Example For the six values recorded in the nurse-in-training study, 10, 6, 5, 14, 6, and 13, find the mode.

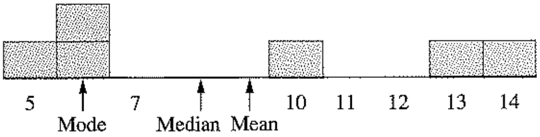
Solution Mode = 6 minutes since this is the most frequently occurring value.

Unfortunately, not all data sets have modes. On the other hand, some sets have more than one. If two modes occur, we refer to the data as **bimodal.** If more than two modes occur, we refer to the data as **multimodal.** ■

The mode, as you can see from the above example, can sometimes be misleading as to the true central tendency of data. Although useful when used in addition to the median and mean, the mode should be viewed with caution when used alone.

Comparison of the Mean, Median, and Mode

The following provides a visual look at the mean, median, and mode using the nurse-in-training data. If you recall, we observed the nurse six times over a period of several months and recorded how long the nurse took to draw a specific series of blood specimens.



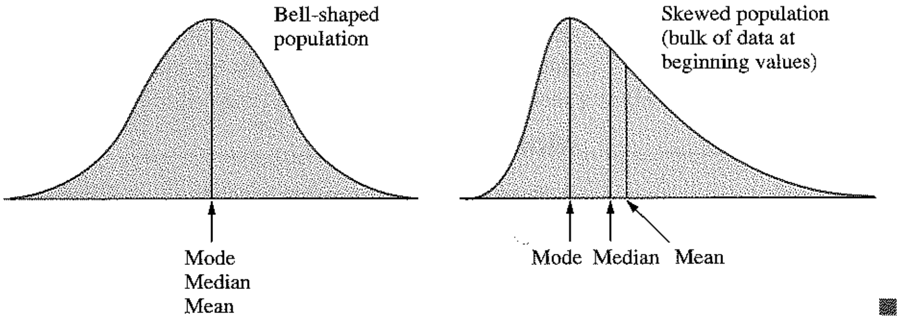
Which do you feel gives a better measure of central tendency for this data? Although the mean is the preferred measure, each adds a little more information.

Example Estimate the position of the mean, median, and mode for a

a. bell-shaped population.

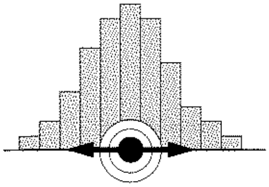
b. skewed population, with the bulk of data at beginning values.

Solution



■

2.3 Measures of Dispersion or Spread (Ungrouped Data)



Whereas measures of central tendency attempt to locate the center or middle of the data, measures of dispersion are designed to measure how widely scattered or spread out the data is. We will study two such measures, the range and standard deviation.

Range

The **range** is the difference between the high and low value in your data set.

> ***Range***
> High value minus low value.

Example For the six values recorded in the nurse-in-training study, 10, 6, 5, 14, 6, and 13 minutes, find the range.

Solution Range = high value minus low value = 14 − 5 = 9

Range = 9 minutes

This can also be expressed as: The data ranged from 5 to 14 minutes. ■

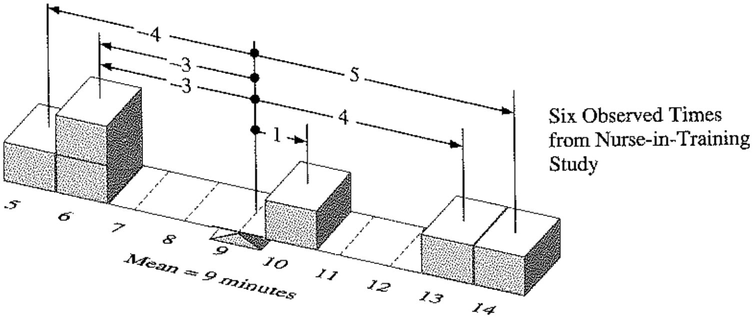
Although easy to compute, the range offers no information about the distribution of data between these two extremes of high and low value and, thus, the range is mostly used as a rough gauge in determining dispersion or spread.

Standard Deviation

> ***Standard Deviation***
> A form of average distance from the mean.

The **standard deviation** is a more complex measure and perhaps best explained through the following example.

Suppose we were to represent the six observations in our nurse-in-training study, 5, 6, 6, 10, 13, and 14, as blocks on a number line, as follows:



Notice in the above diagram, we are measuring the distance (in minutes) each value is away from the mean. Don't read on. Please look at the diagram until you understand it.

For instance, the value recorded as 13 is how many minutes away from the mean? The answer is 4 (13 − 9 = 4). Symbolically, we would represent this as follows:

$$x - \bar{x} = \text{distance from mean}$$
$$13 - 9 = 4 \text{ minutes}$$

The symbol x represents one value in our data, and $\bar{x}$ is the sample mean or average. When you subtract the two, $x - \bar{x}$, you get the distance a value is away from the mean. To calculate all the distances we use the following chart:

x	$\bar{x}$	$x - \bar{x}$
5	9	−4
6	9	−3
6	9	−3
10	9	1
13	9	4
14	9	5

Now, if we wish to calculate the *average distance from the mean,* we simply add up all the distances and divide by n, the number of readings, which is six.

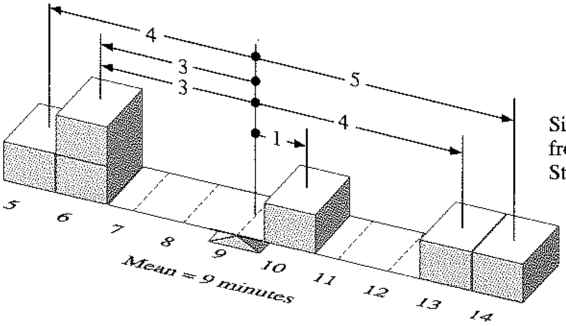
$$\begin{matrix}\text{Average Distance}\\ \text{from the Mean}\end{matrix} = \frac{-4 - 3 - 3 + 1 + 4 + 5}{6} = \frac{0}{6} = 0 \text{ minutes}$$

Unfortunately, if we leave in the negative signs, we always get 0. Notice that the negative values cancel out the positive values. This always occurs. However, since we are interested in the absolute distance each value is away from the mean and not whether it's plus or minus, a simple solution would be to use the distances without the negative signs. To do this, we take the **absolute value** of the distances, as follows:

x	$\bar{x}$	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$
5	9	−4	4
6	9	−3	3
6	9	−3	3
10	9	1	1
13	9	4	4
14	9	5	5

Note that the symbols | |, called *absolute value lines,* allow us to record the distances as positive values.*

Pictorially, we can represent the distances as follows:



Six Observed Times from Nurse-in-Training Study

Now, to calculate the *average distance from the mean,* we would add up the distances and divide by *n,* the number of readings.

$$\text{Average Distance from the Mean} = \frac{\Sigma \lvert x - \bar{x} \rvert}{n}$$

$$= \frac{4 + 3 + 3 + 1 + 4 + 5}{6} = \frac{20}{6} = 3\frac{2}{6}$$

$$= 3\frac{1}{3} \text{ minutes}$$

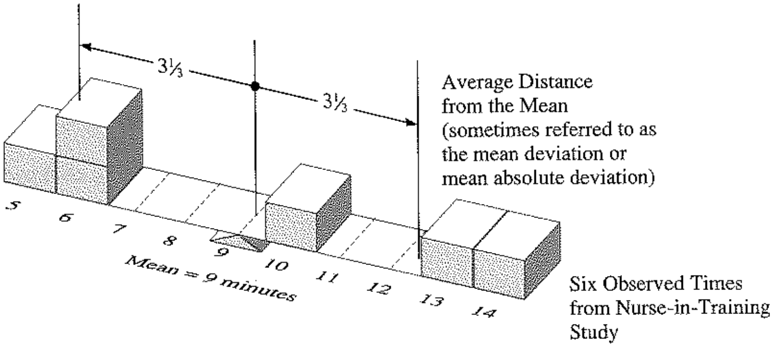
We now have a measure of how scattered or spread out the data is. We can now say, the average distance a value is from the mean is $3\frac{1}{3}$ minutes.

This measure of spread (along with others) was widely used in the 1800s and referred to by a number of names. Today, the value is most often called the

*Technical note: The absolute value of a number *n*, represented by $\lvert n \rvert$, is formally defined as its distance from 0 on the number line without considering direction. For example, $\lvert -3 \rvert = 3$, $\lvert 5 \rvert = 5$.

mean deviation or mean absolute deviation (m.a.d.), however we shall simply refer to it as the *average distance from the mean* because, in effect, that is what it is.

Pictorially, this measure of spread might be represented as follows:


3⅓
3⅓
Average Distance from the Mean (sometimes referred to as the mean deviation or mean absolute deviation)
5
6
7
8
9
10
11
12
13
14
Mean = 9 minutes
Six Observed Times from Nurse-in-Training Study

Despite its simplicity, this method for measuring spread is not in wide use today. The measure has a number of disadvantages, not the least of which are the computational difficulties in advanced work. As statisticians tried to apply this measure of spread to inferential statistics (using samples to estimate population characteristics), the problems grew unmanageable, mostly stemming from the use of the absolute value lines to remove the negative signs. For instance, the process of using absolute distances, $|x - \overline{x}|$'s, to remove the negative signs in complex calculations necessitates breaking up each problem into three cases: (a) when $x < \overline{x}$, (b) when $x = \overline{x}$ and (c) when $x > \overline{x}$. If we pool the absolute distances of many samples, as we do in later work, this mushrooms into a laborious effort. Other difficulties also arose involving discontinuous derivatives when calculus was applied. In other words, it was a statistician's nightmare.

In addition to this practical problem, it has been shown that this measure of spread was not as efficient in estimating the population spread as were other measures. That is, when we calculate the average distance from the mean for many samples drawn from the same population, these values would scatter more loosely around the *true* population value than if we had used other types of measures. If our goal is to use sample data to estimate population characteristics, we want our sample value to be the most efficient estimator of its equivalent population value* and the average distance from the mean was just not as efficient an estimator of its equivalent population value as other available measures.

*The term *most efficient* means sample values cluster closer to the population value.

Fortunately, the work of two great mathematicians of the early 1800s, Legendre (1805) and Gauss (1809, 1823), led to the development and ultimate adoption of another form of average distance from the mean, called the **standard deviation.***

The process of calculating the standard deviation involves squaring each distance, which removes the negative sign, for example, $(-3)^2 = +9$, and thus eliminates the need for the absolute value lines, which greatly reduces the computational difficulties in advanced work. In addition, using this new measure, the sample spread value becomes a more efficient estimator of the population spread value. In other words, sample values now cluster more tightly around the population value. Let's see how it works.

Calculation of Standard Deviation

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
5	9	−4	16
6	9	−3	9
6	9	−3	9
10	9	1	1
13	9	4	16
14	9	5	25

First, we square each distance. For example, the first distance, −4, is squared as follows: (−4)(−4) = +16.

$$\text{Average squared distance (variance)} = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

$$= \frac{16+9+9+1+16+25}{6-1}$$

$$s^2 = \frac{76}{5}$$

$$= 15.2 \text{ squared minutes}$$

Second, we take the average squared distance by summing the squared distances, then dividing† by $n-1$. This averaged squared distance is referred to as s^2, or the *variance*. Notice that this value, 15.2, is in the units of squared minutes

$$\text{Standard deviation of sample} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$s = \sqrt{15.2}$$

$$= 3.899$$

$$= 3.9 \text{ minutes (rounded)}$$

Third, to convert squared minutes back to minutes, we take the square root.‡

*For further historical discussion, refer to chapter 9, section 9.0, under the subheading "Least-Squares Analysis."

†Note that we divided by $n - 1$ (and not n) in our formula for *sample* standard deviation. Experience has shown that dividing by $n - 1$ slightly raises the value of the sample standard deviation and provides, on average, a more accurate estimator of the population standard deviation than if we had divided by n. This has been proven by both experience and theory. However, if we were to calculate the standard deviation of a *population,* we would simply divide by n.

‡Technical note: Actually s^2 (and not s) is the preferred measure of spread, since s^2 is an unbiased estimator of (meaning: on average, equal to) the equivalent population value σ^2. Unfortunately, this is not the case with s. On average, $s \neq \sigma$, however for large samples, the bias is quite small and usually ignored. We will ignore this consideration until later in the text.

The process can be summarized by the following formula:

For Ungrouped Data

$$\text{Sample Standard Deviation} = \sqrt{\frac{\text{sum of the squared distances}}{\text{number of readings minus one}}}$$

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

Let's see how all this might work in an example.

Example

In a medical study, a researcher wished to estimate the average length of time needed for a particular nurse-in-training to draw a series of blood specimens. A sample of the nurse's work over several months yielded the following times: 10, 6, 5, 14, 6, and 13 (in minutes). Calculate the standard deviation.

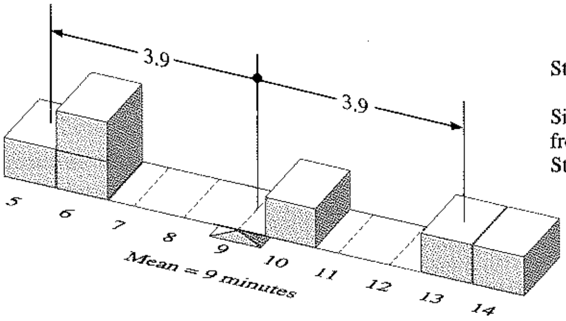
Solution

We would organize our data in chart form and calculate the standard deviation as follows.

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
5	9	−4	16
6	9	−3	9
6	9	−3	9
10	9	1	1
13	9	4	16
14	9	5	25
		$\Sigma(x-\bar{x})^2 =$	76

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{76}{6-1}}$$
$$= \sqrt{15.2} = 3.899$$
$$= 3.9 \text{ minutes}$$

Pictorially, we might represent the results as follows.



Standard Deviation

Six Observed Times from Nurse-in-Training Study

*The standard deviation may also be calculated with the formula

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

known as *the counting formula*. Note that this formula requires only the sum of the x and x^2 columns, however it offers little in the way of understanding the process.

Notice that the standard deviation is still a form of average distance a value is away from the mean, even though we squared the distances, divided by $n - 1$, and took the square root.

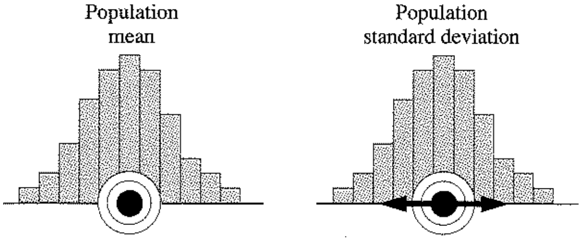
Advantages of Using the Standard Deviation as our Measure of Spread

1. The negative signs are removed without use of the absolute value lines, which greatly reduces the computational difficulties in advanced work.
2. The standard deviation is a more efficient estimator of spread than the average distance from the mean.
3. The standard deviation is a *considerably* more efficient estimator of spread than many other measures considered, such as the median deviation.

   Note: The term *more efficient* means sample values cluster more tightly around the population value—that is, on average, sample values give closer approximations to the population value. For further reading on the standard deviation, refer to section 9.0 under the subheading ''Least-Squares Analysis''; also refer to endnote 16 in chapter 9.

2.4 Estimating Population Characteristics

The purpose of a sample is usually to gather information about a population. Two of the characteristics of a population we most frequently wish to know are the



Unfortunately, small samples (under 30 observations) will sometimes give unreliable approximations of population characteristics, depending on the *shape* of your population histogram. (This will be discussed in greater detail in chapters 7 and 8).

One way to avoid this problem is to keep your sample size at 30 or more observations. If our sample size, *n*, is kept at 30 or more observations, we do not have to worry about the shape of our population. Results from sample sizes of 30 or more give reliable information about population characteristics for almost any shaped population. So, with this in mind, we can state, *it is preferable to keep your sample size at 30 or more.* More specifically, we can state the following.

For a valid random sample of 30 or more observations, drawn from almost any population

$\bar{x} \approx \mu$

The sample average, $\bar{x}$, will be approximately equal to the population average,

μ (mū).

$s \approx \sigma$

The sample standard deviation, *s*, will be approximately equal to the population standard deviation, σ (sigma).

We can also use samples of *smaller than 30* observations, but in that case, we must be assured our population is at least somewhat bell-shaped. In bell-shaped populations, small samples give reliable approximations of population characteristics, or at least reliable enough that after certain adjustments, reasonable conclusions can be drawn. So,

For a valid random sample of *under 30 observations*, we must be assured our population is at least somewhat bell-shaped.

In the preceding examples, we sometimes used samples as small as five or six observations. If our population was somewhat bell-shaped, this would be perfectly okay. However, if our population was far from bell-shaped (let's say, for instance, extremely skewed), then we can *not* depend on these samples to give reliable estimates of population characteristics.

Skip

2.5 Measures of Central Tendency and Dispersion/Spread (Grouped Data)

When we work with large bodies of data, it is sometimes more efficient to group the data into categories. In the following example, we will calculate the mean and standard deviation for such data.

Mean

Example

Say in our nurse-in-training study, the researcher took 36 observations of the nurse. Only now, instead of recording individual times, the researcher chose to record the times as part of a category or group, as follows:

Time Category (in minutes)	**Number of Observations (tally)**
3–5	𝍸 |
6–8	𝍸 𝍸 |
9–11	𝍸 ||
12–14	𝍸 ||||
15–17	|||

Calculate the mean.

over a century of practical experience that valid random samples of 30 or more will give reliable information about μ and σ.

More specifically we can state, *for a valid random sample of 30 or more observations, drawn from almost any shaped population.*

$\bar{x} \approx \mu$: the sample average, $\bar{x}$, will be approximately equal to the population average, μ (mū).

$s \approx \sigma$: the sample standard deviation, s, will be approximately equal to the population standard deviation, σ (sigma).

Samples of under 30 observations can also be used, however *for a valid random sample of under 30 observations, we must be further assured our population is at least somewhat bell-shaped.*

In inferential statistics, we often refer to a particular value in terms of its relative position to the mean. In many instances we use something called a *z* score.

z score: The number of standard deviations a value is away from the mean. For instance, if $\bar{x} = 14$ and $s = 4$, then a value of 18 would be one standard deviation above the mean, expressed as $z = 1.00$.

Exercises

Note that full answers for exercises 1–5 and abbreviated answers for odd-numbered exercises are provided in the Answer Key.

2.1 Countrygirl Makeup is a line of facial products marketed to young women, ages sixteen to twenty-four. Although successful when introduced in 1980, in recent years Countrygirl sales have eroded. An independent research firm was commissioned to gather information about the ages of current users. A nationwide random sample of 50 users yielded the following ages:

19.4	31.3	22.7	27.6	27.9
30.1	23.1	26.4	32.1	22.5
28.2	25.7	33.8	28.9	18.6
26.8	30.5	34.0	21.6	28.2
32.2	27.3	17.5	23.0	32.8
36.0	29.1	42.7	30.5	39.0
26.2	33.2	36.3	22.7	43.1
28.7	26.3	38.6	24.1	21.3
32.1	28.7	25.8	26.0	18.7
18.2	23.9	28.2	20.2	33.1

Age Category	**Number of Observations (tally)**
15–19	
20–24	
25–29	
30–34	
35–39	
40–44	

a. Tally the data into the age categories given above.

b. Construct separately a histogram and frequency polygon.

c. What is the population?

d. What is the sample? How large is *n*, the sample size?

2.2 In exercise 2.1, 21 out of 50 possess the attribute of brown hair.

a. What proportion (or fraction) of the sample, p_s, possesses the attribute of brown hair?

b. Convert this fraction to a percentage.

c. Represent this proportion in a circle graph.

2.3 In a medical study, a researcher wished to estimate the average length of time needed for a particular nurse-in-training to draw a series of blood specimens. A sample of the nurse's work over several months yielded the following times: 11, 7, 13, 7, 5, 8, 10, and 15 (in minutes).

Calculate:

a. Mean

b. Median

c. Mode

d. Range

e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.4 Suppose in our nurse-in-training study, the researcher took 49 observations of the nurse. Only now, instead of recording individual times, the researcher chose to record the times as part of a category or group, as follows.

Calculate:

a. Mean
b. Standard deviation
c. Modal class

Construct:

d. Histogram
e. Frequency polygon

Time category (in minutes)	Number of observations (tally)
3–5	𝍸 ||
6–8	𝍸 𝍸 𝍸
9–11	𝍸 𝍸 𝍸 |
12–14	𝍸 |||
15–17	|||

2.5 Suppose we sampled from a grove of recently planted Indiana poplar trees and measured their heights, obtaining the following:

$\bar{x}$ = 14 feet (average height of trees sampled)
s = 4 feet (standard deviation of trees sampled)

Estimate the z score for a poplar tree of the following height.

a. 20 feet
b. 10 feet
c. 15 feet

2.6 In a study on shyness at the University of Iowa, a team of psychologists asked 7 participants to rank the anxiety created by being at a party with strangers on a scale from 0 (no anxiety) to 20 (maximum anxiety), yielding the following scores: 17, 19, 16, 14, 19, 15, and 12.

Calculate:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.7 In an educational study of second-graders in Westchester County, N.Y., it took 5 students the following times (in seconds) to put together a simple puzzle: 12, 14, 7, 13, and 9.

Calculate:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.8 A sample of n = 100 Countrygirl Makeup users were randomly sampled nationwide, yielding the following ages (taken from the demonstration problem at the beginning of this chapter).

Calculate:

a. Mean
b. Standard deviation
c. Modal class
d. If Countrygirl is currently marketed to 16–24 year olds, and if we managed to achieve a valid random sample above, what effect might these results have on future advertising?

Age category	Number of observations (tally)
15–19	𝍸 𝍸 |
20–24	𝍸 𝍸 𝍸 𝍸 ||||
25–29	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
30–34	𝍸 𝍸 𝍸 |||
35–39	𝍸 𝍸 |
40–44	𝍸
45–49	|
	Total, 100 users

2.9 At n = 70 boutiques randomly selected throughout the New England sales district, the following number of Rolf Laurie designer bed comforters sold last year were

Calculate:

a. Mean
b. Standard deviation
c. Modal class

Construct:

d. Histogram
e. Frequency polygon

Rolf Laurie comforters sold (last year)	Number of New England boutiques (tally)
0–14	𝍸
15–29	𝍸 𝍸 𝍸 ||||
30–44	𝍸 𝍸 𝍸 𝍸 |||
45–59	𝍸 𝍸
60–74	𝍸 |||
75–89	𝍸
	Total, 70 boutiques

2.10 If the average amount wagered per person in state lotteries in a particular year was μ = \$100 with standard deviation σ = \$12, find the z score for a person who wagered

a. \$130.
b. \$96.

2.11 A vending machine is known to fill cups to an average of $\mu = 7.0$ ounces with standard deviation $\sigma = .4$ ounces. Find the z score for a person that gets a cup with

a. 7.6 ounces.
b. 7.1 ounces.
c. 6.7 ounces.

2.12 Say you took five quizzes in Economics and your average was 76. Four of the five quizzes were graded, 80, 72, 86, and 70; however, one quiz grade was lost. Use the formula for calculating an average to determine the missing grade.

2.13 A young couple, Jason and Rebecca, weighed 170 lbs and 138 lbs, respectively. For their age and body structure, a medical association published the following guidelines:

Male	**Female**
$\mu = 155$	$\mu = 118$
$\sigma = 12$	$\sigma = 10$

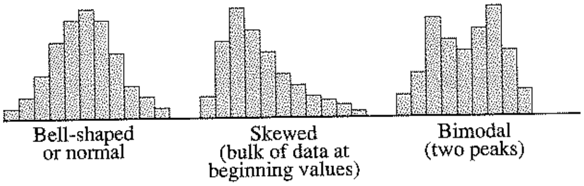
a. Calculate the appropriate z scores for each.
b. According to these guidelines, who would be considered more seriously overweight?

2.14 The Jontnas Company has seven employees with the following annual salaries: \$20,000, \$20,000, \$20,000, \$40,000, \$40,000, \$40,000, and \$240,000.

a. Calculate the mean and median salary.
b. Which do you feel would be a more realistic measure of central tendency, the mean or median salary?

2.15 Experience has shown that many *population* histograms take on a similar appearance. In other words, there are certain common or popular shapes that seem to reoccur.

List population values that might take on each of the following shapes.



2.16 For $n = 200$ randomly selected New England boutiques, 36 carried Rolf Laurie designer blouses.

a. What proportion (or fraction) of the sample, p_s, carries Rolf Laurie designer blouses?
b. Convert this fraction to a percentage.
c. Represent this proportion in a circle graph.

2.17 A West Coast professor's definition of good character includes the following two qualities, "empathy, meaning regards for the needs, rights, and feelings of others, and self-control, meaning the ability to act with reference to the more distant consequences of current behavior." Suppose a test evaluating good character was administered to twenty-one local politicians, resulting in the following scores (10–49 scale):

18 26 23 21 26 29 30 33 32 34
38 38 36 37 41 40 47 41 42 43

a. Construct both a tally and a pictogram (invent your own symbol) using the following categories: 10–19, 20–29, 30–39, and 40–49.
b. Construct a stem-and-leaf display.
c. Construct a box-and-whisker plot.
d. Locate the quartile points, Q_1, Q_2, and Q_3.

(Reference article is *Newsweek,* "A Sterner Kind of Caring," January 13, 1992, p. 68.)