4.1 Measures of Location

To your question concerning the whereabouts of Tom Stevens, a mutual friend responds, "In the library." The answer is informative in that it provides general information about Tom's whereabouts. Even though you don't know what part of the library he is in, the information does provide a focal point—a location of sorts. The information may also permit some inferences regarding what Tom is doing.

Statistically speaking, the idea of location is similar to knowing that Tom is in the library. If we think of a data set as a group of data values that cluster around some central value, then this central value provides a focal point for the data set—a location of sorts. Unfortunately, the notion of *central value* is a vague concept, which is as much defined by the way it is measured as by the notion itself. There are several statistical measures that can be used to define the notion of center: the arithmetic mean, weighted mean, trimmed mean, median, and mode.

The Arithmetic Mean

The **arithmetic mean** is one of the more commonly used statistical measures. It appears every day in newspapers, business publications, and frequently in conversation. For example, after receiving a grade on a test, you might be curious about how the rest of the class performed. You might ask the instructor, *What was the average on the test?* The term *average* is often associated with the arithmetic mean, which is more commonly referred to as just the **mean**.

Arithmetic Mean

Suppose there are n observations in a data set, consisting of the observations $x_1, x_2, \ldots, x_n$; then the **arithmetic mean** is defined to be

$$\frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right).$$

DEFINITION

If we use some common mathematical notation, the formula can be simplified to

$$\frac{\sum x_i}{n}$$

where x_i is the i^{th} data value in the data set and $\sum$ (pronounced sigma) is a mathematical notation for adding values. There are two symbols that are associated with the expression given above:

$\mu = \frac{1}{N}\left(x_1 + x_2 + \ldots + x_N\right)$ the **population mean**, and

$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right)$ the **sample mean**.

Here N refers to the size of the population and n refers to the size of the sample. Otherwise, the calculations are made in precisely the same way. The Greek letter μ, representing the population mean, is pronounced *mu*, and the symbol $\bar{x}$, representing the sample mean, is pronounced *x-bar*.

Technology

To find summary statistics, such as the sample mean, we can use the `1-Var Stats` option on the TI-83/84 calculator. The results are shown here. For instructions please refer to the web resource at **Tech > Descriptive Statistics - One Variable**.

```
1-Var Stats
 x̄=9
 Σx=36
 Σx²=390
 Sx=4.69041576
 σx=4.062019202
↓n=4
```

```
1-Var Stats
↑n=4
 minX=4
 Q1=5.5
 Med=8.5
 Q3=12.5
 maxX=15
```

Example 4.1.1

Calculate the sample mean of the following sample data values: 4, 10, 7, 15.

Solution

$$x_1 = 4,\ x_2 = 10,\ x_3 = 7,\ x_4 = 15, \text{ and } n = 4.$$

Note that

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4+10+7+15}{4} = \frac{36}{4} = 9.$$

The sample mean is 9. But why does adding up a group of numbers and dividing by the number of numbers measure central tendency? As unlikely as it sounds, the answer is related to balancing a scale.

Deviation

Given some point A and a data point x, then $x - A$ represents how far x **deviates** from A. This difference is also called a **deviation**.

DEFINITION

Let's calculate the deviations from the mean (9) for the data in Example 4.1.1. Examining the deviations from the mean in Table 4.1.1, we can see the deviations on the left side (−5 and −2) and right side (1 and 6) are in balance. In fact, the mean is considered a point of centrality because the deviations from the mean on the positive side and the negative side are equal (see Figure 4.1.1). The sample mean can be interpreted as a center of gravity.

Table 4.1.1 – Deviations from the Mean

Data (x_i)	Deviations from the Mean ($x_i - 9$)
4	−5
10	1
7	−2
15	6
	TOTAL $\sum(x_i - 9) = 0$

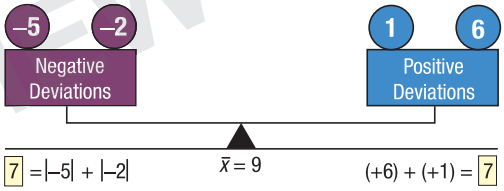


Figure 4.1.1

On the other hand, if we calculate the deviations about any other value the deviations do not balance. For example, assume the central value is 8. The deviation from the alleged central value, 8, for each data value is calculated in Table 4.1.2 and shown in Figure 4.1.2.

The positive deviations (+2 and +7) are not counterbalanced by the negative deviations (−4 and −1). A desirable characteristic of a central value would be to have the positive and negative deviations equal to each other in absolute value.

Table 4.1.2 – Deviations from Some Other Value	
Data (x_i)	**Deviations from 8 ($x_i - 8$)**
4	−4
10	2
7	−1
15	7
TOTAL	$\sum(x_i - 8) = 4$

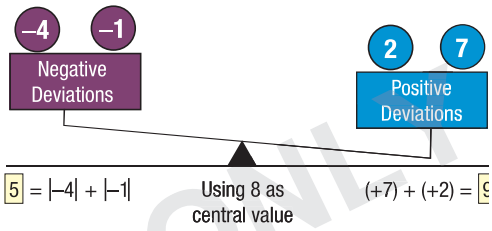


Figure 4.1.2

Although the arithmetic mean is frequently used, there are times when it should not be employed. *Since the mean requires that the data values be added, it should only be used for quantitative data.* Furthermore, if one of the data values is extremely large or small relative to the others, this data value could be considered an **outlier**. An outlier can have a dramatic impact on the value of the mean and dramatically affect its value as a measure of central tendency. (See Section 4.3 for further discussion of outliers.)

Outliers and Resistant Measures

An **outlier** is a data value that is extremely different from other measurements in the data set. Statistical measures which are not affected by outliers are said to be **resistant**.

DEFINITION

The mean is *not* a **resistant measure**.

Weighted Mean

The weighted mean is similar to the arithmetic mean except it allows you to give different weights (or importance) to each data value. The weighted mean gives you the flexibility to assign weights when you find it inappropriate to treat each observation the same. The weights are usually positive numbers that sum to one, with the largest weight being applied to the observation with the greatest importance. The weights can be determined in a variety of ways, such as the number of employees, the market value of a company, or some other objective or subjective method. There are occasions in which it is easier to assign the weights without worrying that they will sum to one. If you are concerned about your weights summing to one, you can make your weights sum to one by dividing each weight by the sum of all the weights.

Weighted Mean

The weighted mean of a data set with values $x_1, x_2, x_3, \ldots, x_n$ is given by

$$\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum(w_i \cdot x_i)}{\sum w_i}$$

where w_i is the weight of observation x_i.

FORMULA

Example 4.1.2

Meghan is a freshman in college and she received the following grades for her first semester.

Meghan's Grades		
Course	**Grade**	**Credit Hours**
Psychology 101	B	3
Probability and Statistics	A	4
Anatomy I	C	5
English 101	A	3

A grade of A is worth 4 points on a 4-point scale. A grade of B is worth 3 points and a grade of C is worth 2 points.

a. Calculate Meghan's GPA using the credit hours as weights. Round to two decimal places.

b. If Meghan's goal was to have a GPA of 3.4, what grade did she need to make in the Anatomy I class to reach her goal?

Solution

a. To calculate Meghan's GPA we use the weighted mean formula with the numerical grade point values as the x-values and the credit hours as weights.

$$\overline{x} = \frac{\sum(w_i \cdot x_i)}{\sum w_i} = \frac{3\cdot 3 + 4\cdot 4 + 2\cdot 5 + 4\cdot 3}{3+4+5+3}$$

$$= \frac{9+16+10+12}{15} = \frac{47}{15}$$

$$\approx 3.13$$

b. To determine the grade that Meghan needed in the Anatomy I class to reach her goal of a 3.4 GPA, let x represent the grade in the weighted mean formula.

$$\overline{x} = 3.4 = \frac{3\cdot 3 + 4\cdot 4 + x\cdot 5 + 4\cdot 3}{15}$$

$$= \frac{9+16+5x+12}{15} = \frac{37+5x}{15}$$

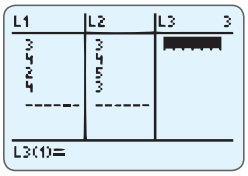
Solving this equation for x gives us the following result.

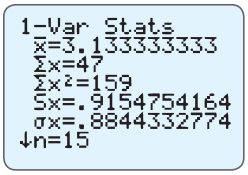
$$\begin{aligned} 3.4(15) &= 37 + 5x \\ 51 &= 37 + 5x \\ 51 - 37 &= 5x \\ 14 &= 5x \\ 2.8 &= x \end{aligned}$$

Since the grading scale only uses integers, Meghan would have had to make 3 points or a grade of B on her Anatomy I class for a GPA of 3.4.

Technology

To find the weighted mean we use the `1-Var Stats` option on the TI-83/84 calculator with the numerical grade point value in L1 and credit hours in L2. The results are shown below. For instructions please refer to the web resource at **Tech > Weighted Mean**.





The Trimmed Mean

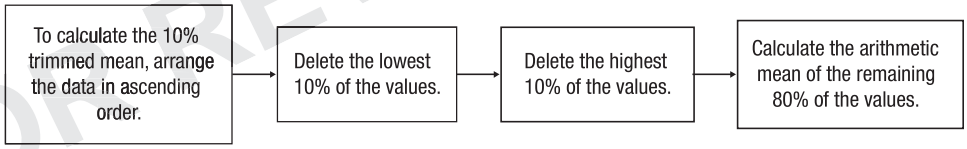
Since outliers can have an enormous effect on the value of the mean, the mean's usefulness as a typical measure of a set of data is diminished if the data contain outliers.

Trimmed Mean

The **trimmed mean** is a modification of the arithmetic mean which ignores an equal percentage of the highest and lowest data values in calculating the mean.

DEFINITION

Finding the 10% Trimmed Mean



Before calculating the trimmed mean, the data are arranged in ascending order of magnitude. A 10% trimmed mean uses the middle 80% of the data values. It is calculated by removing the top 10% *and* the bottom 10% of the data values, then finding the arithmetic average of the remaining values. If the data set does not contain any outliers, the mean and the trimmed mean will be similar. Unlike the arithmetic mean, the trimmed mean is not affected by outliers and is considered a resistant measure.

Example 4.1.3

Consider the following data taken from a poll on how many text messages a person sent in a day.

16 18 20 21 23 23 24 32 36 42

mean = 25.5

Find the 10% trimmed mean.

Solution

Since there are 10 observations, removing the highest 10% and lowest 10% means removing only one observation from each end of the data. That is,

$$10\% \text{ of } 10 = 0.1 \cdot 10 = 1.$$

Note that the data are already sorted. If the mean is calculated without including the values 16 and 42, the resultant measure is called the 10% trimmed mean.

~~16~~ 18 20 21 23 23 24 32 36 ~~42~~

$$10\% \text{ trimmed mean} = \frac{18+20+21+23+23+24+32+36}{8} = 24.625$$

Rounding Rule

A general rule of thumb regarding rounding for the statistics calculated in this chapter is to round to one more decimal place than the largest number of decimal places given in the data. Occasional exceptions to this rule can be made when the type of data lends itself to a more natural rounding scheme, such as rounding values of currency to two decimal places.

Measuring Figure Skating Performances

Almost every figure skating competition has some scoring controversy. The Winter Olympics of 2002 were no exception. French judge, Marie-Reine Le Gougne, said she was "pressured to vote a certain way'' when she scored the Russian couple, Elena Berezhnaya and Anton Sikharulidze, over the Canadian pair, Jamie Sale and David Pelletier. In addition, very few people understood exactly how Sarah Hughes won the gold medal and how Michelle Kwan dropped to third after leading the event.

For almost a century figure skating has used a scoring method that is similar to the methodology of the trimmed mean in order to remove bias. Skaters were scored on a 0 to 6 scale. The highest and the lowest score are discarded (the data is trimmed) and the resulting "score" is computed. The intent of trimming the data is to avoid bias caused by judges with nationalistic or political agendas. The International Skating Union is replacing this scoring method with similar methods that attempt to remove biases in judge's scores.

If there had been 100 data values, the largest 10% and smallest 10% (a total of 20 data values) would have been removed before the mean was calculated.

Example 4.1.4

Consider the same data set, except the last data value is replaced with an outlier.

16 18 20 21 23 23 24 32 36 490

mean = 70.3

Find the 10% trimmed mean.

Solution

Since there are 10 observations, removing the highest 10% and lowest 10% means removing only one observation from each end of the data.

~~16~~ 18 20 21 23 23 24 32 36 ~~490~~

$$10\%\text{ trimmed mean} = \frac{18+20+21+23+23+24+32+36}{8} = 24.625$$

As expected, the trimmed mean is not affected by the addition of the outlier, while the mean increased dramatically. This is why the trimmed mean is considered to be a **resistant measure**.

The Median

The median of a set of data provides another measure of center that is different from a "mean". It is a simple idea. To find the median, place the data in ascending order and then find the observation that has an equal number of data values on either side. That is, half of the observations are less than the median and half of the observations are greater than the median. The median is the middle value.

Median

The **median** of a set of observations is the measure of center that is the middle value of the data when it is arranged in ascending order. The same number of data values lie on either side of the median.

DEFINITION

To determine the median of a set of data, we use the following steps.

Finding the Median of a Data Set

1. Arrange the data in ascending order.
2. Determine the number of values in the data.
3. Find the data value in the middle of the data set.
4. If the number of data values is odd, then the median is the data value that is exactly in the middle of the data set.
5. If the number of data values is even, then the median is the mean of the two middle observations in the data set.

PROCEDURE

Example 4.1.5

Consider the following goal tallies from the last eleven games played by the Charleston Battery soccer team.

2, 3, 5, 4, 1, 7, 3, 3, 1, 2, 6

Find the median.

Solution

First, the data set must be ordered,

1, 1, 2, 2, 3, 3, 3, 4, 5, 6, 7

Since the data set contains an odd number of values, 11, the middle observation in the ordered array must be the sixth observation. Since the median is the sixth observation, the median value is 3.

1, 1, 2, 2, 3, **3**, 3, 4, 5, 6, 7

Example 4.1.6

Consider the following ten test scores from a student taking a high school calculus class.

65, 98, 76, 83, 94, 79, 88, 72, 90, 85

Find the median.

Solution

If there are an even number of observations, average the two center values in the ordered data set.

65, 72, 76, 79, **83**, **85**, 88, 90, 94, 98

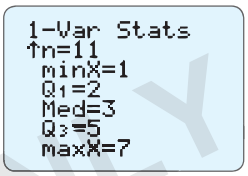
To find the median, average the fifth and sixth observations.

$$\frac{83+85}{2} = 84 \text{ (the median)}$$

Technology

To find summary statistics, such as the median, we can use the `1-Var Stats` option on the TI-83/84 calculator or use Excel. The results are shown here. For instructions please refer to the web resource at **Tech > Descriptive Statistics - One Variable**.

Column1	
Mean	3.363636364
Standard Error	0.591957113
Median	3
Mode	3
Standard Deviation	1.963299634
Sample Variance	3.854545455
Kurtosis	-0.466246885
Skewness	0.636683254
Range	6
Minimum	1
Maximum	7
Sum	37
Count	11



The median possesses a rather obvious notion of centrality, since it is defined as the central value in an ordered list. It is not affected by outliers and is therefore a resistant measure. For example, if we replaced 98 with 200,000,000 in the data set from the previous example, the median would not change at all. The median does possess one limitation: it cannot be applied to nominal data. In order to calculate the median, the data must be placed in order. To accomplish this task meaningfully, the level of measurement must be at least ordinal. Unless the data set is skewed or contains outliers, the median and the mean usually have similar values.

The Mode

The **mode** is another measure of location. It is not used as frequently as the mean or the median, and its relation to the "central tendency" concept and the values of the mean and median are not predictable. The mode is the only measure of location that can be used for nominal data. Of the three measures of location, the mode is used the least due to the limited information it provides. Sometimes sorting the data (in ascending or descending order) makes it easier to find the mode.

Mode

The **mode** of a data set is the most frequently occurring value.

DEFINITION

When reporting the mode for numerical data, do not round. The value of the mode should be the same as the original data value.

Example 4.1.7

Find the mode of the following data regarding the number of power outages reported over a period of eleven days.

0, 1, 4, 3, 9, 8, 10, 0, 1, 3, 0

Solution

Since the value of 0 occurs more than any other value, it is the mode. In this instance, as a measure of location, the modal value is not a particularly appealing choice. However, as noted previously, the mode does possess one very favorable property—it is the only measure of location that can be applied to nominal data. Therefore, for nominal measurements like color preferences, it would be perfectly reasonable to discuss the modal color.

Suppose we added one more value to the data set in Example 4.1.7. If this value were a 1, then both 0 and 1 would be repeated three times and there would be two modes. When this occurs, the data are said to be **bimodal**. Any time a data set has more than two modes, it is said to be **multimodal**. If all observations in a data set occur with the same frequency, the data set has **no mode**.

The Relationship between the Mean, Median, and Mode

Oftentimes, the shape of the data determines how the mean, median, and mode are related. For a bell-shaped distribution, the mean, median, and mode are identical.

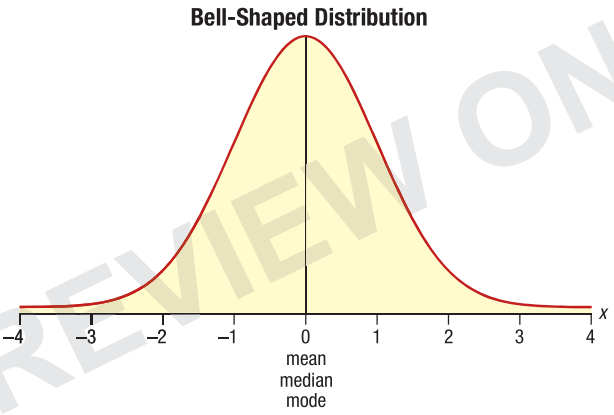


Figure 4.1.3

Certainly, not all data produce distributions that follow a bell-shaped curve. If the distribution of the data has a long tail on the right, it is said to be skewed to the right, or positively skewed. Conversely, if the distribution has a long tail on the left, it is said to be skewed to the left, or negatively skewed. If the data are positively skewed, the median will be smaller than the mean. If the data are negatively skewed, the mean will be smaller than the median.

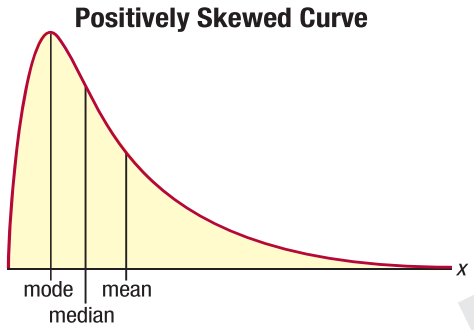


Figure 4.1.4

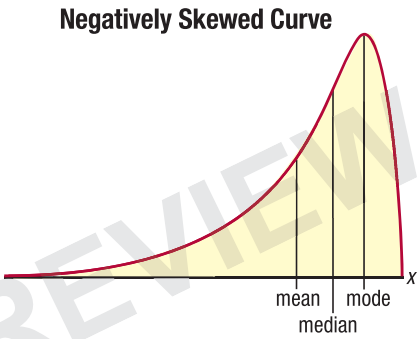
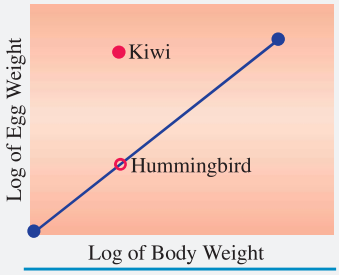


Figure 4.1.5

Where does the mode fall on these graphs? The area containing the greatest number of observations contains the mode. That area is represented by the large peak in the curve. In Figure 4.1.4 and Figure 4.1.5 the obvious peaks in the curves are the portions of the distributions that will contain the mode. The highest point on the curve will be the mode of the distribution. Notice that in a bell-shaped distribution (Figure 4.1.3), the mode is equal to the mean and median. If the distribution is positively skewed (Figure 4.1.4), the mode is less than the mean and median. If the distribution is negatively skewed (Figure 4.1.5), the mode is greater than the mean and median.



Kiwi Eggs Are Outliers!

Both animal and plant kingdoms offer spectacularly odd and beautiful sights. A kiwi bird (one of many interesting life forms from New Zealand) lays eggs that are close to 25% of their body weight and sometimes lays two or three such eggs at a time. For birds, eggs usually correspond to about 5% of the bird's weight among all species.

If you draw a graph relating (log) egg weights against (log) body weight you get a so-called hummingbird-moa curve (moa, an extinct ostrich like bird of the New Zealand area.) In this curve, the kiwi shows up as an outlier. Using the kiwi body weight (of about 5 lbs) one expects an egg weight of about 55 to 100 grams while the real weight of kiwi eggs is about 400 to 435 grams matching an expected body weight of about 40 lbs. Why? And what accounts for such an anomaly?

The most reasonable explanation provided by biologists is that kiwis and moa birds are members of the same species except the kiwis have dwarfed through their evolutionary history. A subarea of biology called "allometry" states that as body size decreases the internal organs decrease relatively slowly which supports the dwarfism hypothesis. The kiwis have lost body weight but not their internal womb structure which still holds large eggs. Outliers are important because they force you to think about your data more seriously.

Source: Gould, S. J. (1991). "Of Kiwi Eggs and the Liberty Bell" in Bully For Brontosaurus, W.W. Norton and Company, New York.

Selecting a Measure of Location

The objective of using descriptive statistics is to provide measures that convey useful summary information about the data. When selecting a statistic to represent the central value of a data set, the first thing to consider is the type of data being analyzed.

The arithmetic mean is used so frequently that its computation is almost a knee-jerk reaction to analyzing data. Unfortunately, it is not always a reasonable measure of centrality or location. Table 4.1.3 defines the applicable levels of measurement for each measure of location. Table 4.1.4 defines the sensitivity to outliers for each measure of location. When the data are qualitative (nominal or ordinal), the mean should not be calculated and, if the data are quantitative and contain outliers, the mean does not convey the notion of typical value as well as some other measures. The only time in which the mean should be used without any explanation is when the distribution of the data is symmetrical or nearly so. In that event, the mean and median should be approximately the same value.

Table 4.1.3 – Applicable Level of Measurement

	Qualitative Nominal	Ordinal	Quantitative Interval	Ratio
Mean			✔	✔
Median		✔	✔	✔
Mode	✔	✔	✔	✔
Trimmed Mean			✔	✔

Table 4.1.4 – Sensitivity to Outliers

	Not Sensitive	Very Sensitive
Mean		✔
Median	✔	
Mode	✔	
Trimmed Mean	✔	

The median is also a good measure of central tendency. It is not sensitive to outliers and can be applied to data gathered from all levels of measurements except nominal.

If the level of measurement of the data is interval or ratio and there are no outliers, the mean is a reasonable choice. If the data set appears to have any unusual values, then the trimmed mean or the median would be more appropriate.

If the level of measurement is nominal or ordinal (the data are qualitative), appropriate measures of center are limited. If the data are ordinal, then the median is the best choice. If the data are nominal, there is only one choice, the mode. The mode is applicable to all data types, although it is not very useful for quantitative data.

Time Series Data and Measures of Centrality

We discussed two types of time series data in Chapter 2, stationary and nonstationary. Stationary time series wobbled around some central value, so calculating a central value is perfectly reasonable, and the methods we previously discussed are applicable. A nonstationary time series is another story. Nonstationary time series possess trend. That means there is no central value for the time series. Instead, the series trends in one direction or another. Computing a central value using the methods discussed earlier would be inappropriate for such data.

Table 4.1.5 shows the first six rows of the data on average US gas price from 1991 to 2015. In this nonstationary time series, the central value of the process is trending upward as shown in Figure 4.1.6. One way to capture this movement is with a **moving average**.