# Summarizing Data

## OUTLINE

1. Measures of Central Tendency
2. Measures of Dispersion
3. Measures of Central Tendency and Dispersion from Grouped Data
4. Measures of Position
5. The Five-Number Summary and Boxplots

## Putting It Together

When we look at a distribution of data, we should consider three characteristics of the distribution: shape, center, and spread. In the last chapter, we discussed methods for organizing raw data into tables and graphs. These graphs (such as the histogram) allow us to identify the shape of the distribution: symmetric (in particular, bell shaped or uniform), skewed right, or skewed left.

The center and spread are numerical summaries of the data. The center of a data set is commonly called the average. There are many ways to describe the average value of a distribution. In addition, there are many ways to measure the spread of a distribution. The most appropriate measure of center and spread depends on the distribution’s shape.

Once these three characteristics of the distribution are known, we can analyze the data for interesting features, including unusual data values, called outliers.

## Section 3.1 Measures of Central Tendency

### Objectives

1. Determine the Arithmetic Mean of a Variable from Raw Data
2. Determine the Median of a Variable from Raw Data
3. Explain What It Means for a Statistic to be Resistant
4. Determine the Mode of a Variable from Raw Data

#### Objective 1: Determine the Arithmetic Mean of a Variable from Raw Data

Introduction, Page 1

*Answer the following after watching the video.*

00:01>> When we look at a distribution of data,

What does a measure of central tendency describe?

we should consider three characteristics

of the distribution--

shape, center, and spread.

In the last chapter, we discussed methods

for organizing raw data into tables and graphs.

These graphs, such as the histogram,

allow us to identify the shape of the distribution.

Distribution shapes can be described as symmetric, such

as you see in the top left and top right graph.

In the top left, we have a uniform distribution.

On the top right, we have a bell-shaped distribution,

or skewed right, or skewed left.

The center and spread are numerical summaries

of the data.

The center of a data set is commonly called the average.

There are many ways to describe the average value

of a distribution.

In addition, there are many ways to measure

the spread of a distribution.

The most appropriate measure of center and spread

depends on the distribution's shape.

Once these three characteristics of the distribution are known,

that is, the shape, center, and spread,

we can analyze the data for interesting features,

including unusual data values called outliers.

In Section 3.1, we focus on measures of central tendency.

A measure of central tendency numerically

describes the average or typical data value.

We hear the word average in the news all the time.

The average miles per gallon for the 2017 Chevrolet Corvette

in highway driving is 22.

According to the US Census Bureau,

the national average commute time to work in 2015

was 25.9 minutes.

According to the US Census Bureau,

the average household income in 2015 was $56,515.

In this section, we discussed the three most widely used

measures of central tendency--

the mean, the median, and the mode.

As we shall see, these three measures of central tendency

can often give very different results.

Objective 1, Page 1

1. Explain how to compute the arithmetic mean of a variable.

The **arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.

1. What symbols are used to represent the population mean and the sample mean? The **population arithmetic mean**, μ (pronounced "mew"), is a [parameter](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_4cf4bb98-c032-71f2-f761-2941bca8c0c2) that is computed using data from all the individuals in a population. The **sample arithmetic mean,**  x¯¯¯ (pronounced "x-bar"), is a [statistic](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_776c87f7-55b1-14f7-b297-a88d1c052694) that is computed using data from individuals in a sample.

Objective 1, Page 2

List the formulas used to compute the population mean and the sample mean.

**Population Mean**

If x1, x2,…, xN are the N observations of a variable from a population, then the population mean, μ (pronounced "mew"), is

μ=x1+x2+⋅⋅⋅+xN/N=∑xi/N

**Sample Mean**

If x1, x2,…, xn are n observations of a variable from a sample, then the sample mean, is

x¯=x1+x2+⋅⋅⋅+xn/n=∑xi/n

**Note:** Throughout this course, we agree to round the mean to one more decimal place than that in the raw data.

Objective 1, Page 3

**Example 1 *Computing a Population Mean and a Sample Mean***

Table 1 shows the first exam scores of the ten students enrolled in Introductory Statistics.

**Table 1**

| **Student** | **Score** |
| --- | --- |
| 1.Michelle | 82 |
| 2. Ryanne | 77 |
| 3. Bilal | 90 |
| 4. Pam | 71 |
| 5. Jennifer | 62 |
| 6. Dave | 68 |
| 7. Joel | 74 |
| 8. Sam | 84 |
| 9. Justine | 94 |
| 10. Juan | 88 |

**Finding the Mean**

enter the raw data into the spreadsheet. Name the column variable.

Select **Stat**, highlight **Summary Stats**, and select **Columns**.

Click on the variable you want to summarize. If you wish to compute certain statistics, hold down the Control (Ctrl) key when selecting the statistic (or Command on an Apple). Click Compute!.

Now we're going to take a sample of size 4.

Press Data, Sample.

Select both columns, student and score.

Type 4 for the sample size.

And under sampling options, select the middle option,

Sample All Columns at One Time.

Press Compute.

And we have our sample of 4 students, Juan, Pam, Bilal,

and Dave.

And their scores are listed next to their names.

Finally, for part C, we're going to find the sample mean.

Just like we did for the first part of the problem,

Stat, Summary Stats, Columns.

Select Sample Score and Mean.

Click Compute.

And we have our sample mean of 79.25.

Now, the sample and sample mean are going to vary.

And yours won't necessarily match this one.

Now we'll go over the StatCrunch steps.

For part A, press Stat.

And from the Summary Stats menu, select Columns.

Select the column containing the data.

Select Mean and click Compute.

For part B, press the Data button and select Sample.

Select the columns you'll be sampling from.

Enter a sample size, and check Sample All Columns at One Time.

Click Compute, and you'll have your sample in another column.

For part C, it's just like part A. Press the Stat button.

And from Summary Stats, select Columns.

Select the column containing the data.

Select Mean and click Compute.

Compute the population mean, .

1. Find a simple random sample of size *n* = 4 students.
2. Compute the sample mean, , of the sample found in part (B).

Objective 1, Page 5

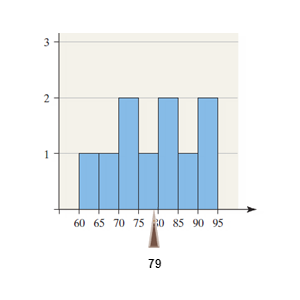
*Answer the following after experimenting with the fulcrum animation*.

1. What is the mean of the data? 79

#### 4. Explain why it is helpful to think of the mean as the center of gravity.

#### Visualizing the Mean as the Center of Gravity: An Animation

It helps to think of the mean as the center of gravity. In other words, the mean is the value such that a histogram of the data is perfectly balanced, with equal weight on each side of the mean. Figure 2 shows a histogram of the data in [Table 1](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_5_8b9dbe3e-3d8d-23e3-abca-6d3ec060e452). Recall, the mean of the data in Table 1 is 79 points. Play with the fulcrum (triangle) to verify that the mean is the balancing point of the data.

**Figure 2**

#### Objective 2: Determine the Median of a Variable from Raw Data

Objective 2, Page 1

5 ) Define the median of a variable.

second measure of central tendency is the *median*. To compute the median of a set of data, the data must be [quantitative](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_1_b4c1bd21-c4fc-66fc-4c6a-2bd3f220d6c6).

**DEFINITION**

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

The next slide shows the steps for finding the median, M, of a data set by hand. It is important to understand how to find the median by hand, before using technology.

1. Objective 2, Page 2

List the three steps in finding the median of a data set.

Step 1. Arrange the data in ascending order.

Step 2. Determine the number of observations, n.

Step 3. Determine the observation in the middle of the data set.

* If the number of observations is odd, then the median is the data value exactly in the middle of the data set. That is, the median is the observation that lies in the  n+12 position.
* If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the n2 position and the n2+1 position.

Objective 2, Page 3

**Example 2 *Determining the Median of a Data Set (Odd Number of Observations)***

Table 2 shows the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

**Table 2**

| **Song Name** | **Length** |
| --- | --- |
| "Sister Golden Hair" | 201 |
| "Black Water" | 257 |
| "Free Bird" | 284 |
| "The Hustle" | 208 |
| "Southern Nights" | 179 |
| "Stayin' Alive" | 222 |
| "We Are Family" | 217 |
| "Heart of Glass" | 206 |
| "My Sharona" | 240 |

Problem

Table 2 shows the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

| **TABLE 2** | |
| --- | --- |
| **Song Name** | **Length** |
| "Sister Golden Hair" | 201 |
| "Black Water" | 257 |
| "Free Bird" | 284 |
| "The Hustle" | 208 |
| "Southern Nights" | 179 |
| "Stayin' Alive" | 222 |
| "We Are Family" | 217 |
| "Heart of Glass" | 206 |
| "My Sharona" | 240 |

**Video Solution**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

**Technology Step-By-Step**

[](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_3_311f2a60-8477-2cbb-b119-452c711b5764)

Approach

Follow the [steps](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_3_3f9f61bf-071b-ebd7-b804-fec6b4a5a259) for finding the median, M.

Solution

**Step 1.**Arrange the data in ascending order:

179,201,206,208,217,222,240,257,284

There are n=9 observations.

Because n is odd, the median is the observation exactly in the middle of the data set with the data written in ascending order. This value lies in the n+12=9+12=5th position. The median appears in blue with four observations on each side of the median.

So M=217.

179,201,206,208,217,222,240,257,284

In this example, we'll learn to determine the median when there

are an odd number of observations, using StatCrunch.

The data in the following table represent the length

in seconds of a random sample of songs released in the 1970s.

Find the median length of the songs.

Here are the data.

Let's open StatCrunch.

I've already typed the data in the first column

under Song Length.

To find the median, we press Stat, Summary Stats, Columns.

Select the column containing the data.

And under Statistics, we want to click on Median.

Click Compute.

And our median song length is 217 seconds.

Now it's time to go over the steps using StatCrunch.

Press the Stat button.

And from Summary Stats, select Columns.

Select the column containing the data.

Select Median.

And click Compute.

And you'll have your result.

Objective 2, Page 5

**Example 3 *Determining the Median of a Data Set (Even Number of Observations)***

Find the median score of the data in Table 1.

**Table 1**

| **Student** | **Score** |
| --- | --- |
| 1. Michelle | 82 |
| 2. Ryanne | 77 |
| 3. Bilal | 90 |
| 4. Pam | 71 |
| 5. Jennifer | 62 |
| 6. Dave | 68 |
| 7. Joel | 74 |
| 8. Sam | 84 |
| 9. Justine | 94 |
| 10. Juan | 88 |

Problem

Find the median score of the data in [Table 1](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_5_78d10160-e329-54d0-9c7c-55a334769f7d).

**Video Solution**

|  |
| --- |
|  |

Approach

Follow the [steps](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_5_0ee11d17-d3c2-1c19-10ab-aa3e6dfa10a9) for finding the median, M.

Solution

**Step 1** Arrange the data in ascending order:

62,68,71,74,77,82,84,88,90,94

There are n=10 observations.

Because nis even, the median is the mean of the two middle observations, the fifth (n2=102=5) and sixth (n2+1=102+1=6) observations with the data written in ascending order. So the median is the mean of 77 and 82:

M=77+822=79.5

62,68,71,74,77,82,84,88,90,94

M =79.5

Notice that there are five observations on each side of the median. We conclude that 50% (or half) of the students scored less than 79.5 and 50% (or half) scored above 79.5

#### Objective 3: Explain What It Means for a Statistic to be Resistant

Objective 3, Page 1

*Answer the following as you work through the Mean versus Median Applet.*

1. When the mean and median are approximately 2, how does adding a single observation near 9 affect the mean? How does it affect the median? AGAIN It radically effects the mean, barely the median at all.
2. When the mean and median are approximately 2, how does adding a single observation near 24 affect the mean? The median? AGAIN It radically effects the mean, littler effect on the median at all.

Objective 3, Page 1 (continued)

1. When the mean and median are approximately 40, how does dragging the new observation from 35 toward 0 affect the mean? How does it affect the median?

Objective 3, Page 2

*Answer the following as you watch the video.*

1. Which measure, the mean or the median, is least affected by extreme observations? Median
2. Define what it means for a numerical summary of data to be resistant. Well, a numerical summary of data

is said to be resistant if values that are extreme--

very large or very small--

relative to the data, do not affect its values

substantially.

1. Which measure, the mean or the median, is resistant? MEDIAN

So far, we've looked at two measures of central tendency--

the mean and the median.

We're going to use a statistical applet

to see the relationship that exists between these two

measures.

The data that I have on the screen

represents the sale price of a random sample of 12 homes sold

in the Seattle area.

What I'm going to do is add a point to the data set.

Let's say, some executive from Amazon

buys a home that costs $4 million.

Take note that the mean currently, is about $422,000,

and the median is about $410,000.

So I'm going to click Add Point and add our $4 million home.

I click OK, and notice what happens.

The median is roughly the same value

as it was before that data point was included,

but the mean jumped up to $697,513.

So this one value, the $4 million home,

had a significant impact on the value of the mean,

but really had no impact whatsoever on the median.

So what we say, based on this result,

is that the mean is not resistant,

but the median is resistant.

What do I mean when I say resistant?

Well, a numerical summary of data

is said to be resistant if values that are extreme--

very large or very small--

relative to the data, do not affect its values

substantially.

Objective 3, Page 3

State the reason that we compute the mean. So the median is resistant, but the mean is not resistant.

You may be asking yourself, "Why would I ever compute the mean?" After all, the mean and median are close in value for symmetric data, and the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of statistical inference is based on the mean.

Objective 3, Page 7

*Answer the following as you work through Activity 2: Relation among the Mean, Median, and Distribution Shape.*

1. If a distribution is skewed left, what is the relation between the mean and median?The MEDIAN is GREATER than THE MEAN
2. If a distribution is skewed right, what is the relation between the mean and median? The MEDIAN is LESSER than THE MEAN

Objective 3, Page 7 (continued)

1. If a distribution is symmetric, what is the relation between the mean and median? EQUAL

Objective 3, Page 11

1. Sketch three graphs showing the relation between the mean and median for distributions that are skewed left, symmetric, and skewed right.

Objective 3, Page 12

**Example 4 *Describing the Shape of a Distribution***

The data in Table 4 represent the birth weights (in pounds) of 50 randomly sampled babies.

1. Find the mean and median birth weight. Mean = 7.5 Median = 7.35
2. Describe the shape of the distribution. Skewed right
3. Which measure of central tendency best describes the average birth weight? Mean

In this example, we'll learn to describe

the shape of the distribution using StatCrunch.

The following data represent the birth weights

in pounds of 50 randomly sampled babies.

Part A, find the mean and the median.

Part B, describe the shape of the distribution.

And Part C, which measure of central tendency

better describes the average birth weight?

Here are the data.

Now let's open StatCrunch to find the mean and the median.

I've already typed the data into the first column

under birth weight.

I want to begin by finding the mean and the median for part A.

Press Stat, Summary Stats, Columns.

Select the column that contains the data.

And for statistics, we'll click On mean and Median.

Click Compute, and we have our results.

The mean is 7.49 pounds, and the median is 7.35 pounds.

So we found that the mean was 7.49 and the median was 7.35.

Now we'll open StatCrunch to create a histogram.

Now, to describe the shape of the distribution,

we want to take a look at a histogram.

To create that, under Graph, Histogram.

Select the column that contains the data.

And I'm going to let StatCrunch do it by its defaults.

Click Compute.

And there we see a histogram of the birth weights.

Looking at the shape that we have here,

this is roughly bell shaped.

So here is a copy of the histogram.

And we can tell that this distribution

is roughly bell shaped.

Because the mean and median are close to each other in value,

**we use the mean as the measure of central tendency.**

**Table 4**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5.8 | 7.4 | 9.2 | 7 | 8.5 | 7.6 |
| 7.9 | 7.8 | 7.9 | 7.7 | 9 | 7.1 |
| 8.7 | 7.2 | 6.1 | 7.2 | 7.1 | 7.2 |
| 7.9 | 5.9 | 7 | 7.8 | 7.2 | 7.5 |
| 7.3 | 6.4 | 7.4 | 8.2 | 9.1 | 7.3 |
| 9.4 | 6.8 | 7 | 8.1 | 8 | 7.5 |
| 7.3 | 6.9 | 6.9 | 6.4 | 7.8 | 8.7 |
| 7.1 | 7 | 7 | 7.4 | 8.2 | 7.2 |
| 7.6 | 6.7 |  |  |  |  |

#### Objective 4: Determine the Mode of a Variable from Raw Data

hird measure of central tendency is the mode, which can be computed for either [quantitative](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj5_1_d5ede0e7-578b-7d34-99e8-0d0b1759e021) or [qualitative](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj5_1_34db8c89-ea6e-f431-6eac-9209a8ee146a) data.

**DEFINITION**

The **mode** of a variable is the observation of the variable that occurs most frequently in the data set.

* To compute the mode, tally the number of observations that occur for each data value.
* The data value that occurs most often is the mode.
* If no observation occurs more than once, we say that the data have **no mode**.
* A set of data can have no mode, one mode, or more than one mode.

Objective 4, Page 1

13)

Define the mode of a variable. The **mode** of a variable is the observation of the variable that occurs most frequently in the data set.

14)Under what conditions will a set of data have no mode? When each data value occurs only once.

15)Under what conditions will a set of data have two modes? When they have two data values that are equal and occur more than once.

Objective 4, Page 2

**Example 5 *Finding the Mode of Quantitative Data***

The following data represent the number of O-ring failures on the shuttle Columbia for the 17 flights prior to its fatal flight:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3

Find the mode number of O-ring failures. 11

Objective 4, Page 3

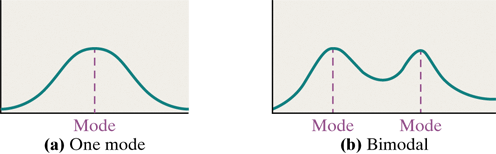
**Example 6 *Finding the Mode of Quantitative Data***

Find the mode of the exam score data listed in Table 1. None

**Table 1**

| **Student** | **Score** |
| --- | --- |
| 1. Michelle | 82 |
| 2. Ryanne | 77 |
| 3. Bilal | 90 |
| 4. Pam | 71 |
| 5. Jennifer | 62 |
| 6. Dave | 68 |
| 7. Joel | 74 |
| 8. Sam | 84 |
| 9. Justine | 94 |
| 10. Juan | 88 |

Objective 4, Page 5

1. What does it mean when we say that a data set is bimodal? Multimodal? 

Objective 4, Page 6

**Example 7 *Finding the Mode of Qualitative Data***

The data in Table 5 represent the location of injuries that required rehabilitation by a physical therapist. Determine the mode location of injury.

**Table 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Back | Back | Hand | Neck | Knee | Knee |
| Wrist | Back | Groin | Shoulder | Shoulder | Back |
| Elbow | Back | Back | Back | Back | Back |
| Back | Shoulder | Shoulder | Knee | Knee | Back |
| Hip | Knee | Hip | Hand | Back | Wrist |

Data from Krystal Catton, student at Joliet Junior College

Objective 4, Page 8

#### ***Summa***ry

1. List the conditions for determining when to use the following measures of central tendency.

|  |  |
| --- | --- |
| **Type of Variable** | **Best measure of central tendency** |
| Nominal | Mode |
| Ordinal | Median |
| Interval/Ratio (not skewed) | Mean |
| Interval/Ratio (skewed) | Median |

## Mean When is the mean the best measure of central tendency?

The mean is usually the best measure of central tendency to use when your data distribution is [continuous](https://statistics.laerd.com/statistical-guides/types-of-variable.php) and symmetrical, such as when your data is normally distributed. However, it all depends on what you are trying to show from your data.

When is the mode the best measure of central tendency?

The mode is the least used of the measures of central tendency and can only be used when dealing with [nominal](https://statistics.laerd.com/statistical-guides/types-of-variable.php) data. For this reason, the mode will be the best measure of central tendency (as it is the only one appropriate to use) when dealing with nominal data. The mean and/or median are usually preferred when dealing with all other types of data, but this does not mean it is never used with these data types.

When is the median the best measure of central tendency?

The median is usually preferred to other measures of central tendency when your data set is skewed (i.e., forms a skewed distribution) or you are dealing with ordinal data. However, the mode can also be appropriate in these situations, but is not as commonly used as the median.

## Section 3.2 Measures of Dispersion

### Objectives

1. Determine the Range of a Variable from Raw Data
2. Determine the Standard Deviation of a Variable from Raw Data
3. Determine the Variance of a Variable from Raw Data
4. Use the Empirical Rule to Describe Data That Are Bell-Shaped

Introduction, Page 1

Measures of central tendency describe the typical value of a variable. We also want to know the amount of dispersion (or spread) in the variable. Dispersion is the degree to which the data are spread out.

Introduction, Page 2

**Example 1 *Comparing Two Sets of Data***

The data tables represent the IQ scores of a random sample of 100 students from two different universities.

For each university, compute the mean IQ score and draw a histogram, using a lower class limit of 55 for the first class and a class width of 15. Comment on the results.

#### Objective 1: Determine the Range of a Variable from Raw Data

Objective 1, Page 1

What is the range of a variable? The **range**, R, of a variable is the difference between the largest and smallest data value. That is,

Range=R=largest data value−smallest data value

Objective 1, Page 2

**Example 2 *Computing the Range of a Set of Data***

The data in the table represent the first exam scores of 10 students enrolled in Introductory Statistics. Compute the range.

| **Student** | **Score** |
| --- | --- |
| 1. Michelle | 82 |
| 2. Ryanne | 77 |
| 3. Bilal | 90 |
| 4. Pam | 71 |
| 5. Jennifer | 62 |
| 6. Dave | 68 |
| 7. Joel | 74 |
| 8. Sam | 84 |
| 9. Justine | 94 |
| 10. Juan | 88 |

#### Objective 2: Determine the Standard Deviation of a Variable from Raw Data

**DEFINITION**

The **population standard deviation** of a variable is the [square root](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_d64832e2-0d45-daf3-528d-a32180540816) of the sum of squared deviations about the population mean divided by the number of observations in the population, N.

In other words, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma). The formula is given as

σ=(x1−μ)2+(x2−μ)2+⋯+(xN−μ)2N−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−√=∑(xi−μ)2N−−−−−−−−−−√

where x1, x2, … , xN are the N observations in the population and μ is the population mean.

Objective 2, Page 1

1. Explain how to compute the population standard deviation  and list its formula.

Objective 2, Page 2

**Example 3 *Computing a Population Standard Deviation***

Compute the population standard deviation of the test scores in Table 6.

**Table 6**

| **Student** | **Score** |
| --- | --- |
| 1. Michelle | 82 |
| 2. Ryanne | 77 |
| 3. Bilal | 90 |
| 4. Pam | 71 |
| 5. Jennifer | 62 |
| 6. Dave | 68 |
| 7. Joel | 74 |
| 8. Sam | 84 |
| 9. Justine | 94 |
| 10. Juan | 88 |

olution

**Step 1** See Table 7. Column 1 lists the observations in the data set, and Column 2 lists the population mean.

| **TABLE 7** | |
| --- | --- |
| **Score,**xi | **Population Mean,**μ |
| 82 | 79 |
| 77 | 79 |
| 90 | 79 |
| 71 | 79 |
| 62 | 79 |
| 68 | 79 |
| 74 | 79 |
| 84 | 79 |
| 94 | 79 |
| 88 | 79 |

Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Michelle is 82−79=3. It is a good idea to add the entries in this column to make sure they sum to 0.

| **TABLE 7** | | |
| --- | --- | --- |
| **Score,**xi | **Population Mean,**μ | **Deviation about the Mean,**xi−μ |
| 82 | 79 | 82−79=3 |
| 77 | 79 | 77−79=−2 |
| 90 | 79 | 11 |
| 71 | 79 | −8 |
| 62 | 79 | −17 |
| 68 | 79 | −11 |
| 74 | 79 | −5 |
| 84 | 79 | 5 |
| 94 | 79 | 15 |
| 88 | 79 | 9 |
|  |  | ∑(xi−μ)=0 |

Column 4 shows the squared deviations about the mean.

| **TABLE 7** | | | |
| --- | --- | --- | --- |
| **Score,**xi | **Population Mean,**μ | **Deviation about the Mean,**xi−μ | **Squared Deviation about the Mean,**(xi−μ)2 |
| 82 | 79 | 82−79=3 | 32=9 |
| 77 | 79 | 77−79=−2 | (−2)2=4 |
| 90 | 79 | 11 | 121 |
| 71 | 79 | −8 | 64 |
| 62 | 79 | −17 | 289 |
| 68 | 79 | −11 | 121 |
| 74 | 79 | −5 | 25 |
| 84 | 79 | 5 | 25 |
| 94 | 79 | 15 | 225 |
| 88 | 79 | 9 | 81 |
|  |  | ∑(xi−μ)=0 | ∑(xi−μ)2=964 |

**Step 4.**Sum the entries in Column 4 to obtain the [numerator](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_92146fab-0d9e-4a5e-e8b8-2108c86304af) of the [formula](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_46f0931f-0f9a-acbb-9f91-1e0b8d637078) for the population standard deviation. Divide this sum by the number of students, 10:

∑(xi−μ)2N=96410=96.4points2

The square root of the result in Step 4 is the population standard deviation.

σ=∑(xi−μ)2N−−−−−−−−−−−√=96.4 points2−−−−−−−−−−√≈9.8 points

StatCrunch

**Finding the Standard Deviation**

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Stat**, highlight **Summary Stats**, and select **Columns**.
3. Click on the variable you want to summarize. Note that Unadj. std. dev. is the population standard deviation. If you want to compute certain statistics, hold down the Control (Ctrl) key (or Command on an Apple) when selecting the statistic. Click Compute!.

Objective 2, Page 5

If a data set has many values that are “far” from the mean, how is the standard deviation affected? Look at [Table 7](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_5_641474b7-8d28-ea05-b035-21f6d6810208). The further an observation is from the mean, 79, the larger the squared deviation. For example, because the second observation, 77, is not “far” from 79, the squared deviation, 4, is not large. However, the fifth observation, 62, is further from 79, so the squared deviation, 289, is much larger.

If a data set has many observations that are “far” from the mean, then the sum of the squared deviations will be large and the standard deviation will be large.

Objective 2, Page 6

Explain how to compute the sample standard deviation *s* and list its formula. Here is the formula for the sample standard deviation.

**EFINITION**

The **sample standard deviation**, s, of a variable is the square root of the sum of squared deviations about the sample mean divided by n−1, where n is the sample size. The formula is given as

s=(x1−x¯¯¯)2+(x2−x¯¯¯)2+⋯+(xn−x¯¯¯)2n−1−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−√=∑(xi−x¯¯¯)2n−1−−−−−−−−−−√

where x1,x2,…,xn are the n observations in the sample and x¯ is the sample mean.

3.2 Measures of Dispersion

**Why Do We Divide by n – 1 ? Understanding Degrees of Freedom**

To find the sample standard deviation, we divide by n- 1   Showing why we divide by N – 1  is beyond the scope of the course. However, the following explanation has intuitive appeal. We already know that the sum of the deviation about the mean, must equal zero. Therefore, if the sample mean is known and the first n – 1 observations are known, then the observation must be the value that causes the sum of the deviations to equal zero. For example, suppose x¯=4 is based on a sample of size n=3. If x1=2 and x2=3, then we can determine x3 as follows:

X1 + x2 +x3 + x / n

2 + 3 + x3 / 3 = 4

5 = x3/ 3 = 4

5 = x3 =12

X3=7

We call N – 1 the **degrees of freedom** because the first N - 1 observations have freedom to be any value, but the *nth* observation has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

[](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_7_6cddcbc0-3db8-c3e0-dbe2-7cec762bcd6e)

3.2 Measures of Dispersion

EXAMPLE 4 Computing a Sample Standard Deviation

In this example, we'll learn to compute a sample standard

deviation using StatCrunch.

The data in the table represent the first exam scores

of 10 students enrolled in Introductory Statistics.

Find the sample standard deviation of the sample

from example 1b in the last section.

Let's go back to StatCrunch.

OK, here's the sample that we took

in example 1 of the previous section.

And now we're going to find the sample standard

deviation for this sample of size 4.

Press Stat, Summary Stats, Columns.

Select the column containing the sample data.

And we want to click on Standard Deviation and Compute.

OK, so there's our sample standard deviation,

and that is 11.4 rounded to 1 decimal place.

Now we'll go over the StatCrunch steps.

Press the Stat button.

And from Summary Stats, select Columns.

Select the column containing the data.

Select Standard Deviation and click Compute,

and you'll have your result.

**The Standard Deviation and Resistance**

Watch the following video, which discusses the resistance of standard deviation.

In this example, we'll see that standard deviation is not

resistant.

The data in the table represent the first exam Scores

In this example, we'll see that standard deviation is not

of 10 students enrolled in introductory statistics.

A random sample of 4 students-- Sam, Pam, Bilal, and Michelle--

is selected.

If Michelle score is incorrectly recorded as 28 instead of 82,

find the standard deviation for the sample.

In other words if this had been typed as 28 instead of 82,

let's see what effect that has on standard deviation.

We're going to go ahead and do this using StatCrunch.

So here's the sample of Sam, Pam, Bilal, and Michelle,

along with their correct scores and the score with Michelle's

82 typed as a 28 by mistake.

And I'm going to find the sample standard deviation of these 2

columns.

Stat, Summary Stats, Columns.

I'm going to select Sample Score and Score With Typo.

This is a sample standard deviation.

And click Compute.

And we see that the correct standard deviation was 7.9.

But with this typo, it jumped up to 28.0.

So the sample standard deviation if Michelle score

is incorrectly entered as 28 instead of 82 is 28.0 points.

The sample standard deviation using the correct value

is 7.9 points.

Here we see that 1 extreme value has

a huge impact on the value of the standard deviation.

That tells us that this measure is not resistantant

**Interpretations of the Standard Deviation**

How does the value of the standard deviation relate to the spread of the distribution?

Standard deviation represents the "typical" deviation from the mean. As such, the standard deviation may be used to judge whether a particular observation is "far away" from the mean of a data set. For example, is a measure of **31** cm far from **25** cm? It depends. If the standard deviation of the data is **6**cm, then the answer is no because **31** cm would be only  **1** standard deviation from **25** cm. However, if the standard deviation of the data is  **2** cm, then the answer is yes because **31** cm would be standard **3** deviations from **25** cm. A good rule of thumb is to consider an observation "far away" if it is more than **2** standard deviations from the other observation (such as the mean).

So, when judging the unusualness of an observation, it is vital that you consider the underlying variation in the data as measured by the standard deviation.

When comparing two populations, the larger the standard deviation, the greater the dispersion, or spread, of the distribution provided the variable of interest from the two populations has the same unit of measure. The units of measure must be the same so that we are comparing "apples with apples." For example**,$100**  is not the same as **100** Japanese yen (because recently **$1**, was equivalent to about **114**  yen). So a standard deviation of  **$100** is substantially higher than a standard deviation of **100** yen.

The standard deviation is used to describe the spread in symmetric distributions (while the mean is used to describe the center of the distribution).

EXAMPLE 5 Comparing the Standard Deviations of Two Data Sets

The following video compares the two standard deviations for the IQ data introduced at the beginning of this section.

The data in the following tables represent the IQ scores

of a random sample of 100 students from two

different universities.

We're going to use the standard deviation

to determine whether University A or University B has more

dispersion in the IQ scores of its students.

Here are the values for University A and the IQ

scores for University B.

Recall in an earlier example, we created histograms

for the same sets of data.

And the histogram for University A

was more dispersed-- more spread out-- than the histogram

for University B, which should suggest that University A has

greater dispersion.

Let's check the standard deviation

to see if that's true.

Here are the values for University A and B

in StatCrunch.

I'm going to use that to calculate

the standard deviation.

STAT, SUMMARY STATS, COLUMNS.

I'll select both columns.

Click on NEXT.

I'm going to unselect everything but standard deviation,

and click CALCULATE.

Here, we see the standard deviation for University A

is 16.1.

And for University B, it's 8.4.

Both rounded to the nearest tenth.

So again, we found that the standard deviation

for University A was 16.1 points.

And at University B, it was 8.4 points.

The IQs at University A have more

dispersion as the scores have a higher standard deviation.

**Activity 1 Standard Deviation as a Measure of Spread**

3.2 Measures of Dispersion

In the previous activity, you should have discovered some important relationships between data sets and their mean, median, and standard deviation. The video below uses a different applet to illustrate the results of the previous activity to reinforce these ideas.

In this activity, we're going to look at the standard deviation

as a measure of spread.

Change the lower limit to 0, and change the upper limit to 10.

Then click Update so that the scale on the number line

changes.

Now, create a data set of 5 observations

such that the mean and the median are roughly 5

and the standard deviation is roughly 1.

Because the mean and the median are the same value,

we know that our dot plot is going to need to be symmetric.

So I start clicking on the applet

with my mouse to add my 5 observations.

And you can see that my mean and median are too large

and my standard deviation is too small.

So I need to move some of my data values to the left

so that the mean and median decrease in value.

Now, my mean and median are getting closer to 5,

but my standard deviation is too small.

So I start creating additional spread in the data set.

So now you can see, I have a mean and a median that

are the same-- roughly 5-- and I have

a standard deviation that is 1.

Note that the data is symmetric, and also note

the spread in the data.

Because now what I want to do is create

a data set of 5 observations such that the mean

and the median are still 5, but the standard deviation

is roughly 2.

To accomplish this, I just create more spread

in the data-- being mindful that I

want to maintain my symmetry because the mean and the median

need to stay the same value.

So as I create more and more spread in the data,

you can see the standard deviation starts to increase.

And so now, the mean and the median are roughly 5,

and the standard deviation is roughly 2.

Note that we now have more spread in the dot plot

than we did when the standard deviation was 1.

So this is how the standard deviation measures spread.

The more spread there is in the distribution,

the higher the standard deviation is going to be.

Next, I want to create a data set of 6 observations--

so add 1 more value-- such that the standard deviation is 0.

Remember, standard deviation as a measure of spread.

Because the standard deviation we're looking for is 0,

it suggests that there will be no spread in the data set.

This means that all the observations

are going to be the same value.

So if I stack all these observations vertically

in my dot plot, I create a data set

where the standard deviation is 0.

So again, standard deviation of 0

suggests that there is no spread in the data set.

Put another way, all the observations are the same.

Now, I want to change the upper limit to 25.

Click Update, and my number line scale changes.

I'm going to create a data set with 6 observations between 0

and 5-- 1, 2, 3, 4, 5, 6-- such that the mean is roughly 3.

So I need to move some of these values

to the left to create a mean of roughly 3.

Note that the standard deviation is about 1.4.

Now, I'm going to add a seventh observation near 10,

and we note that the standard deviation increased

from 1.4 to 2.9.

Now, I grab that point near 10 and slide it to the right

until it's at 25.

Note that the standard deviation ballooned all the way up

to 8.3.

What does this suggest?

This tells us that a single observation can substantially

impact the value of the standard deviation,

and, therefore, the standard deviation

is not a resistant measure of spread.

#### Objective 3: Determine the Variance of a Variable from Raw Data

Objective 3, Page 1

1. Define variance.

Objective 3: Determine the Variance of a Variable from Raw Data

**OBJECTIVE 3** Determine the Variance of a Variable from Raw Data

Up to this point, we have studied two measures of dispersion—the [range](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_1_5440d234-f84a-2e70-a820-07b85e955f26) and the [standard deviation](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_1_eb702067-240b-d316-2683-4233104ad3c2). A third measure of dispersion is called the *variance*.

**DEFINITION**

The **variance** of a variable is the square of the standard deviation. The **population variance** is σ2, and the **sample variance** is s2.

The units of measure in variance are squared values. So if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult.

Objective 3, Page 2

**Example 6 *Determining the Variance of a Variable for a Population and a Sample***

In previous examples, we considered population data of exam scores in a statistics class. For this data, we computed a population mean of  points and a population standard deviation of  points. Then, we obtained a simple random sample of exam scores. For this data, we computed a sample mean of points and a sample standard deviation of *s* = 11.3points. Use the population standard deviation exam score and the sample standard deviation exam score to determine the population and sample variance of scores on the statistics exam.

Objective 3, Page 3

*Answer the following after you watch the video.*

1. Using a rounded value of the standard deviation to obtain the variance results in a round-off error. How should you deal with this issue? e population standard deviation is σ=9.8 points, so the population variance is σ2=(9.8 points)2=96.04 points2. The sample standard deviation is s=11.3 points, so the sample variance is s2=(11.3 points)2=127.7 points2.

We can compute variance in two ways.

One way is to compute the standard deviation first

and then square the result.

For this example, we found that the population

standard deviation was 9.8 points.

If we square that result, we get a variance

of 96.04 points squared.

The expression under the radical in the formula for standard

deviation is the formula for variance.

That is the formula for population variance.

And that is the formula for sample variance.

So if we work this problem by hand,

we would have that the standard deviation

was the square root of 96.4.

That value on the inside, that is our variance.

And if you think about the other way for calculating variance,

we got a result of 96.04, which is different than the result we

obtained here, which is exact.

Using a rounded value of the standard deviation

to obtain the variance results in a round-off error.

To deal with this issue, use as many decimal

places as possible when using standard deviation

to obtain the variance.

Objective 3, Page 5

#### Whenever a statistic consistently underestimates a parameter, it is said to be biased. To obtain an unbiased estimate of the population variance, divide the sum of the squared deviations about the sample mean by .Bias in the Variance and Standard Deviation

The sample variance is obtained using the formula s2=∑(xi−x¯¯¯)2n−1. What if we divided by n instead of n−1 to obtain the sample variance, as one might expect? Then the sample variance would consistently underestimate the population variance. Whenever a statistic consistently underestimates a parameter, it is said to be **biased**. To obtain an unbiased estimate of the population variance, divide the sum of the squared deviations about the sample mean by n−1.

To help understand the concept of a biased estimator, consider the following scenario. Suppose you work for a carnival, guessing people's ages. After 20 people have come to your booth, you notice that you have a tendency to underestimate peoples' ages, or guess too low. What could you do to correct this? You could adjust your guesses higher to avoid underestimating. In other words, originally your

#### Objective 4: Use the Empirical Rule to Describe Data That Are Bell-Shaped

Objective 4, Page 1

According to the Empirical Rule, if a distribution is roughly bell shaped, then approximately what percent of the data will lie within 1 standard deviation of the mean? What percent of the data will lie within 2 standard deviations of the mean? What percent of the data will lie within 3 standard deviations of the mean? Objective 4: Use the Empirical Rule to Describe Data That Are Bell-Shaped

3.2 Measures of Dispersion

OBJECTIVE 4 Use the Empirical Rule to Describe Data That Are Bell-Shaped

If data have a distribution that is bell-shaped, the *Empirical Rule* can be used to determine the percentage of data that will lie within k standard deviations of the mean.

**The Empirical Rule**

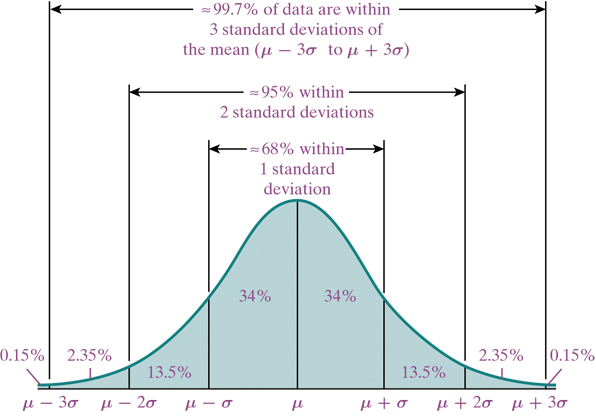
If a distribution is roughly bell shaped, then

* Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data will lie between μ−1σ and μ+1σ.
* Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data will lie between μ−2σ and μ+2σ.
* Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data will lie between μ−3σ and μ+3σ.

**NOTE**

The Empirical Rule can also be used based on sample data with x¯ in place of μ and s in place of σ.

Objective 4, Page 2

1. Sketch the third part of Figure 5. 

Objective 4, Page 3

The data in the following table represent the IQs

of a random sample of 100 students at a university.

Here they are.

And in this example, we'll be using the empirical rule.

I'll re-read these one at a time when I do each part.

To use the empirical rule, the data

must come from a distribution that

is approximately bell-shaped.

Here's the histogram for those IQ scores,

and we can see that it's roughly bell-shaped.

So we can use the empirical rule.

Part A. Determine the percentage of students

who have IQ scores within 3 standard deviations of the mean

according to the empirical rule.

Well, the empirical rule states that approximately 99.7%

of the data are within 3 standard deviations

of the mean, so that will hold true for our IQ scores as well.

Here's a graphic to help you remember this.

Within 3 standard deviations of the mean,

we'll find 99.7% of the data.

Part B. Determine the percentage of students

who have IQ scores between 67.8 and 132.2

according to the empirical rule.

We need to find out how far from the mean each of these values

are.

So 67.8, to determine that, we'll take 67.8,

subtract the mean, which is 100, and divide

by the standard deviation, which is 16.1.

When we work out this calculation,

we find that the result is negative 2.

That means that 67.8 is 2 standard deviations

below the mean.

For 132.2, we'll do the same thing.

Subtract the mean, that tells you how far away it is.

Divide by the standard deviation,

that tells you how many standard deviations away it is.

And this calculates out to be 2.

So what we're really looking for in this problem

is what percentage of the values are

within 2 standard deviations of the mean, either way.

Negative 2 to positive 2.

We know from the empirical rule that 95% of the data

are within 2 standard deviations of the mean.

So here, approximately 95% of all IQ scores

lie between 67.8 and 132.2.

Now we'll move on to determine the actual percentage

of students who fit in that range.

Here are the values again, and, in red,

I'm going to put all the values that

are between 67.8 and 132.2.

It's easier to count the ones that aren't.

1, 2, 3, 4 values are not in that range.

So that leaves 96 within that range.

So exactly 96% of all IQ scores lie between 67.8 and 132.2.

The final part, according to the empirical rule,

what percentage of students has an IQ score

between 116.1 and 148.3.

Let me mark up the picture for this one.

We know that the mean is at 100.

Adding 1 standard deviation-- and recall

that the standard deviation was 16.1.

At 1 standard deviation we're at 116.1.

Adding another 16.1 gets us to 132.2,

and another standard deviation, adding 16.1, is 148.3.

Now we're looking for the percentage

of values that are between these two values.

So we're going to add the area in this portion of the curve,

13.5%, to the area in this portion of the curve,

2.35%, and that leaves us with 15.85% of the students.

**Example 7 *Using the Empirical Rule***

Table 9 represents the IQs of a random sample of 100 students at a university.

1. Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.
2. Determine the percentage of students who have IQ scores between 67.8 and 132.2 according to the Empirical Rule.
3. Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.
4. According to the Empirical Rule, what percentage of students have IQ scores between 116.1 and 148.3?

**Table 9**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 73 | 103 | 91 | 93 | 136 | 108 | 92 | 104 | 90 | 78 |
| 108 | 93 | 91 | 78 | 81 | 130 | 82 | 86 | 111 | 93 |
| 102 | 111 | 125 | 107 | 80 | 90 | 122 | 101 | 82 | 115 |
| 103 | 110 | 84 | 115 | 85 | 83 | 131 | 90 | 103 | 106 |
| 71 | 69 | 97 | 130 | 91 | 62 | 85 | 94 | 110 | 85 |
| 102 | 109 | 105 | 97 | 104 | 94 | 92 | 83 | 94 | 114 |
| 107 | 94 | 1121 | 113 | 115 | 106 | 97 | 106 | 85 | 99 |
| 102 | 109 | 76 | 94 | 103 | 112 | 107 | 101 | 91 | 107 |
| 107 | 110 | 106 | 103 | 93 | 110 | 125 | 101 | 91 | 119 |
| 118 | 85 | 127 | 141 | 129 | 60 | 115 | 80 | 111 | 79 |

## Section 3.3 Measures of Central Tendency and Dispersion from Grouped Data

### Objectives

1. Approximate the Mean of a Variable from Grouped Data
2. Compute the Weighted Mean
3. Approximate the Standard Deviation from a Frequency Distribution

#### Objective 1: Approximate the Mean of a Variable from Grouped Data

In this example, we'll learn to construct frequency

and relative frequency distributions

from discrete data.

The manager of a Wendy's fast food restaurant

wants to know the typical number of customers

who arrive during the lunch hour.

The following data represent the number of customers

who arrive at Wendy's for 40 randomly

selected 15 minute intervals of time during lunch.

Construct a frequency and relative frequency

distribution.

Here are the data.

OK let's go ahead and open up StatCrunch.

I've typed the data in StatCrunch.

Let's begin.

We'll press Stat, Tables, Frequency.

Select the column containing the data, Wendy's.

And I can get both the frequency and relative frequency

distributions at the same time by having

these two highlighted.

For order by, we want to leave that value ascending.

Click Compute.

And there is our frequency distribution

and our relative frequency distribution.

Now, we'll go over the steps for using StatCrunch.

Type the values in one column.

Press the Stat button.

And from table select Frequency.

Select the column containing the data.

Select either frequency or relative frequency.

And click Compute.

And you'll have your results.

>> In this example, we'll learn how to organize continuous data

>> into a frequency and relative frequency distribution.

>> Suppose you were considering investing in a Roth IRA.

>> You collect the following data which

>> represent the five-year rate of return in percent, adjusted

>> for sales charges, for a simple random sample

>> of 40 large blended mutual funds.

>> Here are the data.

>> Construct a frequency and relative frequency

>> distribution.

>> So we look at the data.

>> And one thing we want to pick off

>> is the minimum and maximum values.

>> It appears that 3.22 is the smallest value we have.

>> And 12.03 is the largest.

>> So when we make our classes, we want

>> to make sure that 3.22 fits into the first class.

>> And the last class covers 12.03.

>> Now there's no single correct way to do this.

>> But one thing that we can do here

>> is start with a lower class limit of 3

>> and have a class width of 1 so that the upper class

>> limit will be 3.99.

>> So let the second class go from 4 to 4.99 and so on.

>> Again, the class width is the distance

>> between these two lower class limits, which is 1.

>> Now what this does-- let me finish

>> this off-- is it gives us 10 classes.

>> That's a decent number of classes.

>> If this breakdown doesn't feel right

>> once we see the frequencies, we could

>> go from 3 to 4.99, 5 to 6.99, and so on, we'll

>> have half as many classes that'll be twice as wide.

>> Now we need to tally up the occurrences in each class.

>> Going across the top row, we've got 3.27-- that's in the first

>> class-- 3.53, 3.45, 5.98, 4.55, 3.54, 4.91, and 4.75.

>> Now we continue on until we've reached the end of the table.

>> And then we'll have something that looks like this.

>> So now we'll count up the frequency.

>> Let's see.

>> 5, 10, 15, and 1 is 16 for the first class.

>> 5, 10, and 3 is 13 for the second class.

>> Third class only has 1, 2, 3, 4, and so on.

>> By the way, notice we had two empty classes.

We have to write the 0 there.

This is our frequency distribution.

To convert this to a relative frequency distribution,

we need to divide each frequency by the number of values that

are in our data set, which was 40.

So if I divide 16 by 40, I'll have the relative frequency

for the first class, which is 0.4.

If I divide 13 by 40, I'll have the relative frequency

for the second class, which is 0.325.

And then I keep going until I've completed

the relative frequency distribution.

OBJECTIVE 1 Approximate the Mean of a Variable from Grouped Data

Because raw data cannot be retrieved from a frequency table, we assume that within each class, the mean of the data values is equal to the [class midpoint](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_14d78d48-c14a-1aeb-2f0b-c5d2cf29ce62). Then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. Repeat the process for each class and add the results. This sum approximates the sum of all the data. The [video](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_0035d54b-3a16-d552-aa5c-90bff4440784) explains the formulas.

|  |  |  |
| --- | --- | --- |
| **Population Mean**  μ=∑xifi**/**∑fi =x1f1+x2f2+⋯+xnfn**/**f1+f2+⋯+fn | | **Sample Mean**  x¯¯¯=∑xifi∑fi=x1f1+x2f2+⋯+xnfn**/**f1+f2+⋯+fn |
| where: | xi is the midpoint or value of the ith class | |
|  | fi is the frequency of the ith class | |
|  | n is the number of classes | |

**DEFINITIONS** **Approximate the Mean of a Variable from a Frequency Distribution**

In each formula , x1 f1 approximates the sum of all the data values in the first class, x2 f2 approximates the sum of all the data values in the second class, and so on. Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.

In this example, we'll learn to approximate the mean

for continuous quantitative data from a frequency distribution

using StatCrunch.

The following frequency distribution

represents the five year rate of return

with a random sample of 40large blended mutual funds.

Approximate the mean five year rate of return.

Here are the data.

Let's go to StatCrunch.

I've already type the data into StatCrunch.

In the column labeled Class, I have the classes.

And I typed them in a particular way with the lower limit

space the word to another space and the upper limit.

The frequencies are in the second column.

To do the calculations, press Stat, Summary Stats,

Grouped/Binned Data.

The bins are the classes in the column labeled Class.

The counts are the frequencies.

For the midpoints, we're going to set that

as the average of consecutive lower limits.

Now, all I need to compute is the means.

So I'll select that.

Press Compute.

And there is our result, which rounds to be

13.1 to one decimal place.

So the approximate mean five year rate of return is 13.1%.

Let's go over the StatCrunch steps.

Enter the summarized data into the spreadsheet.

Press Stat.

And from Summary Stats, select Grouped/Binned data.

Choose the column that contains the class under the Bins

In dropdown menu.

Choose the column that contains the frequencies in the Counts

In dropdown menu.

Select the Consecutive Lower Limits radio button

for defining the midpoints.

Click Compute.

And you'll have your mean.

Result

The approximate mean five-year is 13.1%. The mean five-year rate of return from the [raw data](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_edad7c51-9e8c-8c4a-68ac-ccfd4d144d69) is 13.141%. Click the link to view a complete [by-hand solution](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_f710ae76-70e2-5475-c514-834adfab97a0).

**Approximating the Mean and Standard Deviation from Grouped Data**

1. If necessary, enter the summarized data into the spreadsheet. Name the columns.
2. Select **Stat**, highlight **Summary Stats**, and select **Grouped/Binned data**.
3. Choose the column that contains the class under the "Bins in:" drop-down menu. Choose the column that contains the frequencies in the "Counts in:" drop-down menu. Select the "Consecutive lower limits" radio button for defining the midpoints. Click Compute!.

1) Explain how to find the class midpoint.

* 1. List the formulas for approximating the population mean and sample mean from a frequency distribution.

Objective 1, Page 2

**Example 1 *Approximating the Mean for Continuous Quantitative Data from a Frequency Distribution***

The frequency distribution in Table 10 represents the five-year rate of return of a random sample of 40 large-blend mutual funds. Approximate the mean five-year rate of return.

**Table 10**

| **Class (5-year rate of return)** | **Frequency** |
| --- | --- |
| 8-8.99 | 2 |
| 9-9.99 | 2 |
| 10-10.99 | 4 |
| 11-11.99 | 1 |
| 12-12.99 | 6 |
| 13-13.99 | 13 |
| 14-14.99 | 7 |
| 15-15.99 | 3 |
| 16-16.99 | 1 |
| 17-17.99 | 0 |
| 18-18.99 | 0 |
| 19-19.99 | 1 |

Objective 1, Page 2 (Continued)

#### Objective 2: Compute the Weighted Mean

en data values have different importance, or *weights*, associated with them, we compute the *weighted mean*. For example, grade point average is a weighted mean, with weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value. The [video](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_1_b9a3be6d-dcbf-b2f5-731f-3e878e72a2f7) explains the formula for obtaining a weighted mean.

**DEFINITION**

The **weighted mean**, x¯w , of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

x¯¯¯w=Σwixi**/**Σwi = w1x1+w2x2+⋯+wnxn**/**w1+w2+⋯+wn

|  |  |
| --- | --- |
| where | wi is the weight of the ith observation  xi is the value of the ith observation |

Objective 2, Page 1

1. When data values have different importance, or weights, associated with them, we compute the weighted mean. Explain how to compute the weighted mean and list its formula.

Objective 2, Page 2

**Example 2 *Computing the Weighted Mean***

Marissa just completed her first semester in college. She earned an A in her 4-hour statistics course, a B in her 3-hour sociology course, an A in her 3-hour psychology course, a C in her 5-hour computer programming course, and an A in her 1-hour drama course. Determine Marissa’s grade point average.

#### Objective 3: Approximate the Standard Deviation from a Frequency Distribution

Let's go over the steps for using StatCrunch.

Enter the value of the variable in the first column

and the weights in the second column.

Press the Stat button, and from Summary Stats,

select Grouped Binned Data.

Choose the column that contains the values under the Bins

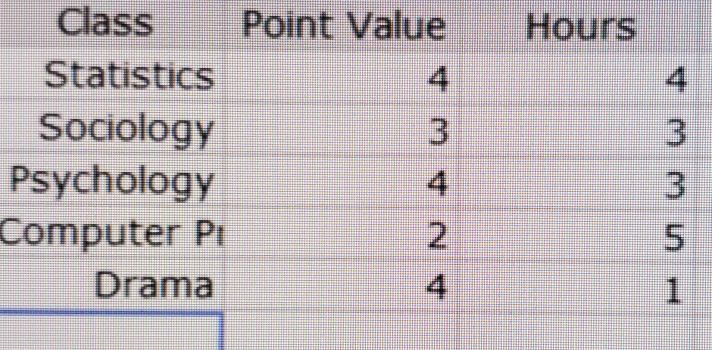
In drop down menu.

Choose the column that contains the weights in the Counts In

and drop down menu.

Select the Limits radio button for defining the midpoints.

Click Compute, and you'll have your result.



Objective 3, Page 1

List the formulas for approximating the population standard deviation and sample standard deviation of a variable from a frequency distribution. 00:00>> Here's the formula to approximate

the standard deviation of a variable from a frequency

distribution.

We'll begin with the population standard deviation.

In this formula, x sub i represents the midpoint

or value of class number i and f sub i represents

the frequency of class i.

In this formula, we subtract the mean mu

from each value where the mean mu is calculated

using Formula 1 from this section.

Once we have this difference, we square it,

multiply it by the frequency for that class,

add all these products together, and divide

by the sum of the frequencies.

The final step is to take the square root of this result.

Now, the difference between the population standard deviation

and the sample standard deviation

is that we subtract 1 from the sum

of the frequencies in the denominator.

Also notice that we use x bar to represent the sample mean

rather than mu to represent the population mean.

1. **DEFINITIONS** **Approximate Standard Deviation of a Variable from a Frequency Distribution**
2. **Population Standard Deviation**

σ=Σ(xi−μ)2fiΣfi−−−−−−−−−−−√

1. where:xi is the midpoint or value of the ith classfi is the frequency of the ith class
2. **Sample Standard Deviation**
3. s=Σ(xi−x¯¯¯)2fi(Σfi)−1−−−−−−−−−−−√

In this example, we'll learn to approximate

the standard deviation from a frequency distribution

using StatCrunch.

The following frequency distribution

represents the five-year rate of return of a random sample of 40

large-blended mutual funds.

Approximate the standard deviation of the five-year rate

of return.

Here is the frequency distribution.

Let's go to StatCrunch.

Here we are in StatCrunch.

I have the classes typed into the first column

and the frequencies in the second.

Again, notice how I type the classes, the lower limit

and then a space and then the word

"to" followed by another space, so that StatCrunch

can pick off the lower limit and understand that this

is binned or group data.

To do the calculation, press Stat, Summary Stats,

Grouped/Binned Data.

The bins are the classes, and the counts are the frequencies.

The midpoints, we want those to be

defined by the average of consecutive lower limits.

Finally, for statistics, I just need the standard deviation

here.

Click Compute, and there is our result.

To three decimal places, we have 2.181.

So we found that the standard deviation of the five-year rate

of return is 2.181%.

A quick note-- this is going to differ slightly than some

of the other techniques, which would return

a standard deviation of 2.182%.

And that is due to the way that StatCrunch

is computing the midpoint of each class,

including the last class.

Let's go over the steps.

Enter the summarized data into the spreadsheet,

press the Stat button, And from Summary Stats,

select Grouped/Binned Data.

Choose the column that contains the class under the Bins

In dropdown menu.

For Counts In, select the column that contains the frequencies,

and select the Consecutive Lower Limits Radio button

for defining the midpoints.

Click Compute, and you'll have your result.

**Approximating the Mean and Standard Deviation from Grouped Data**

1. If necessary, enter the summarized data into the spreadsheet. Name the columns.
2. Select **Stat**, highlight **Summary Stats**, and select**Grouped/Binned data**.
3. Choose the column that contains the class under the "Bins in:" drop-down menu. Choose the column that contains the frequencies in the "Counts in:" drop-down menu. Select the "Consecutive lower limits" radio button for defining the midpoints. Click Compute!.

Objective 3, Page 2

**Example 3 *Approximating the Standard Deviation from a Frequency Distribution***

The frequency distribution in Table 11 represents the five-year rate of return of a random sample of 40 large-blend mutual funds. Approximate the standard deviation five-year rate of return.

**Table 11**

| **Class (5-year rate of return)** | **Frequency** |
| --- | --- |
| 8-8.99 | 2 |
| 9-9.99 | 2 |
| 10-10.99 | 4 |
| 11-11.99 | 1 |
| 12-12.99 | 6 |
| 13-13.99 | 13 |
| 14-14.99 | 7 |
| 15-15.99 | 3 |
| 16-16.99 | 1 |
| 17-17.99 | 0 |
| 18-18.99 | 0 |
| 19-19.99 | 1 |

## Section 3.4 Measures of Position

### Objective

1. Determine and Interpret *z*-Scores
2. Interpret Percentiles
3. Determine and Interpret Quartiles
4. Determine and Interpret the Interquartile Range
5. Check a Set of Data for Outliers

#### Objective 1: Determine and Interpret z-Scores

Objective 1, Page 1

The z-score represents the distance

that a data value is from the mean,

in terms of the number of standard deviations.

Whether we're working with a population or a sample,

we find it by subtracting the mean from the value

and then dividing that difference

by the standard deviation.

The z-score is unitless.

It has a mean of 0 and a standard deviation of 1.

**DEFINITION**

The ***z*-score**represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation.

**Population**  z-**score**

z=x−μ**/**σ

**Sample**  z-**score**

z=x−x¯**/**s

The z-score is unitless. It has mean 0 and standard deviation 1.

1. If a data value is larger than the mean, the z-score is positive. If a data value is smaller than the mean, the z-score is negative. If the data value equals the mean, the z-score is zero. A z-score measures the number of standard deviations an observation is above or below the mean. For example, a z-score of 1.24 means the data value is 1.24 standard deviations above the mean. A z-score of −2.31 means the data value is 2.31 standard deviations below the mean.

Explain how to find a *z*-score and list the formulas for computing a population *z*-score and a sample *z*-score. **Population**  z-**score**

z=x−μ**/**σ

**Sample**  z-**score**

z=x−x¯**/**s

1. What does a positive *z*-score for a data value indicate? What does a negative z-score indicate? If a data value is larger than the mean, the z-score is positive. If a data value is smaller than the mean, the z-score is negative.
2. What does a *z*-score measure? A z-score measures the number of standard deviations an observation is above or below the mean.

Objective 1, Page 1 (continued)

How are *z*-scores rounded?  **NOTE**

Round z-scores to the nearest hundredth.

Objective 1, Page 2

**Example 1 *Determine and Interpret z-Scores***

Determine whether the Boston Red Sox or the Colorado Rockies had a relatively better run-producing season. The Red Sox scored 878 runs and play in the American League, where the mean number of runs scored was  and the standard deviation was  runs. The Rockies scored 845 runs and play in the National League, where the mean number of runs scored was  and the standard deviation was  runs. places.

X = 878 Red Sox's z-score=x−μ/σ=878−731.3/54.9=2.67

X= 845 Rockies' z-score=x−μ/σ =845−718.3/61.7=2.05

So the Red Sox had a run production 2.67 standard deviations above the mean, whereas the Rockies had a run production 2.05 standard deviations above the mean. Therefore, the Red Sox had a relatively better year at scoring runs than did the Rockies.

Objective 1, Page 5

With negative *z*-scores, we need to be careful when deciding the better outcome. For example, when comparing finishing times for a marathon the lower score is better because it is more standard deviations below the mean.

#### Objective 2: Interpret Percentiles

Objective 2, Page 1

What does the *k*th percentile represent? Recall that the [median](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_3227bde4-a224-e30b-03b9-9dd9e575f192) divides the lower 50% of a data set from the upper 50%. The median is a special case of a general concept called the *percentile*.

**DEFINITION**

The kth**percentile**, denoted Pk, of a set of data is a value such that k percent of the observations are less than or equal to the value.

Percentiles divide a set of data, written in ascending order, into 100 parts such that 99 percentiles can be determined. For example, P1 divides the bottom 1% of the observations from the top 99%,P2 divides the bottom 2% of the observations from the top 98%, and so on. Figure 8 displays the 99 possible percentiles.

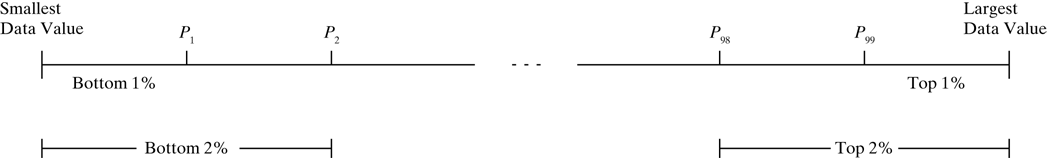


Figure 8

Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT, use percentiles to let students know how they scored on the exam in relation to all others who took the exam.

Objective 2, Page 2

**Example 2 *Interpreting a Percentile***

Jennifer just received the results of her SAT exam. Her math score of 600 is at the 74th percentile. Interpret this result. Interpretation

A percentile rank of 74 means that 74% of the SAT math scores are less than or equal to 600 and 26% of the scores are greater than 600. So 26% of the students who took the exam scored better than Jennifer.

#### Objective 3: Determine and Interpret Quartiles

Objective 3, Page 1

Define the first, second, and third quartiles. The most common percentiles are **quartiles**, which divide data sets into fourths, or four equal parts.

* The first quartile, denoted Q1, divides the bottom 25% of the data from the top 75%.
* The second quartile, Q2, divides the bottom 50% of the data from the top 50%.
* The third quartile, Q3, divides the bottom 75% of the data from the top 25%.

Figure 9 illustrates the concept of quartiles.

**In Other Words**

[](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_1_cbcf0016-3acf-26bb-09b4-bbd254eca0af)

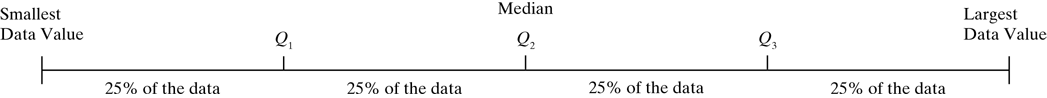


Figure 9

1. [Alt-Text](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx)Objective 3, Page 2

List the three steps for finding quartiles. **Finding Quartiles**

1. Arrange the data in ascending order.
2. Determine the median, M, or second quartile, Q2.
3. Divide the data set into two halves: the observations less than M and the observations greater than M. The first quartile, Q1, is the median of the bottom half, and the third quartile, Q3, is the median of the top half. Do not include M in these halves
4. Objective 3, Page 3

**Example 3 *Finding and Interpreting Quartiles***

The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual’s vehicle. Table 12 represents a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute for 2007 models. Find and interpret the first, second, and third quartiles for collision coverage claims.

**Table 12**

|  |  |  |
| --- | --- | --- |
| $6751 | $9908 | $3461 |
| $2336 | $21,147 | $2,332 |
| $189 | $1185 | $370 |
| $1414 | $4668 | $1953 |
| $10,034 | $735 | $802 |
| $618 | $180 | $1657 |

In this example, we'll learn to find and interpret quartiles

using StatCrunch.

The Highway Loss Data Institute routinely

collects data on collision coverage claims.

Collision coverage insures against physical damage

to an insured individual's vehicle.

The following data represent a random sample of 18 collision

coverage claims based on data obtained from the Highway Loss

Data Institute for 2007 models.

Here are the data.

Find and interpret the first, second, and third quartiles

for collision coverage claims.

Let's go ahead and enter this in StatCrunch

I've already typed the claims in column 1.

To compute the quartiles, I press Stat, Summary Stats,

Columns, select the column containing the data.

And I want to compute the median, which

is the second quartile, as well as Q1 and Q3.

Press compute, and we have our result.

Q1 is 735.

The median, or Q2 is 1,805, and Q3 is 4,668.

So the first quartile, Q1, was $735.

The second quartile, or the median, was $1,805.

And the third quartile, Q3, was $4,668.

Let's interpret these results.

Interpreting these results, we know

that 25% of the collision claims are

less than or equal to the first quartile, $735.

And we also know that 75% of the collision claims

are greater than this value.

For the second quarter or median,

we know that 50% or half of the collision claims

are less than or equal to the second quartile $1,805.

And we also know that 550% percent of the values

are greater than the second quartile.

Finally, for the third quartile, 75% of the collision claims

are less than or equal to the third quartile $4,668.

And since we know that 75% are less than or equal to it,

we know that 25% of the collision claims

are greater than this value.

OK let's go over the StatCrunch steps.

Press the Stat button.

And from Summary Stats, select Columns.

Select the column containing the data,

select median, Q1, and Q3, and click Compute.

And you'll have your result.

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Stat**, highlight **Summary Stats**, and select **Columns**.
3. Click on the variable you want to summarize. Deselect any statistics you do not wish to compute by clicking on the statistic name. If you wish to compute certain statistics, hold down the Control (Ctrl) key (or Command on an Apple) when selecting the statistic. Click Compute!.

Solution

**Step 1.** Write the data in ascending order:

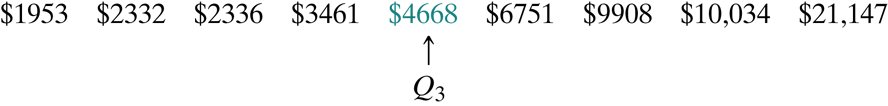
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $180 | $189 | $370 | $618 | $735 | $802 | $1185 | $1414 | $1657 |
| $1953 | $2332 | $2336 | $3461 | $4668 | $6751 | $9908 | $10,034 | $21,147 |

There are n=18 observations, so the median, or second quartile, Q2, is the mean of the 9th and 10th observations. Therefore, M=Q2=$1657+$19532=$1805.

The median of the bottom half of the data is the first quartile, Q1, which is the 5th observation, so Q1=$735.



The median of the top half of the data is the third quartile, Q3, which is the 5th observation, so Q3=$4668.



Interpretation

* For the first quartile, 25% of the collision claims are less than or equal to $735, and 75% of the collision claims are greater than $735.
* For the second quartile, 50% of the collision claims are less than or equal to $1805, and 50% of the collision claims are greater than $1805.
* For the third quartile, 75% of the collision claims are less than or equal to $4668, and 25% of the collision claims are greater than $4668.

#### Objective 4: Determine and Interpret the Interquartile Range

Objective 4, Page 1

1. Which measure of dispersion is resistant? Quartiles For this reason, quartiles are used to define a resistant measure of dispersion.

Define the interquartile range, IQR. **DEFINITION**

**The interquartile range,** IQR, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the first and third quartiles and is found using this formula IQR=Q3−Q1.

Solution

The interquartile range is

IQR=Q3−Q1=$4668−$735=$3933

Interpretation

The IQR, or range of the middle 50% of the observations, for the collision claim data is $3933. The IQR is used as the measure of dispersion for data sets that are skewed or contain extreme observations because it is a resistant measure of dispersion.

Solution

The interquartile range is

IQR=Q3−Q1=$4668−$735=$3933

Interpretation

The IQR, or range of the middle 50% of the observations, for the collision claim data is $3933. The IQR is used as the measure of dispersion for data sets that are skewed or contain extreme observations because it is a resistant measure of dispersion.

**Deciding Which Measure of Central Tendency and Dispersion to Report**

Let’s compare the measures of central tendency and dispersion discussed for the collision claim data.

* The median $1805  is a more representative measure of "center" than the mean because the data are skewed to the right (only 5 of the 18 observations are greater than the mean, which is $3874.4 ).
* The range is $21,147 - $180 = $20,967. The standard deviation is $ 5301.60 and the interquartile range is $3933. The values of the range and standard deviation are affected by the extreme claim of $ 21,147. In fact, if this claim were $120,000 (let’s say the claim was for a totaled Mercedes S-class AMG), then the range and standard deviation would increase to $119,820 and $27,782.5  respectively. The interquartile range would not be affected.

Therefore, when the distribution of data is highly skewed or contains extreme observations, it is best to use the median as the measure of central tendency and the interquartile range as the measure of dispersion because these measures are resistant.

| **SUMMARY: WHICH MEASURES TO REPORT** | | |
| --- | --- | --- |
| **Shape of Distribution** | **Measure of Central Tendency** | **Measure of Dispersion** |
| Symmetric | Mean | Standard Deviation |
| Skewed Left Or Skewed Right | Median | Interquartile Range |

For the remainder of this course, the direction **describe the distribution** will mean to describe its shape (skewed left, skewed right, or symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

Objective 4, Page 2

**Example 4 *Finding and Interpreting the Interquartile Range***

Determine and interpret the interquartile range of the collision claim data from Table 12 in Example 3.

**Table 12**

|  |  |  |
| --- | --- | --- |
| $6751 | $9908 | $3461 |
| $2336 | $21,147 | $2332 |
| $189 | $1185 | $370 |
| $1414 | $4668 | $1953 |
| $10,034 | $735 | $802 |
| $618 | $180 | $1657 |

Objective 4, Page 4

1. If the shape of a distribution is symmetric, which measure of central tendency and which measure of dispersion should be reported?
2. If the shape of a distribution is skewed left or skewed right, which measure of central tendency and which measure of dispersion should be reported? Why?

#### Objective 5: Check a Set of Data for Outliers

Objective 5, Page 1

What is an outlier? When analyzing data, we must check for extreme observations, called **outliers**. Outliers can occur by chance, because of errors in the measurement of a variable, during data entry, or from errors in sampling.

Outliers aren't always due to error or chance. Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European cars. If our sample included a Ferrari F430 Spider (approximately $175,000), it probably would be an outlier because this car costs much more than the typical European car. The value of this car would be considered *unusual* because it is not a typical value from the data set.

1. Objective 5, Page 2

List the four steps for checking for outliers by using quartiles. **ecking for Outliers by Using Quartiles**

**Step 1.** Determine the first and third quartiles of the data.

2. Compute the interquartile range.

3. Determine the fences. **Fences** serve as cutoff points for determining outliers.

Lower fence=Q1−1.5(IQR)

Upper fence=Q3+1.5(IQR)

If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

3.4 Measures of Position

Objective 5, Page 3

**Example 5 *Checking for Outliers***

Check the data in Table 12 on collision coverage claims for outliers.

**Table 12**

|  |  |  |
| --- | --- | --- |
| $6751 | $9,908 | $3461 |
| $2336 | $21,147 | $2332 |
| $189 | $1185 | $370 |
| $1414 | $4668 | $1953 |
| $10034 | $735 | $802 |
| $618 | $180 | $1657 |

## Section 3.5 The Five-Number Summary and Boxplots

Let's consider what we've learned so far.

In Chapter 2, we discussed techniques

for graphically representing data.

These summaries included bar graphs, pie charts, histograms,

stem and leaf plots, and time series graphs.

In Sections 3.1 to 3.4, we presented techniques

for measuring the center of a distribution,

spread in a distribution, and relative position

of observations in a distribution of data.

Why do we want these summaries?

What purpose do they serve?

Well, we want these summaries to see what the data can tell us.

We explore the data to see if they contain

interesting information that may be useful in our research.

The summaries make this exploration much easier.

In fact, because these summaries represent an exploration,

a famous statistician named John Tukey

called this material exploratory data analysis.

Tukey defined exploratory data analysis

as detective work, numerical detective work,

or graphical detective work.

He believed exploration of data is best carried out the way

a detective searches for evidence

when investigating a crime.

Our goal is only to collect and present evidence.

Drawing conclusions or inference is

like the deliberations of the jury.

What we have done so far falls under the category

of exploratory data analysis.

We have only collected information

and presented summaries, not reach

### Objectives

1. Determine the Five-Number Summary
2. Draw and Interpret Boxplots

#### Objective 1: Determine the Five-Number Summary

Objective 1, Page 1

What values does the five-number summary consist of? Remember that the [median](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_3c55b7f8-21ce-ce26-ae94-45926d62dac2) is [resistant](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_e29a0f17-8162-eecb-bd63-ff4f61e87e9d) to extreme values, so it is the preferred measure of central tendency when data are [skewed right](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_1599826a-b566-f56d-a67a-e695424b2e4a) or [skewed left](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_d8987967-d978-36af-a91a-6cf79e920a1c).

The three measures of dispersion that are not resistant are the [range](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_cbf0f743-2e77-f4ba-3694-812851a4a1ec), [standard deviation](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_4a11391a-4c57-880d-bda5-6b8f1d522d01), and [variance](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_384be5a5-3035-32a6-b8f9-24e6ed53c041). The [interquartile range](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_1_ff0c279d-a2b0-84f1-fc33-572166d275a8) is resistant. However, the median, Q1, and Q3 do not provide information about the extremes of the data. For this, we need the smallest and largest values in the data set.

The **five-number summary** of a set of data consists of the smallest data value, Q1, the median, Q3, and the largest data value. We use the five-number summary to learn information about the extremes of the data set. The summary is organized as follows:

Objective 1, Page 2

**Example 1 *Obtaining the Five-Number Summary***

Table 13 shows the finishing times (in minutes) of the men in the 60- to 64-year-old age group in a 5-kilometer race. Determine the five-number summary of the data.

**Table 13**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 19.95 | 23.25 | 23.32 | 25.55 | 25.83 | 26.28 | 42.47 |
| 28.58 | 28.72 | 30.18 | 30.35 | 30.95 | 32.13 | 49.17 |
| 33.23 | 33.53 | 36.68 | 37.05 | 37.43 | 41.42 | 54.63 |

Data from Laura Gillogly, student at Joliet Junior College

#### Objective 2: Draw and Interpret Boxplots

Objective 2, Page 1

List the five steps for drawing a boxplot. **Drawing a Boxplot**

Determine the lower and upper fences:

Lower Fence =Q1−1.5(IQR

Upper Fence = )=Q3+1.5(IQR)

where IQR=Q3−Q1

2.Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q1,M, and Q3. Enclose these vertical lines in a box.

3.Label the lower and upper fences with a temporary mark.

4.Draw a line from Q1 to the smallest data value that is larger than the lower fence. Draw a line from Q3 to the largest data value that is smaller than the upper fence. These lines are called **whiskers**.

5.Plot any data values less than the lower fence or greater than the upper fence as outliers. Outliers are marked with an asterisk (\*). Remove the temporary marks labeling the fences.

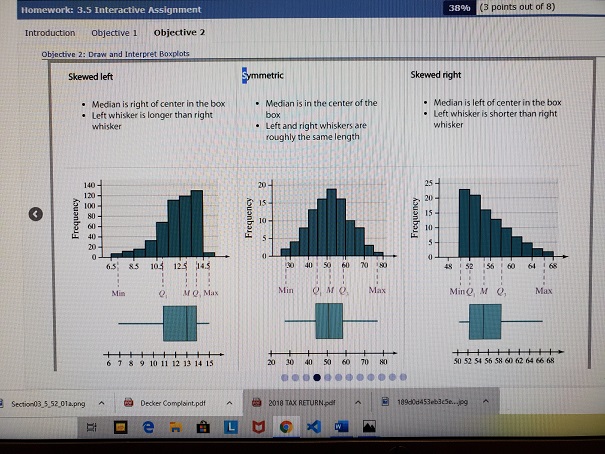
Objective 2, Page 2

**Example 2 *Constructing a Boxplot***

Use the results of Example 1 to construct a boxplot of the finishing times of the men in the 60- to 64-year-old age group.

(The five-number summary is: 19.95, 26.06, 30.95, 37.24, 54.63.)

#### Objective 2, Page 4 Drawing Boxplots

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Boxplot**.
3. Click on the variable whose boxplot you want to draw. Check the boxes "Use fences to identify outliers" and "Draw boxes horizontally." Enter label for the X-axis. Enter a title for the graph. Click Compute!.
4. If the right whisker of a boxplot is longer than the left whisker and the median is left of the center of the box, what is the most likely shape of the distribution? 

Objective 2, Page 5

When describing the shape of a distribution from a boxplot, be sure to justify your conclusion. Possible areas to discuss:

* Compare the length of the left whisker to the length of the right whisker
* The position of the median in the box
* Compare the distance between the median and the first quartile to the distance between the median and the third quartile
* Compare the distance between the median and the minimum value to the distance between the median and the maximum value

Objective 2, Page 10

**Example 3 *Comparing Two Distributions Using Boxplots***

Table 14 shows the red blood cell mass (in millimeters) for 14 rats sent into space (flight group) and for 14 rats that were not sent into space (control group). Construct side-by-side boxplots for red blood cell mass for the flight group and control group. Does it appear that space flight affects the rats' red blood cell mass?

**Table 14**

| **Flight** | **Control** |
| --- | --- |
| 7.43 | 8.65 |
| 7.21 | 6.99 |
| 8.59 | 8.40 |
| 8.64 | 9.66 |
| 9.79 | 7.62 |
| 6.85 | 7.44 |
| 6.87 | 8.55 |
| 7.89 | 8.70 |
| 9.30 | 7.33 |
| 8.03 | 8.58 |
| 7.00 | 9.88 |
| 8.80 | 9.94 |
| 6.39 | 7.14 |
| 7.54 | 9.14 |

Data from NASA Life Sciences Data Archive

**Using a Boxplot to Determine Skewness**

The boxplot in [Figure 10(d)](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_5_164f10d1-ec27-464e-2b0b-cd877e7359a2)suggests that the distribution is skewed right because the right whisker is longer than the left whisker and the median is left of center in the box. We can also assess the shape of a distribution using quartiles. The distance from M to Q1 is 4.89 (=30.95–26.06), whereas the distance from M to Q3 is 6.29 (=37.24–30.95). Also, the distance from M to the minimum value is 11 (=30.95–19.95), whereas the distance from M to the maximum value is 23.68 (=54.63–30.95).

When describing the shape of a distribution from a boxplot, be sure to justify your conclusion as we did above.



**Drawing Boxplots**

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and highlight **Boxplot**.
3. Click on the variable whose boxplot you want to draw. If you want to draw side-by-side boxplots, hold the Control (or Command) key down while clicking the variable. Check the boxes "Use fences to identify outliers" and "Draw boxes horizontally." Enter label for the X-axis. Enter a title for the graph. Click Compute!.