight, such as



A set of data that is stretched to the left is **negatively skewed**. A set of data is negatively skewed if the values to the left of the median are more spread out than the values on the right side of the median. A low outlier(s) can cause a set of data to be negatively skewed. In such a situation, the mean will be less than the median. (Why?) If a set of data is stretched to the right, we say that it is **positively skewed**.

We have a measure that calculates how skewed a set of data is—the **coefficient of skewness**. Here is the formula.

$$sk = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

If a set of data is positively skewed, its mean will be greater than its median, causing the coefficient of skewness to be positive. A negative coefficient of skewness indicates that a set of data is negatively skewed. The coefficient of skewness will be a number between –3 and 3. The closer it is to 0, the more symmetric the data are. The farther away from 0 the coefficient of skewness is, the more skewed the data are.

EXAMPLE 2.31

Here are the ages of ten randomly selected women at a college orientation. Find the coefficient of skewness for these data.

18 25 31 19 22 21 19 25 18 27

The data represent a sample. To calculate the coefficient of skewness we need to find the mean, median, and standard deviation. First the mean.

$$\begin{aligned}\bar{x} &= \frac{\Sigma x}{n}\\ &= \frac{225}{10}\\ &= 22.5\end{aligned}$$

Next, find the median.

18 18 19 19 21   22 25 25 27 31

$$\begin{aligned}\text{Median} &= \frac{21 + 22}{2}\\ &= 21.5\end{aligned}$$

Finally, find the standard deviation. You could use your calculator for this; however, you should also be able to calculate standard deviation through the use of the formula.

x	$x - \bar{x}$	$(x - \bar{x})^2$
18	–4.5	20.25
18	–4.5	20.25
19	–3.5	12.25
19	–3.5	12.25
21	–1.5	2.25
22	–0.5	0.25
25	2.5	6.25
25	2.5	6.25
27	4.5	20.25
31	8.5	72.25
$\bar{x}$ = 22.5		172.5

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

$$= \sqrt{\frac{172.5}{9}}$$

$$= 4.38$$

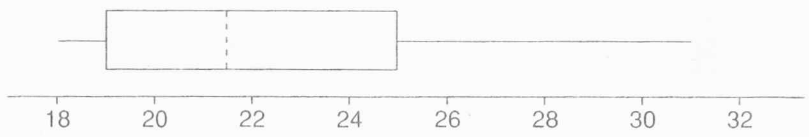
Now we can calculate the coefficient of skewness.

$$sk = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

$$= \frac{3(22.5 - 21.5)}{4.38}$$

$$= 0.68$$

These data have a coefficient of skewness of 0.68. They are moderately positively skewed. Here is a boxplot to demonstrate the skewness.



Note that the side to the right of the median is more stretched out than the side to the left of the median. ■

Standard Units

Standard deviation can be used to help us compare two values from two different sets of data. For instance, suppose a man is 6′4″ tall, and his IQ is 130. Both of these values are above the mean, but which stands out the most? The mean height of adult males is 69.0 inches (5′9″), so the man is 7 inches above the mean height. The mean IQ is 100 points, so the man's IQ is 30 points above the mean IQ. We cannot compare the 7 inches to the 30 points, because these are based on two different scales. A fair comparison would be to compare how many standard deviations above the mean these two values are. Finding how many standard deviations away from the mean a value is converts the value to **standard units**, or its ***z*-score**. To convert a value to standard units, we subtract the mean from the value, and then divide by the standard deviation.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

Adult male heights have a mean of 69.0 inches, with a standard deviation of 2.8 inches. Here is the *z*-score that corresponds to 6′4″:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

$$= \frac{76 - 69.0}{2.8}$$

$$= 2.5$$