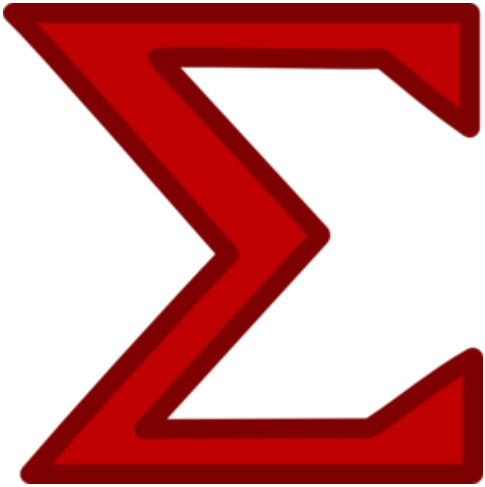
Disclaimer

Chapter 4

SUMMARIZING DATA



CASE STUDY:

What is the average height of Dr. Z's female patients?

To answer the question, Dr. Z collected and summarized the data. It was found that the average height was 64 inches with a standard deviation of 2 inches.

So, what was the methodology used to determine the values?

Well, he had to do the following:

1. Compute the mean.
2. Compute the standard deviation.

In the following chapter, you will learn about these concepts and their applications.

You will also learn how to:

3. Compute measures of central tendency.
4. Compute measures of variation.
5. Compute measures of position.
6. Understand and apply the Empirical Rule.
7. Understand and apply Exploratory Data Analysis.

Section 4.1:

Measures of Central Tendency

Student Learning Outcomes

By the end of the section,

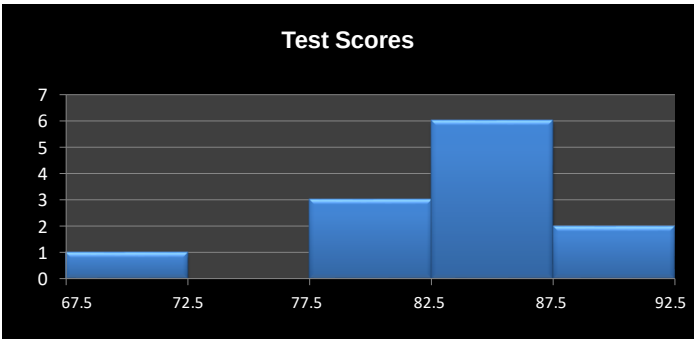
1. **You will compute the mean, median and mode.**
2. **You will determine which measure of central tendency should be reported.**

In Chapter 2, concepts related to organizing data were examined. As a result, graphical representations of data were constructed. Graphs can be used to show the shape of a distribution. In addition to determining the shape of a distribution, researchers can determine the center and spread (i.e. variation). Let's explore the following example:

Mr. Scott wanted to know whether the use of calculators on a test would help improve math scores. He sampled 12 of his students. The following scores were obtained:

90	92	87
83	80	81
86	85	81
68	87	86

A graphical representation resulted in the following:



The graph shows the data is skewed left.

What other information can be gathered about the distribution in addition to understanding the shape?

To obtain a *good* description of the overall pattern of data, researchers must also discuss the center and spread (i.e. variation). In this section, measures of central tendency will be discussed. In the next section, measures of variation will be

discussed. These measures can be gathered from a population or a sample.

DEFINITION

Parameter: a numerical summary gathered from a population.

Statistic: a numerical summary gathered from a sample.

Since most of the time researchers will be examining samples to make inferences about the population, statistics will be computed.

Nonetheless, the notation for parameters will also be discussed. It will be helpful for understanding inferences in the later chapters.

There are 3 common measures used to describe the center of a distribution:

- Mean
- Median
- Mode

DEFINITION

Mean: a value that is calculated by summing the values of the observations, then dividing by the total number of observations.

Population Mean: $\mu = \frac{\sum x}{N}$

Sample Mean: $\bar{x} = \frac{\sum x}{n}$

Summary of Notations:

μ	**lower case Greek letter pronounced as "mu"**
$\bar{x}$	**read as "x-bar"**
$\sum$	**upper case Greek letter Sigma, which means sum**
x	**each x-value**
n	**sample size**
N	**population size**

Therefore, to compute the mean for Mr. Scott's class, the formula for the sample mean will be used. The formula for the *sample mean* is used because Mr. Scott selected a *sample* of 12 test scores from his class to analyze. If he used test scores from *all* students in his class, then the population mean would have been computed.

Thus, the sample mean is the following:

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{x} = \frac{90+83+86+68+92+80+85+87+87+81+81+86}{12}$$

$$\bar{x} = \frac{1006}{12}$$

$$\bar{x} = 83.83$$

The sample mean is 83.8. Thus, the mean test score is 83.8.

A Note about Rounding: Round one or two more decimal place than the original data values. Also, round off at the FINAL step of calculations. Rounding too soon in the computation will result in an incorrect answer! This rule will ALWAYS be used throughout the text unless otherwise noted.

DEFINITION

Median: a value that lies in the middle of the data set. It divides the data set into two equal parts.

To compute the median, perform the following steps:

Computing the Median

Step 1: Order the data from least to greatest.

Step 2: If n is odd, the median is the middle observation. If n is even, the median is the sum of the two middle observations divided by 2.

Using the data from Mr. Scott's class, let's compute the median.

Step 1: *Order the data from least to greatest.*

68	83	87
80	85	87
81	86	90
81	86	92

Step 2: *If n is odd, the median is the middle observation. If n is even, the median is the sum of the two middle observations divided by 2.* In this example, n is even. The two middle observations are 85 and 86. Add them together. The result is

171. Finally, divide by 2. 171 ÷ 2 = 85.5. Thus, the median is 85.5.

DEFINITION

Mode: the most frequently occurring observation.

By looking at the data set, the most frequently occurring observations are 81, 86 and 87.

If there is one mode, the distribution is considered unimodal. If there are two modes, the distribution is considered bimodal. For 3 or more modes, the distribution is considered multimodal. If each observation occurs once, the distribution has no mode. For this example, the distribution is considered **multimodal**.

Suppose Mr. Scott decided to offer extra credit by adding 5 points to each score. What would happen to each measure of central tendency? Let's re-examine the test scores with the extra credit added.

95	97	92
88	85	86
91	90	86
73	92	91

The mean = 88.8, the median = 90.5 and the modes are 86, 91, 92.

The mean and median increased by 5 points. In addition, the values of the modes increased by 5 points.

Therefore, adding "k" units to each data value will increase the values of the mean, median and mode by "k" units. Conversely, subtracting "k" units from each data value will decrease the

values of the mean, median and mode by "k" units. The same is applied when using multiplication and division.

As stated, the mean, median and mode can be used to describe the center of a distribution. There are times when one measure is preferred over the others. The mode is rarely used when describing quantitative data. It is most preferred when describing qualitative data.

Therefore, ***mean vs. median***, which is best? The answer depends on the shape of the distribution. If the distribution is symmetrical, the mean is the best choice because it takes into account every observation. The mean is the most common measure of central tendency. However, if the distribution is skewed, report the median.

Let's look at an example to see why that is the case.

Suppose a sample of 5 students in a class is randomly chosen. They are asked to report their annual salary. The results are below (in dollars).

21,000 17,000 25,000 20,000 110,000

The mean is 38,600 and the median is 21,000.

The following is a graphical representation of the data (in thousands of dollars):

Stem	Leaf
1	7
2	0 1 5
3	
4	
5	
6	
7	
8	
9	
10	
11	0

So, which measure should be reported? The answer is the median since the distribution is skewed. The median value of $21,000 should be reported because it is a better reflection of the typical value in the data set (i.e. most of the salaries are in the $20K range). A typical student in the class is not making $38,600.

In addition, the mean is larger than the median. The mean will always be larger than the median in skewed right distributions because the extreme value is "pulling" the mean in the direction of the skew.

Conversely, when the mean is smaller than the median, the distribution will be skewed left. When the mean and median are the same, the distribution will be symmetrical.

In conclusion, the term "average" does not always refer to the mean. The median can be used to represent the "average" when the distribution is skewed.

The following table can be used as a reference:

Central Measure to Report	Shape of Distribution
Mean	Symmetrical
Median	Skewed left
Median	Skewed right

For qualitative data, only report the mode.

Section 4.2:

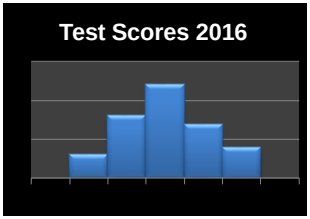
Measures of Variation

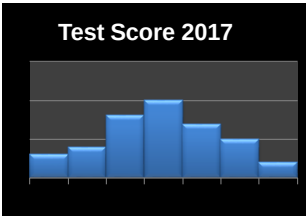
Student Learning Outcomes

By the end of the section,

1. **You will compute the range, variance and standard deviation.**
2. **You will apply the Empirical Rule to a set of data.**
3. **You will use the Empirical Rule to find probabilities.**

Let's examine the following two graphs.





Both graphs have a mean of 80. However, the graph on the right is more "spread out".

Spread, also known as variation, can tell researchers valuable information about a data set. There are 3 common measures used to describe spread:

- Range
- Variance
- Standard deviation

Using the data from Mr. Scott's class (Section 4.1: Measures of Central Tendency), let's determine the above measures.

Recall, Mr. Scott wanted to know whether the use of calculators on a test would help improve math scores. He sampled 12 of his students. The following scores were obtained:

90	92	87
83	80	81
86	85	81
68	87	86

DEFINITION

Range: a value that is determined by taking the difference between the largest value (Maximum) and the smallest value (Minimum).

Range = Maximum – Minimum

The largest value in the data set is 92. The smallest value is 68. Thus, the range is 92 – 68 = 24.

The range can be useful when comparing more than one distribution. For example, a distribution with a range of 48 is more "spread out" (i.e. variable) than a distribution with a range of 16. However, the range doesn't typically provide much information about a single distribution since only two values are used.

DEFINITION

Variance: a value that is the average of the squares of the deviations from the mean of a data set.

Population Variance: $\sigma^2 = \dfrac{\sum(x-\mu)^2}{N}$

Sample Variance: $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$

Note: σ is the lower case Greek letter Sigma.

The variance is rarely used in reports. However, it is used to compute the standard deviation, which is most useful when discussing the spread of a distribution.

Since there was a *sample* of 12 students, the *sample* variance will be computed.

To compute the sample variance, first determine the numerator. The numerator is the sum of the squared deviations. A *deviation* is found by taking a data value minus the mean. A *squared deviation* is found by simply squaring each deviation. Recall, in Section 4.1: Measures of Central Tendency, the mean was 83.83.

x	$x - \bar{x}$ (Deviations)	$(x - \bar{x})^2$ (Square of the Deviations)
90	90 − 83.83 = 6.17	$(6.17)^2$ = 38.0689
92	92 − 83.83 = 8.17	$(8.17)^2$ = 66.7489
87	87 − 83.83 = 3.17	$(3.17)^2$ = 10.0489
83	83 − 83.83 = −0.83	$(-0.83)^2$ = 0.6889
80	80 − 83.83 = −3.83	$(-3.83)^2$ = 14.6689
81	81 − 83.83 = −2.83	$(-2.83)^2$ = 8.0089
86	86 − 83.83 = 2.17	$(2.17)^2$ = 4.7089
85	85 − 83.83 = 1.17	$(1.17)^2$ = 1.3689
81	81 − 83.83 = −2.83	$(-2.83)^2$ = 8.0089
68	68 − 83.83 = −15.83	$(-15.83)^2$ = 250.5889
87	87 − 83.83 = 3.17	$(3.17)^2$ = 10.0489
86	86 − 83.83 = 2.17	$(2.17)^2$ = 4.7089

$\sum(x - \bar{x})^2$ = 38.0689 + 66.7489 + 10.0489 + 0.6889 + 14.6689 + 8.0089 + 4.7089 + 1.3689 + 8.0089 + 250.5889 + 10.0489 + 4.7089 = 417.6668

Now that the numerator is known, place this value into the formula for the sample variance. The denominator is determined by taking the sample size minus 1.

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

$$s^2 = \frac{417.6668}{12-1}$$

$$s^2 = \frac{417.6668}{11}$$

$$s^2 = 37.97$$

The sample variance is 37.97.

The sample variance can be used to compute the standard deviation. As with the variance, either the population standard deviation or sample standard deviation can be computed.

DEFINITION

Standard Deviation: a value that is the square root of the variance. The value represents the typical distance between a data value and the mean.

Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$

Sample Standard Deviation: $s = \sqrt{s^2}$

Thus, the sample standard deviation is the following:

$$s = \sqrt{s^2}$$

$$s = \sqrt{37.97}$$

$$s = 6.16$$

So, what does this value tell researchers? Well, it means that, on average, test scores deviate about 6 points from the mean. The standard deviation tells researchers the measure of spread around the mean (i.e. center) of the data set. In addition, it tells researchers on average how much the observations deviate from the mean.

The value of ***s*** represents 1 standard deviation.

Note: When the standard deviation is equal to 0, there is no spread. Thus, all observations are equal and there are no deviations from the mean. If the observations are not equal, the standard deviation will be positive. Moreover, a larger standard deviation denotes more spread.

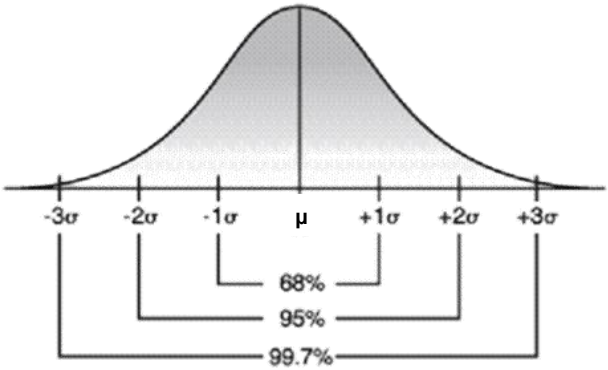
The standard deviation can also be useful in understanding the overall pattern of the distribution by using the Empirical Rule.

DEFINITION

Empirical Rule (*applies to symmetrical distributions ONLY*):

- **About 68% of the observations lie within 1 standard deviation of the mean.**
- **About 95% of the observations lie within 2 standard deviations of the mean.**
- **About 99.7% of the observations lie within 3 standard deviations of the mean.**

The following is an illustration of the Empirical Rule:



- −1σ can be found by subtracting the standard deviation once from the mean
- −2σ can be found by subtracting the standard deviation twice from the mean
- −3σ can be found by subtracting the standard deviation three times from the mean
- +1σ can be found by adding the standard deviation once to the mean
- +2σ can be found by adding the standard deviation twice to the mean
- +3σ can be found by adding the standard deviation three times to the mean

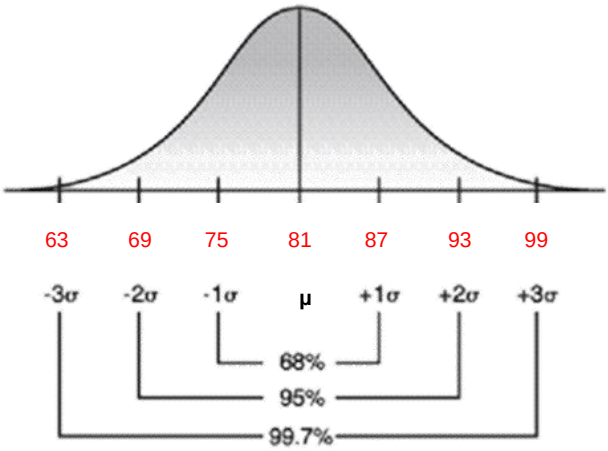
Note: *s* can be used in place of σ and the sample mean can be used in place of the population mean when applicable.

How does this rule apply to a set of data? Well, this is a general rule that applies to most distributions, specifically distributions that are symmetrical. Unfortunately, this rule cannot be applied to the data from Mr. Scott's class because the distribution was skewed left. Therefore, let's explore another example.

Miss Brown wanted to know if take-home tests would result in higher test scores than traditional in-class tests. She sampled 9 of her students. The test scores are the following:

83	89	77
83	72	75
90	81	79

The sample mean = 81. The sample standard deviation = 6.0. Since the sample mean and median = 81, it can be concluded that this distribution is symmetrical. The following illustration shows the rule applied to this set of data:



The values in red were determined by performing the following:

Recall, the sample mean = 81 and the sample standard deviation = 6.0.

−1s = 81 – 6.0 = 75

−2s = 81 – 6.0 – 6.0 = 69

−3s = 81 – 6.0 – 6.0 – 6.0 = 63

+1s = 81 + 6.0 = 87

+2s = 81 + 6.0 + 6.0 = 93

+3s = 81 + 6.0 + 6.0 + 6.0 = 99

Thus, about 68% of the test scores are between 75 and 87, about 95% of the test scores are between 69 and 93 and about 99.7% of the test scores are between 63 and 99.

Using the Empirical Rule provides great information about a distribution. For this example, it shows that most of the observations are between 63 and 99.

Section 4.3:

Measures of Position

Student Learning Outcomes

By the end of the section,

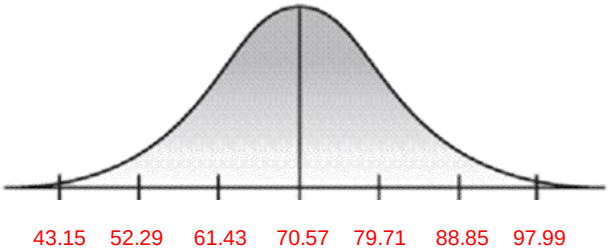
1. **You will compute z-scores.**
2. **You will compute percentiles and quartiles.**

Thus far, the center and spread of a distribution has been discussed. In this section, an observation's relative *position* within a data set will be explored. There are 2 common ways to describe an observation's position, in relation to others, within a data set. First, z-scores will be examined.

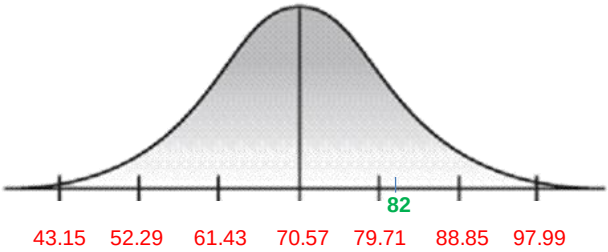
Suppose there is a group of students who took a final exam in an Ethics course. The scores of a sample of 7 students follow a symmetrical distribution and are listed below.

61 80 58 82 71 75 67

The sample mean = 70.57. The sample standard deviation = 9.14. The following chart shows the Empirical Rule applied to this set of data:



For a student who scored an **82**, the score falls between 1 and 2 standard deviations **ABOVE** the mean. This can be seen below.



Although the score of 82 is between 1 and 2 standard deviations above the mean, it's best to be exact.

To determine exactly how many standard deviations an observation is away from the mean, a z-score is computed.

DEFINITION

z-score: a value that represents the number of standard deviations an observation is away from the mean.

The formula for the z-score is the following:

$$z = \frac{(x - \bar{x})}{s}$$

Thus, the z-score for the test score of 82 is the following:

$$z = \frac{(82 - 70.57)}{9.14} = 1.25$$

The test score of 82 is 1.25 standard deviations **ABOVE** the mean. The term above is used because the z-score is positive.

What is the z-score for a student who scored 58 on the exam? How many standard deviations is the observation away from the mean? To determine this, compute a z-score.

$$z = \frac{(58 - 70.57)}{9.14} = -1.38$$

The test score of 58 is 1.38 standard deviations **BELOW** the mean. The term below is used because the z-score is negative.

Z-scores can be very informative for researchers. If researchers were using final exam scores as an example, a person's final

exam z-score equal to 3 performed much better than a person whose final exam z-score was equal to −2.8. However, if researchers were analyzing final scores of a major golf tournament, smaller scores equated to better performance. Thus, in this situation, a person whose z-score was −1.75 performed much better than a person whose z-score was 2.75.

If a z-score is equal to 0, then the observation is equal to the mean. Z-scores that are less than −2 or greater than 2 are considered *unusual*. Z-scores that are less than −3 or greater than 3 are considered *very unusual*.

Another way of determining an observation's relative position within a data set is to compute a percentile. Percentiles are commonly used in education and health.

DEFINITION

Percentile: the percentage of observations that fall *below* a particular data value.

The percentile is often referred to as the pth *percentile* with p representing a number ranging between 0 and 100. Selecting the value of the percentile can vary based on a researcher's preference.

Using the final exam scores from a sample of 7 students in an Ethics course, the percentile of each score can be determined. What is the score that corresponds to the 36th percentile? To answer this question, perform the following steps:

Determining the pth Percentile

Step 1: Order the data from least to greatest.

Step 2: Compute: $\frac{n \cdot p}{100}$ **. This value will correspond to the position of the observations which were ordered in step 1.**

Step 3: If the value determined in step 2 is an integer, find the observation in the data set which represents that position. Also, find the next highest observation. Add them together, then divide by 2. The observation will be the pth percentile. If the value determined in step 2 is *not* an integer, round *up* to the *next highest* integer. Find the observation in the data set which represents that position. The observation will be the pth percentile.

Using the steps above, determine the 36th percentile.

Step 1:

58 61 67 71 75 80 82

Step 2:

$$\frac{n \cdot p}{100} = \frac{7 \cdot 36}{100} = 2.52$$

Step 3: 2.52 is not an integer. Thus, round up to 3. The 3rd observation in the data set is 67. Therefore, 67 represents the 36th percentile.

What is the score that corresponds to the 75th percentile?

Step 1:

58 61 67 71 75 80 82

Step 2:

$$\frac{n \cdot p}{100} = \frac{7 \cdot 75}{100} = 5.25$$

Step 3: 5.25 is not an integer. Thus, round up to 6. The 6th observation in the data set is 80. Therefore, 80 represents the 75th percentile.

The 75th percentile is also known as the 3rd quartile.

DEFINITION

Quartiles: specific percentiles that divide the data set into 4 parts.

The quartiles are the following:

1st quartile (Q_1) = 25th percentile

2nd quartile (Q_2) = 50th percentile (also represents the Median)

3rd quartile (Q_3) = 75th percentile

Section 4.4:

Exploratory Data Analysis

Student Learning Outcomes

By the end of the section,

1. **You will demonstrate the calculation of the five-number summary.**
2. **You will explore the construction of a boxplot.**
3. **You will determine whether an observation is an outlier.**
4. **You will identify statistics that are resistant measures.**

In traditional statistics, researchers generally begin with a research question and would like to prove a claim to be true. To prove the claim, data must be collected, organized, summarized and analyzed.

When collecting data, researchers like to discover various aspects about the data set. Typically, researchers organize the data set using frequency distributions and graphs such as histograms to help provide a visualization of the data. In addition, researchers typically summarize a data set by using the mean and standard deviation. Based on the analysis of the data, researchers draw conclusions about the claim.

However, rather than developing research questions and using data to prove or disprove a claim, researchers can view a variety of characteristics of the data set to develop a hypothesis.

Various attributes about the data set can be viewed by using **Exploratory Data Analysis (EDA)**. By using EDA, researchers can view the center, spread and possible extreme observations, which are known as outliers.

In EDA, researchers use the median to describe the center, the interquartile range to describe the spread and a boxplot to provide a graphical representation of the data.

DEFINITION

Boxplot: **a graphical representation of data using the five-number summary.**

Five-Number Summary: **a summary that consists of the minimum value, the first quartile (Q_1), the median (Med), the third quartile (Q_3) and the maximum value.**

To construct a boxplot, perform the following steps:

Constructing a Boxplot

Step 1: Denote the minimum and maximum values as dots.
Step 2: Draw a box using Q_1 and Q_3 as the ends.
Step 3: Draw a vertical line inside the box at the median.
Step 4: Draw a horizontal line from the minimum value to Q_1 and a horizontal line from Q_3 to the maximum value.
Step 5: Denote any outliers with an asterisk.

By viewing a boxplot, researchers can also determine the shape of the distribution.

Boxplots and Shapes of Distributions

Symmetrical:

Skewed right:

Skewed left:

In Step 5 of constructing a boxplot, outliers are mentioned.

Outliers, extremely large or small observations, can be determined mathematically by creating upper and lower fences.

Upper fence: $Q_3 + 1.5 \times IQR$

Lower fence: $Q_1 - 1.5 \times IQR$

where, **IQR (Interquartile Range)** $= Q_3 - Q_1$

An observation is considered an outlier if the value of the observation is greater than the upper fence or less than the lower fence.

Outliers can affect summary statistics such as the mean and standard deviation because they fall outside the overall pattern of the distribution. Let's view an example.

On April 18, 2009, a search on Orbitz.com showed the following prices for one-way flights with 1-stop from Tampa International Airport to Atlanta International Airport. The following table shows the airlines along with their corresponding price:

Airline	Price ($)
Delta Airlines	428
US Airways	323
AirTran Airways	353
Continental Airlines	509
American Airlines	401
United Airlines	424
Northwest Airlines	988

Source: Orbitz.com

Let's determine if any of the observations are considered outliers. To do so, let's compute the following:

Upper fence: $Q_3 + 1.5 \times IQR$

Lower fence: $Q_1 - 1.5 \times IQR$

Using the techniques learned in Section 4.3: Measures of Position, $Q_1 = 353$ and $Q_3 = 509$. The IQR = 156.

Therefore,

Upper fence: $Q_3 + 1.5 \times IQR$

$$509 + 1.5 \times 156 = 743$$

Lower fence: $Q_1 - 1.5 \times IQR$

$$353 - 1.5 \times 156 = 119$$

Next, search the data set to determine if there are any values greater than the upper fence (743) or less than the lower fence (119).

The price of $988 is considered an outlier because the value is greater than 743. Now, let's examine how this affects the mean and median.

Using the techniques learned in Section 4.1: Measures of Central Tendency, the mean = 489.4 and the median is 424. Thus, the distribution is skewed right.

Suppose the outlier's value changed from 988 to 588. Let's re-compute the mean and median. The new mean = 432.3 and the new median = 424.

The median was not affected by changing the extreme value because the median examines the middle of the data set. However, the mean was affected.

Summary statistics that are not affected by outliers are said to be **resistant**. Summary statistics that are affected by outliers are said to be **nonresistant**.

The median and interquartile range are resistant measures. The mean and standard deviation are nonresistant measures.

Therefore, it's best to report the median and interquartile range if the distribution is skewed. Otherwise, if the distribution is symmetrical, it's best to report the mean and standard deviation.