Measures of Center for Ungrouped Data

A **measure of center** gives the center of a histogram or a frequency distribution curve. This section discusses five different measures of center,

1. The Mean
2. The Median
3. The Mode

The Mean:

Calculating Mean for Ungrouped Data The *mean for ungrouped data* is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data:} \quad \mu = \frac{\Sigma x}{N}$$

$$\text{Mean for sample data:} \quad \bar{x} = \frac{\Sigma x}{n}$$

where Σx is the sum of all values, N is the population size, n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean.

Example: 2014 Profits of 10 U.S. Companies

Table 3.1 lists the total profits (in million dollars) of 10 U.S. companies for the year 2014 (www.fortune.com).

Table 3.1 2014 Profits of 10 U.S. Companies

Company	Profits (million of dollars)
Apple	37,037
AT&T	18,249
Bank of America	11,431
Exxon Mobil	32,580
General Motors	5346
General Electric	13,057
Hewlett-Packard	5113
Home Depot	5385
IBM	16,483
Wal-Mart	16,022

Find the mean of the 2014 profits for these 10 companies.

Solution The variable in this example is 2014 profits of a company. Let us denote this variable by x. The 10 values of x are given in the above table. By adding these 10 values, we obtain the sum of x values, that is:

$$\Sigma x = 37{,}037 + 18{,}249 + 11{,}431 + 32{,}580 + 5346 + 13{,}057 + 5113 + 5385 + 16{,}483 + 16{,}022 = 160{,}703$$

Note that the given data include only 10 companies. Hence, it represents a sample with $n = 10$. Substituting the values of Σx and n in the sample formula, we obtain the mean of 2014 profits of 10 companies as follows:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{160{,}703}{10} = 16{,}070.3 = \textbf{\$16,070.3 million}$$

Thus, these 10 companies earned an average of \$16,070.3 million profits in 2014. ■

Example: Ages of Employees of a Company

The following are the ages (in years) of all eight employees of a small company:

53 32 61 27 39 44 49 57

Find the mean age of these employees.

Solution Because the given data set includes *all* eight employees of the company, it represents the population. Hence, $N = 8$. We have

$$\sum x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

The population mean is

$$\mu = \frac{\sum x}{N} = \frac{362}{8} = \mathbf{45.25\ years}$$

Thus, the mean age of all eight employees of this company is 45.25 years, or 45 years and 3 months. ■

Illustrating the effect of an outlier on the mean

Example: Prices of Eight Homes

Following are the list prices of eight homes randomly selected from all homes for sale in a city.

$245,670	176,200	360,280	272,440
450,394	310,160	393,610	3,874,480

Note that the price of the last house is $3,874,480, which is an outlier. Show how the inclusion of this outlier affects the value of the mean.

Solution If we do not include the price of the most expensive house (the outlier), the mean of the prices of the other seven homes is:

$$\text{Mean without the outlier} = \frac{245{,}670 + 176{,}200 + 360{,}280 + 272{,}440 + 450{,}394 + 310{,}160 + 393{,}610}{7}$$

$$= \frac{2{,}208{,}754}{7} = \mathbf{\$315{,}536.29}$$

Now, to see the impact of the outlier on the value of the mean, we include the price of the most expensive home and find the mean price of eight homes. This mean is:

Mean with the outlier

$$= \frac{245{,}670 + 176{,}200 + 360{,}280 + 272{,}440 + 450{,}394 + 310{,}160 + 393{,}610 + 3{,}874{,}480}{8}$$

$$= \frac{6{,}083{,}234}{8} = \mathbf{\$760{,}404.25}$$

Thus, when we include the price of the most expensive home, the mean more than doubles, as it increases from $315,536.29 to $760,404.25. ■

We should remember that the

Note:

Mean is not always the best measure of center because it is heavily influenced by outliers. Sometimes other measures of center give a more accurate impression of a data set. For example, when a data set has outliers, instead of using the mean, we can use either the trimmed mean or the median as a measure of center.

The Median:

Median The **median** is the value that divides a data set that has been ranked in increasing order in two equal halves. If the data set has an odd number of values, the median is given by the value of the middle term in the ranked data set. If the data set has an even number of values, the median is given by the average of the two middle values in the ranked data set.

As is obvious from the definition of the median, it divides a ranked data set into two equal parts. The calculation of the median consists of the following two steps:

1. Rank the given data set in increasing order.
2. Find the value that divides the ranked data set in two equal parts. This value gives the median.[1]

Note that if the number of observations in a data set is *odd*, then the median is given by the value of the middle term in the ranked data. However, if the number of observations is *even*, then the median is given by the average of the values of the two middle terms.

Calculating the median for ungrouped data: odd number of data values

Example: Compensations of Female CEOs

Table 3.2 lists the 2014 compensations of female CEOs of 11 American companies (*USA TODAY*, May 1, 2015). (The compensation of Carol Meyrowitz of TJX is for the fiscal year ending in January 2015.)

Table 3.2 **Compensations of 11 Female CEOs**

Company & CEO	2014 Compensation (millions of dollars)
General Dynamics, Phebe Novakovic	19.3
GM, Mary Barra	16.2
Hewlett-Packard, Meg Whitman	19.6
IBM, Virginia Rometty	19.3
Lockheed Martin, Marillyn Hewson	33.7
Mondelez, Irene Rosenfeld	21.0
PepsiCo, Indra Nooyi	22.5
Sempra, Debra Reed	16.9
TJX, Carol Meyrowitz	28.7
Yahoo, Marissa Mayer	42.1
Xerox, Ursula Burns	22.2

Find the median for these data.

Solution To calculate the median of this data set, we perform the following two steps.

Step 1: The first step is to rank the given data. We rank the given data in increasing order as follows:

16.2 16.9 19.3 19.3 19.6 21.0 22.2 22.5 28.7 33.7 42.1

Step 2: The second step is to find the value that divides this ranked data set in two equal parts. Here there are 11 data values. The sixth value divides these 11 values in two equal parts. Hence, the sixth value gives the median as shown below.

16.2 16.9 19.3 19.3 19.6 **21.0** 22.2 22.5 28.7 33.7 42.1

↑ Median

Thus, the median of 2014 compensations for these 11 female CEOs is \$21.0 million. Note that in this example, there are 11 data values, which is an odd number. Hence, there is one value in the middle that is given by the sixth term, and its value is the median. Using the value of the median, we can say that half of these CEOs made less than \$21.0 million and the other half made more than \$21.0 million in 2014. ■

Calculating the median for ungrouped data: even number of data values.

Example: Cell Phone Minutes Used

The following data give the cell phone minutes used last month by 12 randomly selected persons.

230 2053 160 397 510 380 263 3864 184 201 326 721

Find the median for these data.

Solution To calculate the median, we perform the following two steps.

Step 1: In the first step, we rank the given data in increasing order as follows:

160 184 201 230 263 326 380 397 510 721 2053 3864

Step 2: In the second step, we find the value that divides the ranked data set in two equal parts. This value will be the median. The value that divides 12 data values in two equal parts falls between the sixth and the seventh values. Thus, the median will be given by the average of the sixth and the seventh values as follows.

160 184 201 230 263 **326 380** 397 510 721 2053 3864

↑ Median = 353

$$\text{Median} = \text{Average of the two middle values} = \frac{326 + 380}{2} = \textbf{353 minutes}$$

Thus, the median cell phone minutes used last month by these 12 persons was 353. We can state that half of these 12 persons used less than 353 cell phone minutes and the other half used more than 353 cell phone minutes last month. Note that this data set has two outliers, 2053 and 3864 minutes, but these outliers do not affect the value of the median. ■

Note:

The median gives the center of a histogram, with half of the data values to the left of the median and half to the right of the median. The advantage of using the median as a measure of center is that it is not influenced by outliers. Consequently, the median is preferred over the mean as a measure of center for data sets that contain outliers.

The Mode:

Mode The *mode* is the value that occurs with the highest frequency in a data set.

Example: Speeds of Cars

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77 82 74 81 79 84 74 78

Find the mode.

Solution In this data set, 74 occurs twice, and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode. Therefore,

Mode = **74 miles per hour** ■

Data set with no mode

Example: Incomes of Families

Last year's incomes of five randomly selected families were \$76,150, \$95,750, \$124,985, \$87,490, and \$53,740. Find the mode.

Solution Because each value in this data set occurs only once, this data set contains **no mode**. ■

Data set with two modes

Example: Commuting Times of Employees

A small company has 12 employees. Their commuting times (rounded to the nearest minute) from home to work are 23, 36, 14, 23, 47, 32, 8, 14, 26, 31, 18, and 28, respectively. Find the mode for these data.

Solution In the given data on the commuting times of these 12 employees, each of the values 14 and 23 occurs twice, and each of the remaining values occurs only once. Therefore, this data set has two modes: 14 and 23 minutes. ■

Data set with three modes

Example: Ages of Students

The ages of 10 randomly selected students from a class are 21, 19, 27, 22, 29, 19, 25, 21, 22, and 30 years, respectively. Find the mode.

Solution This data set has three modes: **19**, **21**, and **22**. Each of these three values occurs with a (highest) frequency of 2. ■

The mode for qualitative data

Example: Status of Students

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, and senior, respectively. Find the mode.

Solution Because **senior** occurs more frequently than the other categories, it is the mode for this data set. We cannot calculate the mean and median for this data set. ■

Note:

A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median.
One advantage of the mode is that it can be calculated for both kinds of data—quantitative and qualitative—whereas the mean and median can be calculated for only quantitative data.

Note:

We cannot say for sure which of the various measures of center is a better measure overall. Each of them may be better under different situations. Probably the mean is the most used measure of center, followed by the median. The mean has the advantage that its calculation includes each value of the data set. The median and trimmed mean are better measures when a data set includes outliers. The mode is simple to locate, but it is not of much use in practical applications.

Relationships Among the Mean, Median, and Mode

1. For a **symmetric histogram** and frequency distribution curve with one peak (see Figure 3.2), the values of the mean, median, and mode are identical, and they lie at the center of the distribution.

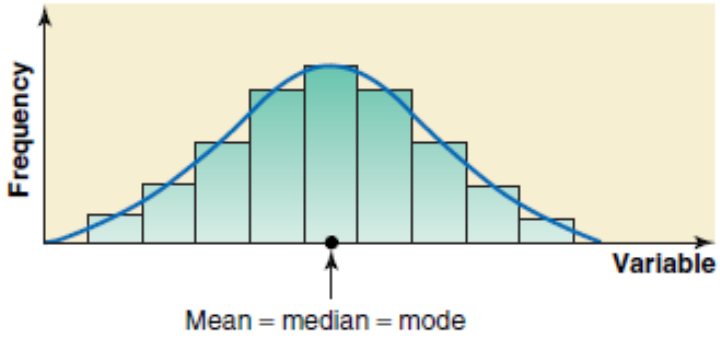


Figure 3.2 Mean, median, and mode for a symmetric histogram and frequency distribution curve.

2. For a histogram and a frequency distribution curve **skewed to the right** (see Figure 3.3), the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two. (Notice that the mode always occurs at the peak point.) The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail. These outliers pull the mean to the right.

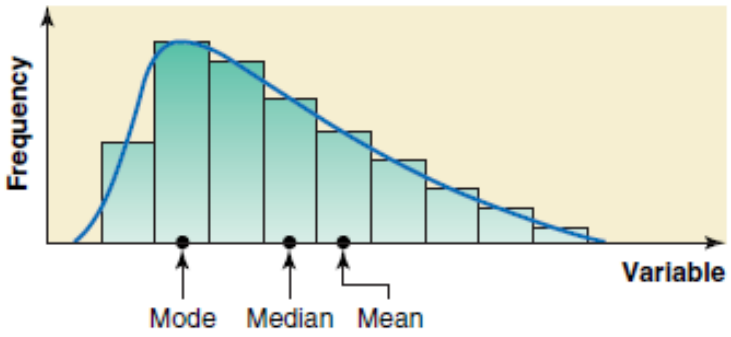


Figure 3.3 Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.

3. If a histogram and a frequency distribution curve are **skewed to the left** (see Figure 3.4), the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two. In this case, the outliers in the left tail pull the mean to the left.

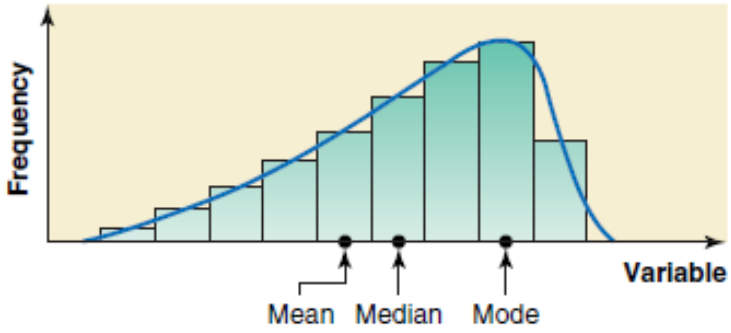


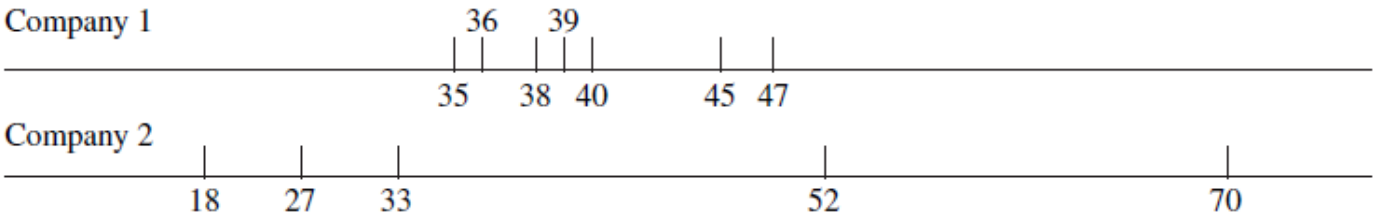
Figure 3.4 Mean, median, and mode for a histogram and frequency distribution curve skewed to the left.

Measures of Dispersion for Ungrouped Data

The measures of center, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than for the other data set. (Note that the words dispersion, spread, and variation have similar meanings.) Consider the following two data sets on the ages (in years) of all workers at each of two small companies.

Company 1:	47	38	35	40	36	45	39
Company 2:	70	33	18	52	27		

The mean age of workers in both these companies is the same, 40 years. If we do not know the ages of individual workers at these two companies and are told only that the mean age of the workers at both companies is the same, we may deduce that the workers at these two companies have a similar age distribution. As we can observe, however, the variation in the worker's ages for each of these two companies is very different. As illustrated in the diagram, the ages of the workers at the second company have a much larger variation than the ages of the workers at the first company.



Thus, a summary measure such as the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set. We also need a measure that can provide some information about the variation among data values. The measures that help us learn about the spread of a data set are called the measures of dispersion. The measures of center and dispersion taken together give a better picture of a data set than the measures of center alone.

This section discusses four measures of dispersion:

4. Range
5. Variance
6. Standard deviation
7. Coefficient of variation.

The Range:

Finding the Range for Ungrouped Data

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Example: Total Areas of Four States

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

Solution The largest total area for a state in this data set is 267,277 square miles, and the smallest area is 49,651 square miles. Therefore,

$$\begin{aligned}\text{Range} &= \text{Largest value} - \text{Smallest value} \\ &= 267{,}277 - 49{,}651 = \mathbf{217{,}626 \text{ square miles}}\end{aligned}$$

Thus, the total areas of these four states are spread over a range of 217,626 square miles. ■

Note:

The range, like the mean, has the disadvantage of being influenced by outliers. the range is not a good measure of dispersion to use for a data set that contains outliers. This indicates that the range is a nonresistant measure of dispersion.

Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

Variance and Standard Deviation:

The standard deviation is the most-used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around the mean. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean. In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

The standard deviation is obtained by taking the positive square root of the variance.

Short-Cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\Sigma x^2 - \left(\frac{(\Sigma x)^2}{N}\right)}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \left(\frac{(\Sigma x)^2}{n}\right)}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\frac{\Sigma x^2 - \left(\frac{(\Sigma x)^2}{N}\right)}{N}}$

Sample standard deviation: $s = \sqrt{\frac{\Sigma x^2 - \left(\frac{(\Sigma x)^2}{n}\right)}{n-1}}$

Two Observations:

1. The values of the variance and the standard deviation are never negative
2. The measurement units of the variance are always the square of the measurement units of the original data.

Calculating the sample variance and standard deviation for ungrouped data

Example: Compensations of Female CEOs

Refer to the 2014 compensations of 11 female CEOs of American companies given in Example 3–4. The table from that example is reproduced below.

Company & CEO	2014 Compensation (millions of dollars)
General Dynamics, Phebe Novakovic	19.3
GM, Mary Barra	16.2
Hewlett-Packard, Meg Whitman	19.6
IBM, Virginia Rometty	19.3
Lockheed Martin, Marillyn Hewson	33.7
Mondelez, Irene Rosenfeld	21.0
PepsiCo, Indra Nooyi	22.5
Sempra, Debra Reed	16.9
TJX, Carol Meyrowitz	28.7
Yahoo, Marissa Mayer	42.1
Xerox, Ursula Burns	22.2

Find the variance and standard deviation for these data.

Solution Let x denote the 2014 compensations (in millions of dollars) of female CEOs of American companies. The calculation of Σx and Σx^2 is shown in Table 3.6.

Table 3.6 Calculation of Σx and Σx^2

x	x^2
19.3	372.49
16.2	262.44
19.6	384.16
19.3	372.49
33.7	1135.69
21.0	441.00
22.5	506.25
16.9	285.61
28.7	823.69
42.1	1772.41
22.2	492.84
$\Sigma x = 261.5$	$\Sigma x^2 = 6849.07$

Calculation of the variance involves the following three steps.

Step 1. *Calculate* Σx.

The sum of the values in the first column of Table 3.6 gives the value of Σx, which is 261.5.

Step 2. *Find* Σx^2.

The value of Σx^2 is obtained by squaring each value of x and then adding the squared values. The results of this step are shown in the second column of Table 3.6. Notice that $\Sigma x^2 = 6849.07$.

Step 3. *Determine the variance.*

Substitute the values of n, Σx, and Σx^2 in the variance formula and simplify. Because the given data are on the 2014 compensations of 11 female CEOs of American companies, we use the formula for the sample variance using $n = 11$.

$$s^2 = \frac{\Sigma x^2 - \left(\frac{(\Sigma x)^2}{n}\right)}{n-1} = \frac{6849.07 - \left(\frac{(261.5)^2}{11}\right)}{11-1} = \frac{6849.07 - 6216.5682}{10} = \mathbf{63.2502}$$

Now to obtain the standard deviation, we take the (positive) square root of the variance. Thus,

$$s = \sqrt{63.2502} = 7.952999 = \mathbf{\$7.95 \text{ million}}$$

Thus, the standard deviation of the 2014 compensations of these 11 female CEOs of American companies is \$7.95 million. ■

Calculating the population variance and standard deviation for ungrouped data.

Example: Earnings of Employees

The following data give the 2015 earnings (in thousands of dollars) before taxes for all six employees of a small company.

88.50 108.40 65.50 52.50 79.80 54.60

Calculate the variance and standard deviation for these data.

Solution Let x denote the 2015 earnings before taxes of an employee of this company. The values of $\sum x$ and $\sum x^2$ are calculated in Table 3.7.

Table 3.7

x	x^2
88.50	7832.25
108.40	11,750.56
65.50	4290.25
52.50	2756.25
79.80	6368.04
54.60	2981.16
$\sum x = 449.30$	$\sum x^2 = 35{,}978.51$

Because the data in this example are on earnings of *all* employees of this company, we use the population formula to compute the variance using $N = 6$. Thus, the variance is

$$\sigma^2 = \frac{\sum x^2 - \left(\frac{(\sum x)^2}{N}\right)}{N} = \frac{35{,}978.51 - \left(\frac{(449.30)^2}{6}\right)}{6} = \mathbf{388.90}$$

The standard deviation is obtained by taking the (positive) square root of the variance:

$$\sigma = \sqrt{\frac{\sum x^2 - \left(\frac{(\sum x)^2}{N}\right)}{N}} = \sqrt{388.90} = \mathbf{19.721 \text{ thousand} = \$19{,}721}$$

Thus, the standard deviation of the 2015 earnings of all six employees of this company is \$19,721. ■

Coefficient of Variation

One disadvantage of the standard deviation as a measure of dispersion is that it is a measure of absolute variability and not of relative variability. Sometimes we may need to compare the variability for two different data sets that have different units of measurement. In such cases, a measure of relative variability is preferable. One such measure is the **coefficient of variation**.

Coefficient of Variation

The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean and is computed as follows.

For population data: $$CV = \frac{\sigma}{\mu} \times 100\%$$

For sample data: $$CV = \frac{s}{\bar{x}} \times 100\%$$

Note that the coefficient of variation does not have any units of measurement, as it is always expressed as a percent.

Example: Salaries and Education

The yearly salaries of all employees working for a large company have a mean of $72,350 and a standard deviation of $12,820. The years of schooling (education) for the same employees have a mean of 15 years and a standard deviation of 2 years. Is the relative variation in the salaries higher or lower than that in years of schooling for these employees? Answer the question by calculating the coefficient of variation for each variable.

Solution Because the two variables (salary and years of schooling) have different units of measurement (dollars and years, respectively), we cannot directly compare the two standard deviations. Hence, we calculate the coefficient of variation for each of these data sets.

$$\text{CV for salaries} = \frac{\sigma}{\mu} \times 100\% = \frac{12{,}820}{72{,}350} \times 100\% = \mathbf{17.72\%}$$

$$\text{CV for years of schooling} = \frac{\sigma}{\mu} \times 100\% = \frac{2}{15} \times 100\% = \mathbf{13.33\%}$$

Thus, the standard deviation for salaries is 17.72% of its mean and that for years of schooling is 13.33% of its mean. Since the coefficient of variation for salaries has a higher value than the coefficient of variation for years of schooling, the salaries have a higher relative variation than the years of schooling. ■

Population Parameters and Sample Statistics

A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a population parameter, or simply a parameter. A summary measure calculated for a sample data set is called a sample statistic, or simply a statistic. Thus, μ and σ are population parameters, and x and s are sample statistics.