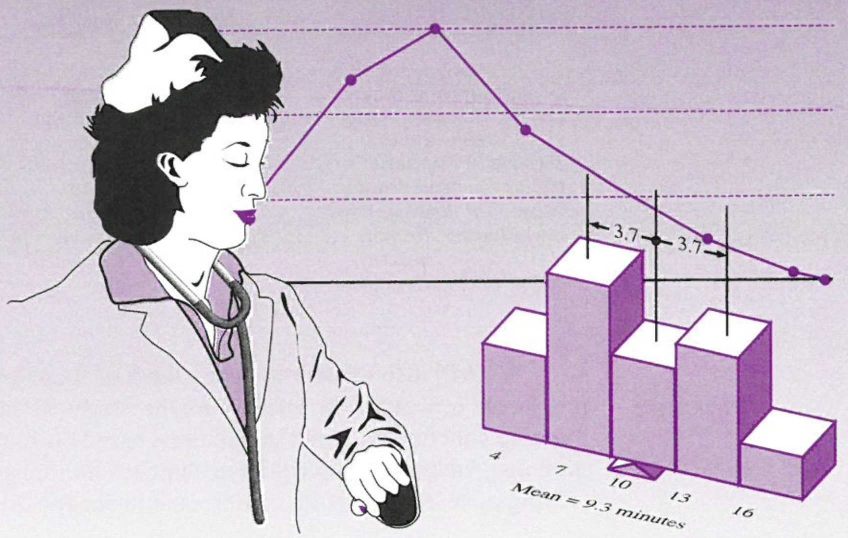
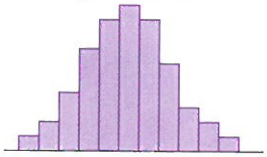
CHAPTER 2

Organizing and Analyzing Data



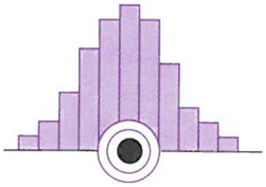
This chapter is concerned with the techniques used to summarize data. Specifically, we will study

Graphical representations



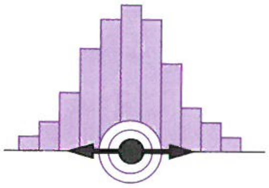
Graphical representations provide the reader with a visual sense of the data, allowing an immediate registration of the data's impact. The drawing above is a histogram. We will also study the frequency polygon and circle graph.

Measures of central tendency



Measures of central tendency provide some sense of the middle or central value, which best typifies the data. We shall study the arithmetic mean, the median, and the mode.

Measures of dispersion or spread



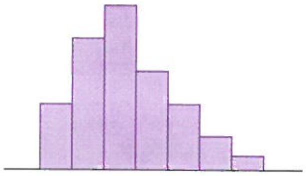
Measures of dispersion or spread provide a value that measures how widely scattered the data is. Specifically, we will present the range and standard deviation.

We will also explore the usefulness of these graphical representations and measures, especially in relation to the study of inferential statistics, that is, drawing conclusions about populations based on samples taken from those populations. Additional descriptive techniques are presented in section 2.7 and the writing of research reports is presented in section 2.8. ▼

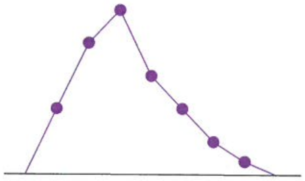
2.1 Graphical Representations

Whether it be population or sample data, the importance of representing data in graphical or picture form cannot be overemphasized. These graphs or picture representations provide the immediate impact of the data in an easy-to-interpret format. Specifically, in this section, we introduce the histogram, frequency polygon, and circle graph shown as follows.

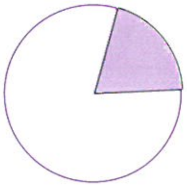
Histogram



Frequency polygon



Circle graph



These graphical representations are used often throughout the text, so understanding them is essential.

Histogram

One of the most frequently used graphical representations is the histogram. Let's look at the following example.

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 100 current users yielded the following ages:

Ages of 100 randomly selected users of Countrygirl Makeup

21.4	18.1	17.6	26.6	47.1
24.5	22.2	22.9	32.7	15.8
31.1	23.1	27.7	29.0	21.6
17.7	32.6	21.6	26.4	31.4
22.1	22.3	31.9	25.7	35.2
21.9	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
40.7	40.7	16.6	18.1	42.7
31.1	16.0	38.9	26.7	36.6
21.7	26.7	36.0	37.3	27.1
23.1	28.2	20.6	25.7	26.7
26.9	35.8	23.7	38.2	20.9

When data is presented in this form, we call it **ungrouped.** Each value is recorded exactly as measured. Unfortunately, ungrouped data is difficult to represent graphically. Because graphical representations are often essential for a complete understanding, we may choose to tally such data into groups or categories, in this case, age categories. We would then use the results of these groupings to construct a histogram.

To construct a histogram from the above set of 100 ages, let us proceed as follows.

Age Category
15–19
20–24
25–29
30–34
35–39
40–44
45–49

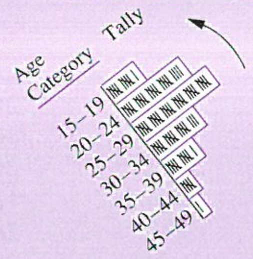
First, we must select suitable age categories or groups. More is discussed on this, but for now, say we select the above age categories.

Age Category	Tally
15–19	
20–24	|
25–29	|
30–34	
35–39	
40–44	
45–49	

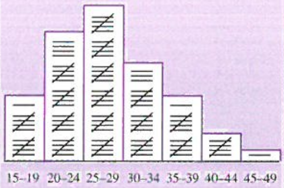
Second, we tally the data. For example, the first age on our list is 21.4 years. We represent this by a slash in the 20–24 age category. The next age, 24.5, belongs in the 25–29 category. Note: Had this age been 24.4, it would have been put in the 20–24 age category.

Age Category	Tally	Frequency or Category Total
15–19	𝍸𝍸|	11
20–24	𝍸𝍸𝍸𝍸||||	24
25–29	𝍸𝍸𝍸𝍸𝍸𝍸	30
30–34	𝍸𝍸𝍸|||	18
35–39	𝍸𝍸|	11
40–44	𝍸	5
45–49	|	1

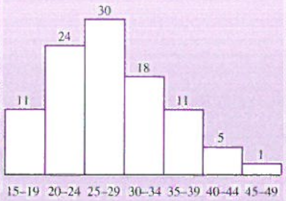
Third, we tally all 100 ages. When finished, the tally would appear as shown above. Also shown is the final count for each category called the frequency or category total.



Fourth, we encase the slashes in rectangular bars, and rotate the tally ¼ turn counterclockwise.



The final position appears as above. For aesthetic reasons, the bars may be widened.

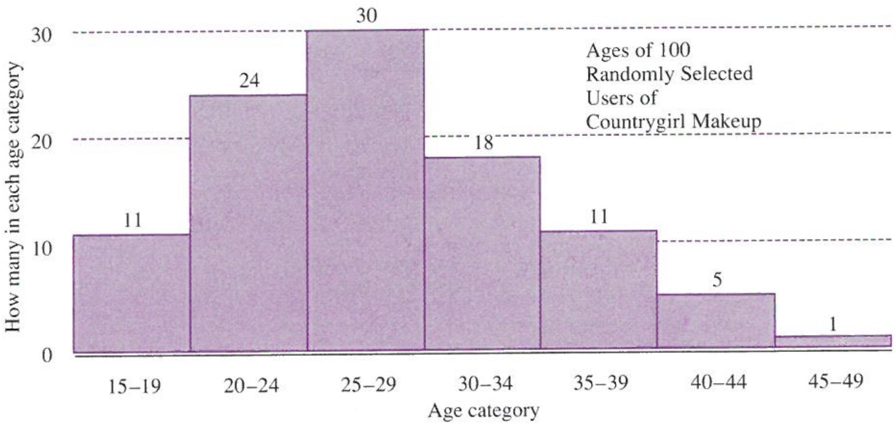


Last, we remove the tally slashes and, often, for ease of viewing, lightly shade or color the bars and put category totals atop each bar.

To complete the histogram, we add the following components.

1. An overall identifying title, explaining what the histogram represents.
2. A horizontal scale, identifying the attribute we are measuring.
3. A vertical scale, representing *how many in each category.*

The final presentation of the histogram might then be as follows:



Notice that the histogram bar of the 25–29 age category is almost three times as high as the histogram bar for the 15–19 age category. This means that there were almost three times as many people in the 25–29 age category as were in the 15–19 age category.

Also did you notice that 48 out of a total of 100 in our sample (30 plus 18) were between the ages of 25 and 34? This represents almost half or 48% of our sample. Also did you notice very few users were ages 15–19?

Histograms are quite valuable in helping us unlock the secrets of a population. If a sample is large enough, the shape of the sample histogram will give a good approximation of the shape of the population histogram. In other words, if we were to measure *all* current users of our product, we might find the histogram above to be quite similar in shape to the histogram of our entire population of users.

With this in mind, do you feel Countrygirl executives might reevaluate their strategy of marketing their products to young women ages 16–24? Actually, this is a complex question and executives at companies like Revlon and L'Oreal, along with their advertising agency counterparts, grapple with questions like this on a routine basis.

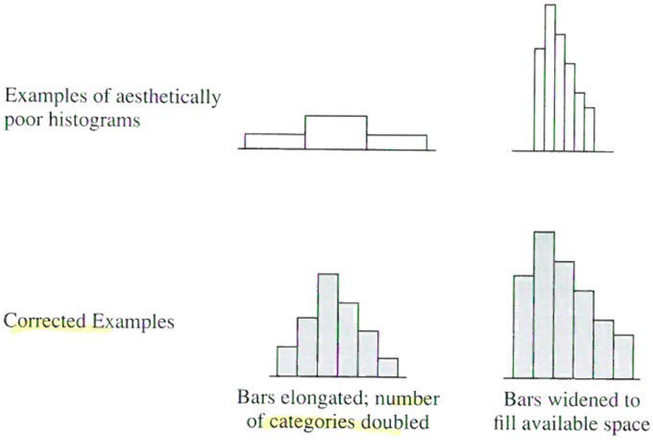
Basic guidelines in the construction of histograms are as follows:

1. Each category should be the same numerical width. In other words, in our histogram of 100 users of Countrygirl, the first age category, 15–19, contains five ages, 15, 16, 17, 18, and 19, therefore the next category, 20–24, should also contain five ages, which it does. And each subsequent category should contain five ages in sequence, which they do.

2. There should be no overlap. For instance, in our Countrygirl histogram, someone 24.4 years old fits into the 20–24 age category, whereas someone 24.5 years old fits into the 25–29 age category. Each reading fits into *one* and only one category.

Aesthetic guidelines in the construction of histograms would include

3. The histogram should be made to fill as much of the available space as possible by widening or proportionally elongating the histogram bars. Good examples would be the histograms presented in this chapter.
4. Histograms with 5 to 12 categories tend to be the most visually appealing.

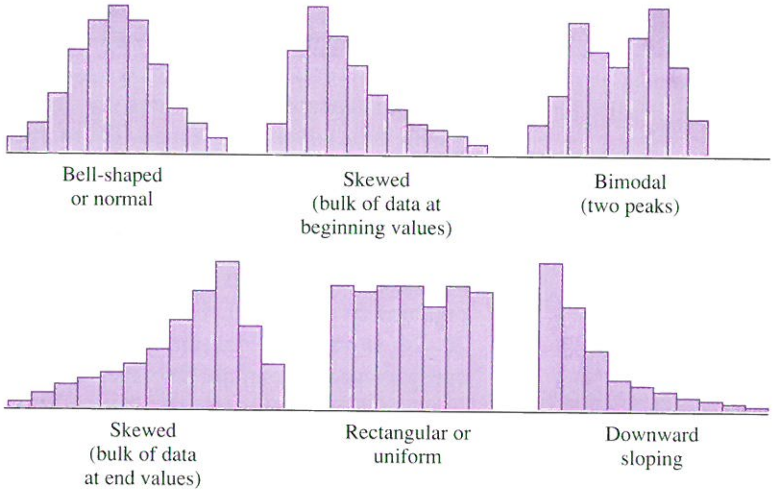


see p160

Population Histograms

Experience has shown that many *population* histograms take on repeating shapes. In other words, there are certain common shapes that seem to reoccur.

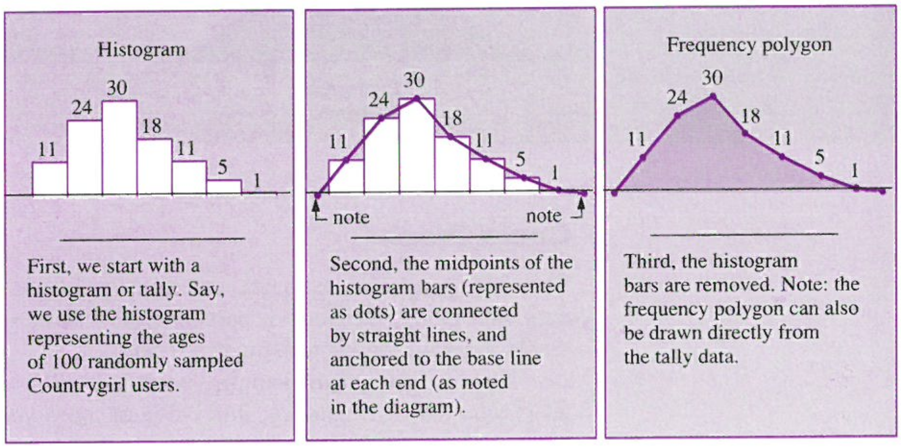
Common Shapes of Population Histograms



Can you think of populations that might take on the shapes represented in the above figures? For instance, if we measured the heights of all the men or all the women in your town or city, chances are the histograms would be bell-shaped, whereas, if we measured the salaries of all employees in a company, the histogram might be skewed with the bulk of data at beginning values. Can you think of other populations that might fit these shapes?

Frequency Polygon

A frequency polygon is merely a line representation of a histogram. Let's see how it works.

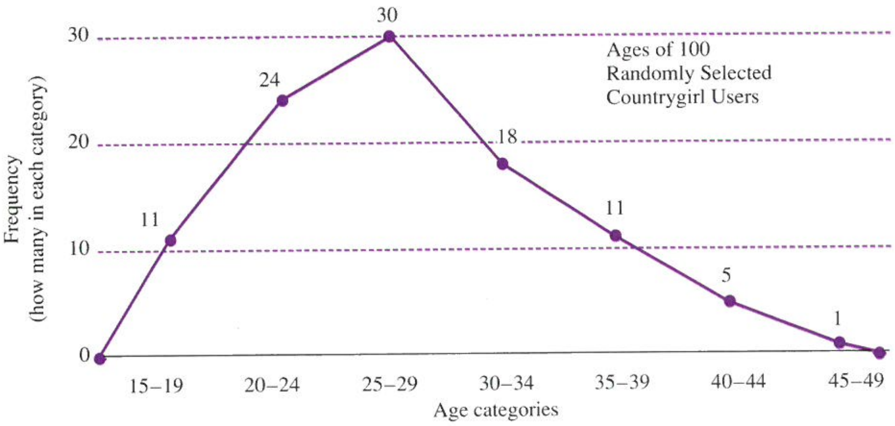


First, we start with a histogram or tally. Say, we use the histogram representing the ages of 100 randomly sampled Countrygirl users.

Second, the midpoints of the histogram bars (represented as dots) are connected by straight lines, and anchored to the base line at each end (as noted in the diagram).

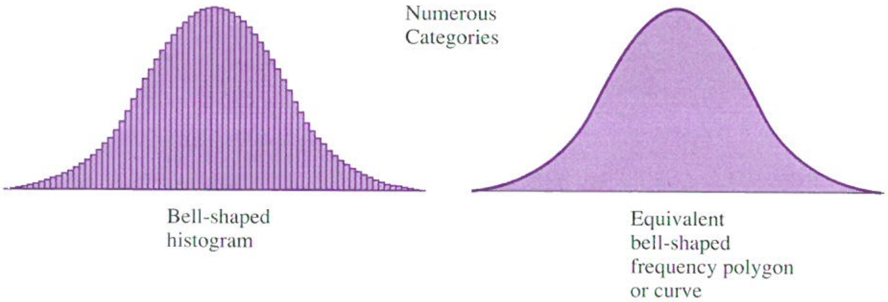
Third, the histogram bars are removed. Note: the frequency polygon can also be drawn directly from the tally data.

The final frequency polygon, titled, might appear as follows.



Notice that the word **frequency** was used to title the vertical scale. This is a commonly used word in statistics, meaning *how many in each category*.

Frequency polygons are clear, concise, and easy to sketch and are especially useful when the number of categories in a histogram grows too large to be easily represented. When the number of categories in a histogram grows into the hundreds or more, often the frequency polygon takes on the appearance of a smooth sloping line, which is often referred to as a **curve.**



Circle Graph

The circle graph (or pie chart) is a common pictorial representation for proportion data, that is, the fraction (or percentage) of some sample or population that possesses a certain characteristic or attribute.

In our Countrygirl sample, we might wish to represent the proportion of users who are, for instance, thirty-five or more years old. We would first count how many were thirty-five or more years old, change that number to a percentage, then represent that percentage as a slice in a circle graph, as follows:

Age category	Tally	Frequency or category total
15–19	卌 卌 |	11
20–24	卌 卌 卌 卌 ||||	24
25–29	卌 卌 卌 卌 卌 卌	30
30–34	卌 卌 卌 |||	18
35–39	卌 卌 |	11
40–44	卌	5
45–49	|	1

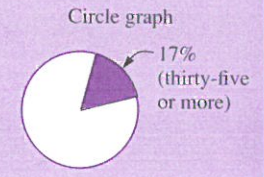
First, we start with our raw data, or in this case, we can use our tally, and count the number of users who are thirty-five or more years old. The answer is 11 + 5 + 1 = 17

$$\left[\begin{array}{l}\text{Percentage of}\\\text{sample who}\\\text{are thirty-five}\\\text{or more}\end{array}\right] = \left[\begin{array}{l}\text{Fraction}\\\text{who are}\\\text{thirty-five}\\\text{or more}\end{array}\right] \times 100$$

$$P = \frac{17}{100} \times 100$$

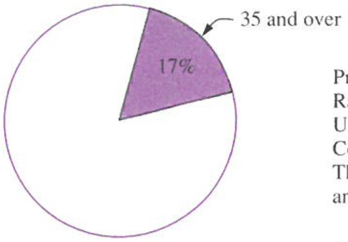
$$P = 17\%$$

Second, since 17 users were thirty-five or over, we use our formula, given above, to calculate what percentage this is of our total sample. Answer: 17%.



Third, since a circle has 360°, 17% of 360° = 61.2°, so we shade a slice of the circle representing 61.2°. A protractor can be used to draw the angle.

The final circle graph would be titled and might appear as follows:

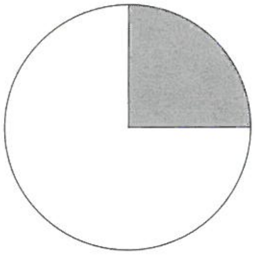


Proportion of 100
Randomly Selected
Users of
Countrygirl Makeup
Thirty-Five Years Old
and Over

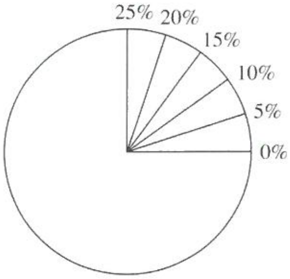
If we represent the sample proportion with the symbol p_s we can state,

$$p_s = 17\% \text{ (or in decimal form, } p_s = .17)$$

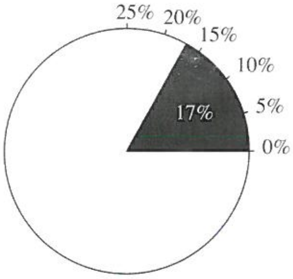
Instead of calculating 17% of 360° = 61.2° and then measuring 61.2° with a protractor, the slice in the circle graph can be estimated as follows:



25% is ¼
of a circle.

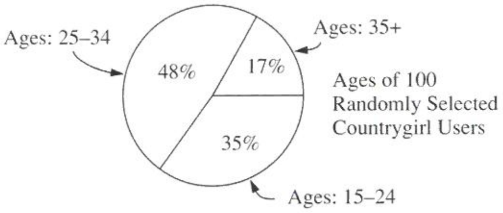


Starting from
0%, draw equal
5% intervals on
the circle until
you reach 25%.



Estimate where
17% lies and shade.

The circle graph is quite versatile. We may wish to represent multiple categories, such as the percentage of users 35 or more, the percentage of users 25 to 34, and the percentage of users 15 to 24. This would be represented as follows:



Ages of 100
Randomly Selected
Countrygirl Users

Ages	**Totals**	**Change to Percentage**	**Equivalent Degrees***
15–24	35	$\frac{35}{100} \times 100 = 35\%$	126.0°
25–34	48	$\frac{48}{100} \times 100 = 48\%$	172.8°
35+	17	$\frac{17}{100} \times 100 = 17\%$	61.2°

The following example is presented for practice.

*To change to degrees, take 35% of 360°: .35 × 360 = 126.0°.

Example

Out of 50 randomly selected Countrygirl users, 12 possessed the attribute of red hair. Draw a circle graph representing this sample proportion.

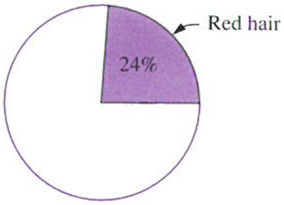
Solution

We first convert the quantity 12 out of 50 into a percentage.

$$\text{Percentage} = \text{Fraction} \times 100$$
$$P = \frac{12}{50} \times 100$$
$$P = 24\%$$

Now we know 24% of the sample had red hair, 24% of 360° = 86.4°. We can measure this angle with a protractor or we can estimate 24% as *almost* 25% or almost $\frac{1}{4}$ of a circle.

Answer



The proportion of 50 randomly selected users of Countrygirl Makeup who possess the attribute of red hair is 24%, expressed as

$$p_s = 24\%$$

or in decimal form, as

$$p_s = .24$$

■

2.2 Measures of Central Tendency (Ungrouped Data)

In addition to graphical representations, we often wish to get some sense of the *middle* or *central value* of our data. This is called **central tendency** and the most used measures of central tendency are the arithmetic mean, the median, and the mode.

Arithmetic Mean

Perhaps the most widely used measure of central tendency is the arithmetic mean, what most people refer to as the *average*.

> The **arithmetic mean** is the result obtained when a collection of values are added, then divided by *n*, the number of values. The formula for the arithmetic mean of a *sample* is
>
> $$\bar{x} = \frac{\Sigma x \longleftarrow \text{Sum of the values (Note: the symbol } \Sigma \text{ means ``sum of.'')}}{n \longleftarrow \text{Number of values}}$$

Notice that the symbol $\bar{x}$ (x bar) was used to represent the arithmetic mean of a sample. If we were to calculate the arithmetic mean of a population, we would use the symbol μ.

Because we shall employ the arithmetic mean throughout the remainder of the text, we shall simply refer to it as the *mean* or *average*. Technically, the words mean and average each refer to a broad number of measures, however we shall use these words only to represent the arithmetic mean.

Example

Suppose 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 mean.

Solution

Since the sample average, $\bar{x}$, is equal to the sum of the values divided by n, we proceed as follows.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{10 + 6 + 5 + 14 + 6 + 13}{6} = 9 \text{ minutes}$$

The results would be summarized by stating,

$$n = 6 \text{ readings}$$
$$\bar{x} = 9 \text{ minutes}$$

This simply means that for six observations, the average was calculated to be 9 minutes.

Now, the question arises, can we use the sample average to evaluate the nurse-in-training? In other words, because our sample produced an average of 9 minutes, can we assume that if we had measured all the hundreds of times this nurse drew blood specimens, the average of all these hundreds of readings would also be approximately 9 minutes?

In order to use sample values as representative of population values, we must be careful to ensure that a valid random sample was taken. So, the answer to the question would be, yes, provided the sample times were randomly selected over several months, accurately recorded, and the researcher's presence and methods in no way interfered with the normal functioning of the nurse's procedures. In other words, provided we were successful in achieving a valid random sample (actually, for small samples, $n < 30$, other considerations apply, which are addressed in section 2.4; note the expression $n < 30$ means n *less than* 30). ■

The arithmetic mean or average is often the preferred measure of central tendency for a number of reasons.

Advantages of Using the Arithmetic Mean

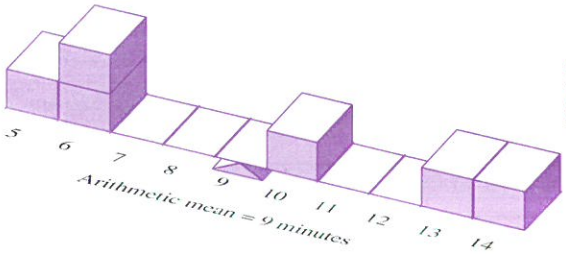
1. It is unique and always exists. In other words, in any set of data, there is one and only *one* arithmetic mean and that value always exists.

2. It takes into consideration all the data. No value is left out and every value influences the calculation of the arithmetic mean in some way.

3. It readily lends itself to inferential analysis, that is, in using sample data to estimate population characteristics. For example, with bell-shaped populations, one of the most frequent population shapes we encounter, the sample means cluster much closer to the population mean than, say, the sample medians cluster to the population median. This is an important advantage when we begin to use samples to estimate population characteristics.

4. It is algebraically tractable. In other words, the mean is easily manipulated in complex equations, more so than other measures. This is a vital consideration in advanced work and probably in large part what has accounted for the mean's widespread appeal throughout the centuries.

5. There are even aesthetic advantages to using the mean. It strongly appeals to our sense of balance. In other words, if we represent each value as a block on a number line, the blocks would perfectly balance at one point, and that point always turns out to be the arithmetic mean.



Six Observed Times from Nurse-in-Training Study

Other measures of central tendency do not offer such a powerful array of advantages. Mostly for these reasons, has the arithmetic mean evolved into the preferred measure of central tendency.

However, we must consider other measures, because in certain instances, these other measures offer us information about our data that substantially adds to our understanding. Two such other measures are the median and the mode.

Median

Another commonly used measure of central tendency is the **median.**

Median
The middle value when data is arranged from lowest to highest value.

Example

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

Solution

First, line up the data according to size.

5, 6, 6, 10, 13, 14

Because there is no one middle value, we take the average of the two middle values.

$$\text{Median} = \frac{6 + 10}{2} = 8 \text{ minutes}$$

As a general rule, when the number of values is even, you will have to average the two middle values. When the number of values is odd, you will have *one* middle value and that will be used as the median. ■

The median is especially useful as a measure of central tendency when data is highly skewed, such as in the following example.

Example

The Technic Company has five employees, which includes the president, yielding the following annual salaries: $30,000; $30,000; $30,000; $30,000; and $200,000. Calculate the median.

Solution

Arrange the salaries in order, then select the middle value.

$30,000 $30,000 $30,000 $30,000 $200,000
↑

Median = $30,000

Had we calculated the mean for the above salaries, the mean would have been $64,000 [(30,000 + 30,000 + 30,000 + 30,000 + 200,000)/5 = $64,000]. Picture yourself answering an employment ad: Join Technic Company, average salary of employees, $64,000. Do you think the arithmetic mean in this case gives a fair representation of central tendency? Many would consider the median, $30,000, to be a more realistic value here in defining central tendency. ■

The advantage of the median is that it is unaffected by extreme values and, in certain instances, gives a more realistic measure of central tendency. This is especially true for highly skewed data, such as in the case above. Unfortunately, the median does not lend itself quite as well as the mean to inferential analysis, that is, in using sample data to estimate population characteristics.

Mode

A third measure of central tendency is the **mode.**

Mode
The most occurring value.

Example

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

Solution

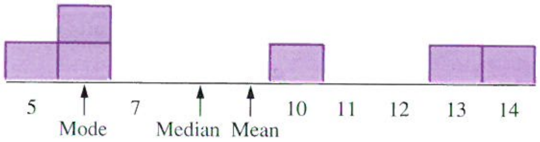
Mode = 6 minutes since this is the most frequently occurring value.

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.



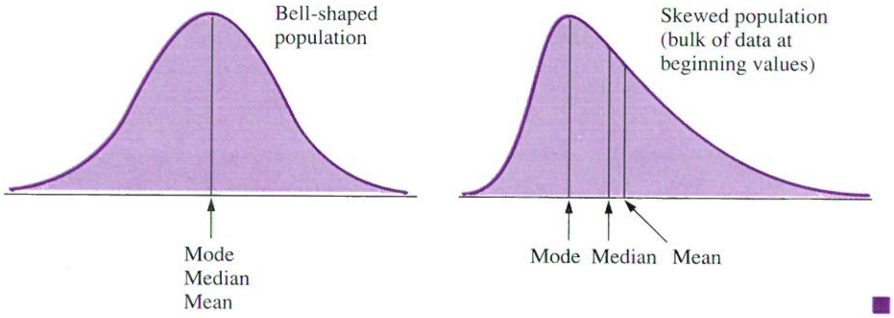
Which do you feel gives a better measure of central tendency for this data? Although the mean is the preferred measure, each adds a little more information.

Example

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

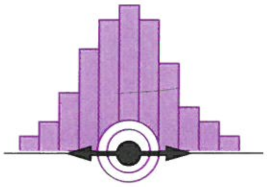
a. bell-shaped population.
b. skewed population, with the bulk of data at beginning values.

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 = $14 - 5 = 9$

$$\text{Range} = 9 \text{ minutes}$$

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

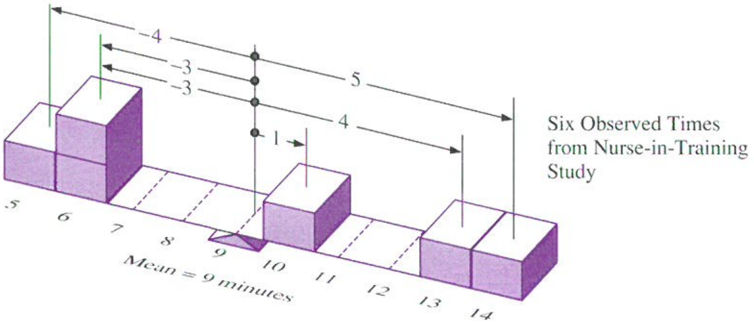
Although easy to compute, the range offers no information about the distribution of data between these two extremes of high and low value and, thus, the range is mostly used as a rough gauge in determining dispersion or spread.

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:

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

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.

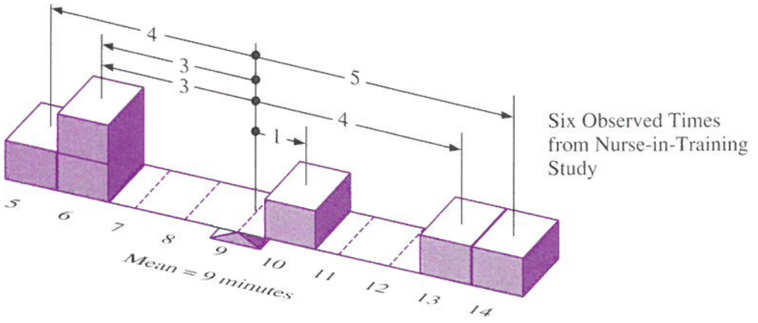
$$\begin{array}{l}\text{Average Distance} \\ \text{from the Mean}\end{array} = \frac{-4 - 3 - 3 + 1 + 4 + 5}{6} = \frac{0}{6} = 0 \text{ minutes}$$

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	$\overline{x}$	$x - \overline{x}$	$\lvert x - \overline{x} \rvert$
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.

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

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

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

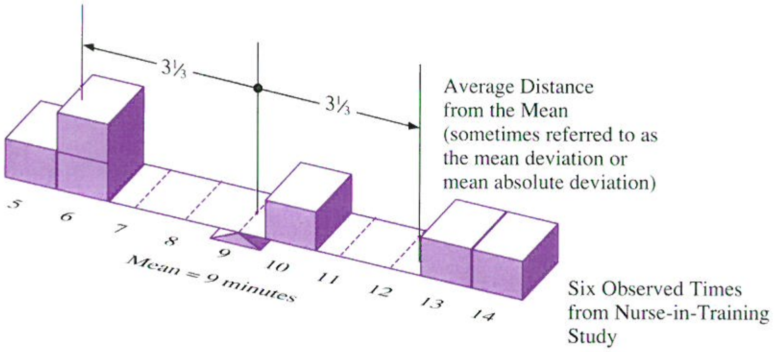
We now have a measure of how scattered or spread out the data is. We can now say, the average distance a value is from the mean is $3\frac{1}{3}$ minutes.

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, $|x - \bar{x}|$'s, to remove the negative signs in complex calculations necessitates breaking up each problem into three cases: (a) when $x < \bar{x}$, (b) when $x = \bar{x}$ and (c) when $x > \bar{x}$. 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.

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

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

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

$= 15.2$ 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

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

$$s = \sqrt{15.2}$$

$$= 3.899$$

$= 3.9$ 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*, σ, we would simply divide by N in the formula:

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

‡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.

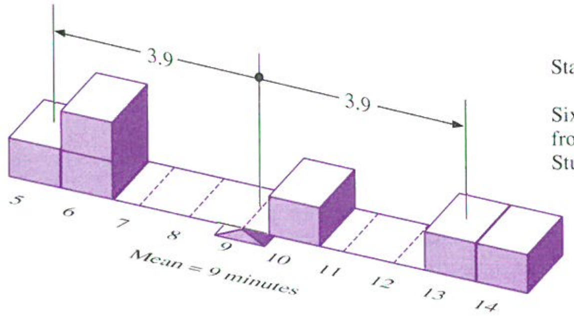
x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
5	9	−4	16
6	9	−3	9
6	9	−3	9
10	9	1	1
13	9	4	16
14	9	5	25
		$\Sigma(x - \bar{x})^2 =$	76

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

$$= \sqrt{15.2} = 3.899$$

$$= 3.9 \text{ minutes}$$

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.

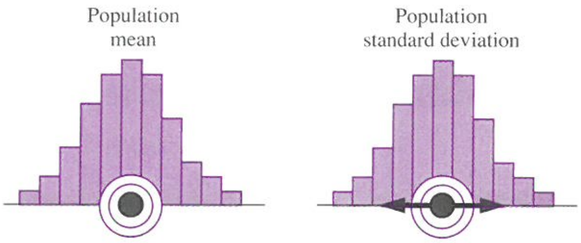
Advantages of Using the Standard Deviation as our Measure of Spread

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.

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

skip

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 I
6–8	𝍸 𝍸 I
9–11	𝍸 II
12–14	𝍸 IIII
15–17	III

Calculate the mean.