

7.1 Central Tendency

DISCOVER

THINK

APPLY

SOLVE

REFLECT

Purpose

The role this topic plays in quantitative reasoning

One of the goals of descriptive statistics is to describe the characteristics of a large set of data with a few numbers. Perhaps the most relevant number is a measure of central tendency. It is a number that describes where the data cluster around a central area. Measures of central tendency provide useful information such as, “How much can I expect to earn at that company?” or “Can I afford to purchase a house in that neighborhood?” There are various measures of central tendency that can be defined to represent a data set. It is critical to determine in each situation which measure — the mean, median or mode — is the best measure of central tendency. Understanding the concepts together with the computational methods will enable you to describe the behavior of a data set and make predictions about future data elements drawn from the same population.

Learning Goals

What you should learn while completing this activity

1. Become proficient in computing each major measure of central tendency.
2. Determine which measure of central tendency is most appropriate for a given data set.
3. Present central tendency visually
4. Predict the relationship among the mean, median and modes from different data sets

Discovery

Finding out for yourself

Which single number would best describe the heights of adult males in the USA? Which single number would best describe the heights of adult females in the USA? How different would those two numbers be? Would you expect the means and medians to be similar? Would you expect that the central tendency numbers would be the same for various ethnic groups?

If you wanted to purchase a home in a certain community what measure of central tendency would you use to determine if you could afford such a purchase? Would you expect the means and medians to be different?

What Do You Already Know?

Tapping into your existing knowledge

1. What are five types of averages that you know?
2. Is the average a mean or median?
3. Do you know how to calculate the mean of a series of numbers?
4. Do you know how to calculate the median of a series of numbers?
5. Have you heard of a *mode*? What is it?

Mathematical Language

Terms and notation

a measure of central tendency — a number that describes the central behavior of a set of data. In particular it describes where data points cluster. There are three common measures of central tendency, the mean, median and mode.

mean — the arithmetic average of a set of numbers. It is determined by finding the sum of the elements in a data set and then dividing by the number of elements in a data set. It is often denoted as: \bar{x} .

median — the number in the middle of the data set. It is the 50th percentile. It is determined by listing all of the data points in order from lowest to highest. If there are an odd number of data points, there is a unique number in the middle and that is the median. If there is an even number of data points, there are two numbers in the middle. Add those two numbers and then divide by two to determine the median.

mode — the data point that occurs most often. A set of data may have more than one mode.

nominal data — data is listed in categories and there is no ordering scheme. Example: 0 = female, 1 = male.

ordinal data — data is listed in categories but differences are meaningless. Example: List favorite sports in order from 1 – 5. 1 = football, 2 = soccer, 3 = baseball, 4 = basketball, 5 = boxing.

interval data — the differences between data values are meaningful but there is no “zero” so ratios are meaningless. Example: 0 degrees Celsius does not mean lack of all heat, so data in degrees Celsius is interval.

ratio data — differences are relevant and there is a natural zero. Examples: Ages, degrees Kelvin, odometer mileage.

symmetric — the left half of the histogram is a mirror image of the right half of the histogram

skewed — a data set is not symmetric and it extends much more to either the right or left side

outlier — data points that are vastly different from the great majority of the other data points

Information

What you need to know

 READINGS

 RESOURCES

METHODOLOGY

CONSTRUCTING MEASURES OF CENTRAL TENDENCY (CT)

 **Scenario:** Random Sample of Male Heights

Step	Explanation
1. Describe the data set	<i>Is the set a random collection of data from a larger sample or is it a convenience sample? What characteristic is being measured? What is the relevant unit? Are there any specific conditions?</i>

 WATCH IT WORK!

The set is a random collection of heights of males in the USA. The unit is in inches. All data points are rounded to the nearest inch.

Step	Explanation
2. Describe the type of data	<i>Is the data nominal, ordinal, interval or ratio?</i>



Since heights are measured in inches, there is a natural zero and ratios are relevant so the data has the ratio level of measurement.

3. Construct a histogram or bar graph	<i>Determine if the data has outliers or if it is skewed. Use a histogram for ratio, or interval data. Use a bar graph for nominal or ordinal data. If some numbers are drastically different from the great majority of data, the set has outliers. If the left and right sides are not mirror images and if either tail is much longer than the other, the data set is skewed</i>
---------------------------------------	---



Looking at the histogram, it appears that the data set is symmetric.

4. Determine which measures of CT are reasonable	<p><i>If the data has the ratio or interval level of measurement, then the mean, median and mode are relevant.</i></p> <p><i>If the ratio/interval data set is skewed or has outliers, the median is most reasonable.</i></p> <p><i>If the ratio/interval data is symmetric, the mean is most reasonable</i></p>
--	--



The heights are symmetric and ratio, therefore the mean is most relevant. Since the data is ratio, it is fair to construct the median and the mode.

5. Compute the measures of CT and determine which is best	<p><i>If it is fair to construct the mean, median, or mode do so:</i></p> <p><i>Mean = Sum(var)/Num</i></p> <p><i>Median is the middle value of a sorted set of values if an odd number of values or the average of two middle values if an even number of values.</i></p> <p><i>Mode is determined by identifying the data point that occurs most often. A set of data may have more than one mode.</i></p>
---	--



The mean is 69.8 inches. The median is 69.5 inches. The mode is 72 inches. Since the data set is symmetric, the mean of 69.8 inches is best.

6. Validate	<i>If the data is a random sample from a larger population, collect a second sample and see if the measures of central tendency are similar. If possible, research the question to see what others have found. Compare with and without outliers.</i>
-------------	---




According to the CDC, the average height for men in the USA is 69.3 inches. (Reference is online.)



Scenario 1: Symmetric Data


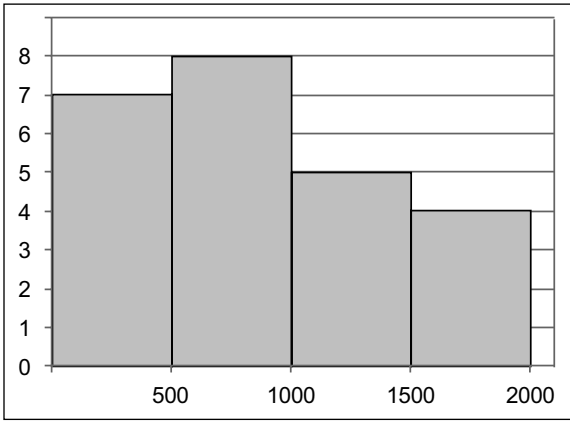
A fashion designer wants to know the average height of their female clients. They will use this information as part of the design process.


Step	Watch it Work!								
1. Describe the data set	<p>The following is a list of heights of 16 randomly selected female clients of a fashion designer. Unit is inches and rounded to the nearest inch.</p> <p style="text-align: center;">66 65 69 66 64 72 66 64 67 62 68 62 68 65 65 64</p>								
2. Describe the type of data	<p>Heights of women are ratio level of measurement. The differences are relevant and 0 inches means the complete lack of height.</p>								
3. Construct a histogram or bar graph	<div style="display: flex; align-items: flex-start;"> <div style="flex: 1;"> <table border="1" style="margin-top: 10px;"> <caption>Histogram Data</caption> <thead> <tr> <th>Height (inches)</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>62-64</td> <td>5</td> </tr> <tr> <td>66</td> <td>10</td> </tr> <tr> <td>70</td> <td>1</td> </tr> </tbody> </table> </div> <div style="flex: 1; padding-left: 20px;"> <p>Use technology to construct a histogram. Determine if the data has outliers or if it is skewed.</p> <p>There are no outliers. The data does not seem to be symmetric. There is not a long tail on one side so we cannot say for certain that the data set is skewed.</p> </div> </div>	Height (inches)	Frequency	62-64	5	66	10	70	1
Height (inches)	Frequency								
62-64	5								
66	10								
70	1								
4. Determine which measures of CT are reasonable	<p>Since the data is ratio interval of measurement, the mean, median and mode are reasonable.</p> <p>The data set is neither symmetric nor strongly skewed so both the mean and median may be reasonable.</p>								
5. Compute the measures of CT and determine which is best	<p>To compute the mean, add the numbers in the data set and divide by the number of elements in the data set.</p> $66 + 65 + 69 + 66 + 64 + 72 + 66 + 64 + 67 + 62 + 68 + 62 + 68 + 65 + 65 + 64 = 1053$ $\text{Mean} = \frac{1053}{16} = 65.8125$ <p>There is an even number of elements in this sorted set so there are two numbers straddling the middle.</p> <p style="text-align: center;">62 62 64 64 64 64 65 65 65 66 66 66 67 68 68 69 72</p> <p>The median is obtained by adding those together and dividing by 2.</p> $\text{Median} = \frac{65 + 66}{2} = 65.5$ <p>Notice that the mean and median are similar but not identical. Both measures are reasonable in this example.</p> <p>The mode corresponds to the date element that appears most often. In this example 64, 65, and 66 are all modes since they appear three times. The mode may not be useful in this context.</p>								

Step	 Watch it Work!
6. Validate	According to the CDC, the average height for all US women is 63.8 inches. It appears that the clients for the fashion designer are not a random sample of the entire population. In order to validate this study a second sample of clients should be examined.

Scenario 2: Skewed Data

A person wants to purchase a single-family home in an established community. She is interested in the selling prices of recently sold homes in that community. She wants to know what the “typical” price is for such a home. She researches the list prices on zillow.com.

Step	 Watch it Work!																									
1. Describe the data set	<p>The following is a set of list prices for homes in an affluent neighborhood.</p> <table style="margin-left: auto; margin-right: auto;"> <tbody> <tr> <td>\$709,000</td> <td>\$1,500,000</td> <td>\$1,199,993</td> <td>\$1,299,000</td> <td>\$1,595,000</td> </tr> <tr> <td>\$585,000</td> <td>\$615,000</td> <td>\$1,650,000</td> <td>\$555,000</td> <td>\$495,000</td> </tr> <tr> <td>\$1,199,000</td> <td>\$549,000</td> <td>\$499,999</td> <td>\$1,275,000</td> <td>\$420,000</td> </tr> <tr> <td>\$199,900</td> <td>\$449,000</td> <td>\$720,000</td> <td>\$1,199,000</td> <td>\$264,993</td> </tr> <tr> <td>\$1,549,993</td> <td>\$624,993</td> <td>\$585,000</td> <td>\$289,000</td> <td>\$395,000</td> </tr> </tbody> </table>	\$709,000	\$1,500,000	\$1,199,993	\$1,299,000	\$1,595,000	\$585,000	\$615,000	\$1,650,000	\$555,000	\$495,000	\$1,199,000	\$549,000	\$499,999	\$1,275,000	\$420,000	\$199,900	\$449,000	\$720,000	\$1,199,000	\$264,993	\$1,549,993	\$624,993	\$585,000	\$289,000	\$395,000
\$709,000	\$1,500,000	\$1,199,993	\$1,299,000	\$1,595,000																						
\$585,000	\$615,000	\$1,650,000	\$555,000	\$495,000																						
\$1,199,000	\$549,000	\$499,999	\$1,275,000	\$420,000																						
\$199,900	\$449,000	\$720,000	\$1,199,000	\$264,993																						
\$1,549,993	\$624,993	\$585,000	\$289,000	\$395,000																						
2. Describe the type of data	Selling prices are ratio level of measurement. The differences are relevant and \$0 is a meaningful measure.																									
3. Construct a histogram or bar graph	<div style="display: flex; align-items: center;">  <div style="margin-left: 20px;"> <p>Use technology to construct a histogram. Determine if the data has outliers or if it is skewed.</p> <p>There are no outliers. The data does not seem to be symmetric.</p> </div> </div>																									
4. Determine which measures of CT are reasonable	<p>Since the data is ratio interval of measurement, the mean, median and mode are reasonable.</p> <p>The data set seems skewed positively so the median may be the best measure of central tendency.</p>																									


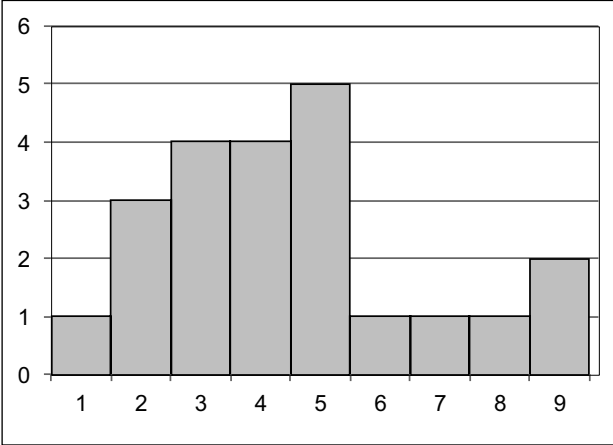
Step	 Watch it Work!									
5. Compute the measures of CT and determine which is best	<table border="1"> <thead> <tr> <th data-bbox="492 226 873 268">Compute the mean:</th> <th data-bbox="889 226 1458 268">Compute the median:</th> </tr> </thead> <tbody> <tr> <td data-bbox="492 279 873 394"> Add the numbers in the data set and divide by the number of elements in the data set. </td> <td data-bbox="889 279 1458 394"> Put the elements of the data set in order from lowest to highest. </td> </tr> <tr> <td data-bbox="492 405 873 562"> The sum of the elements in the data set is 20027871. There are 25 elements in the data set, so </td> <td data-bbox="889 405 1458 562"> There are an odd number of elements in this set so there is one number in the middle. </td> </tr> <tr> <td data-bbox="492 573 873 709"> $\text{Mean} = \frac{\\$2,042,2871}{25}$ $= \\$816,915$ </td> <td data-bbox="889 573 1458 1178"> $\text{Median} = \\$615,000$ \$199,900.00 \$264,993.00 \$289,000.00 \$395,000.00 \$420,000.00 \$449,000.00 \$495,000.00 \$499,999.00 \$549,000.00 \$555,000.00 \$585,000.00 \$585,000.00 \$615,000.00 \$624,993.00 \$709,000.00 \$720,000.00 \$1,199,000.00 \$1,199,000.00 \$1,199,993.00 \$1,275,000.00 \$1,299,000.00 \$1,500,000.00 \$1,549,993.00 \$1,595,000.00 \$1,650,000.00 </td> </tr> </tbody> </table>	Compute the mean:	Compute the median:	Add the numbers in the data set and divide by the number of elements in the data set.	Put the elements of the data set in order from lowest to highest.	The sum of the elements in the data set is 20027871. There are 25 elements in the data set, so	There are an odd number of elements in this set so there is one number in the middle.	$\text{Mean} = \frac{\$2,042,2871}{25}$ $= \$816,915$	$\text{Median} = \$615,000$ \$199,900.00 \$264,993.00 \$289,000.00 \$395,000.00 \$420,000.00 \$449,000.00 \$495,000.00 \$499,999.00 \$549,000.00 \$555,000.00 \$585,000.00 \$585,000.00 \$615,000.00 \$624,993.00 \$709,000.00 \$720,000.00 \$1,199,000.00 \$1,199,000.00 \$1,199,993.00 \$1,275,000.00 \$1,299,000.00 \$1,500,000.00 \$1,549,993.00 \$1,595,000.00 \$1,650,000.00	<p>Notice that the mean and median are not similar. Which number is more important to the buyer? If half the houses are under \$615,000, she will have many choices at or under that price range. Relatively few houses cost more than the mean of \$816,915. So in this case the median is a much better representative of central tendency.</p> <p>The mode corresponds to the price that appears most often. In this example \$585,000 and \$1,199,000 each appear twice. Both \$585,000 and \$1,199,000 are modes. The mode may not be useful in this context.</p>
Compute the mean:	Compute the median:									
Add the numbers in the data set and divide by the number of elements in the data set.	Put the elements of the data set in order from lowest to highest.									
The sum of the elements in the data set is 20027871. There are 25 elements in the data set, so	There are an odd number of elements in this set so there is one number in the middle.									
$\text{Mean} = \frac{\$2,042,2871}{25}$ $= \$816,915$	$\text{Median} = \$615,000$ \$199,900.00 \$264,993.00 \$289,000.00 \$395,000.00 \$420,000.00 \$449,000.00 \$495,000.00 \$499,999.00 \$549,000.00 \$555,000.00 \$585,000.00 \$585,000.00 \$615,000.00 \$624,993.00 \$709,000.00 \$720,000.00 \$1,199,000.00 \$1,199,000.00 \$1,199,993.00 \$1,275,000.00 \$1,299,000.00 \$1,500,000.00 \$1,549,993.00 \$1,595,000.00 \$1,650,000.00									
6. Validate	In order to validate this study a second sample of list prices should be examined.									



Scenario 3: Nominal Data

A cosmetologist is interested in the hair color of women in a certain ethnic community. She designs the following coding scheme.

Level 1: Black Level 4: Dark Brown Level 7: Dark Blonde
 Level 2: Darkest Brown Level 5: Brown Level 8: Medium Blonde
 Level 3: Very Dark Brown Level 6: Light Brown Level 9: Blonde

Step	 Watch it Work!																				
1. Describe the data set	A trained cosmetologist is able to categorize each woman. Her results follow: <p style="text-align: center;">5 5 4 2 3 1 9 5 4 7 8 2 3 4 4 3 5 3 2 6 9 5</p>																				
2. Describe the type of data	The differences and the ratios are irrelevant. Since this data is used to establish categories, it is nominal data.																				
3. Construct a histogram or bar graph	Use technology to construct a bar graph. <div style="text-align: right;">  <table border="1" style="display: none;"> <caption>Bar Graph Data</caption> <thead> <tr> <th>Category</th> <th>Frequency</th> </tr> </thead> <tbody> <tr><td>1</td><td>1</td></tr> <tr><td>2</td><td>3</td></tr> <tr><td>3</td><td>4</td></tr> <tr><td>4</td><td>4</td></tr> <tr><td>5</td><td>5</td></tr> <tr><td>6</td><td>1</td></tr> <tr><td>7</td><td>1</td></tr> <tr><td>8</td><td>1</td></tr> <tr><td>9</td><td>2</td></tr> </tbody> </table> </div>	Category	Frequency	1	1	2	3	3	4	4	4	5	5	6	1	7	1	8	1	9	2
Category	Frequency																				
1	1																				
2	3																				
3	4																				
4	4																				
5	5																				
6	1																				
7	1																				
8	1																				
9	2																				
4. Determine which measures of CT are reasonable	Since the data set is nominal, the only reasonable measure of central tendency is the mode.																				
5. Compute the measures of CT and determine which is best	The mode is the most occurring response. In this example, category 5 is the mode. The most common hair color was brown.																				
6. Validate	Repeat the process with a second set of women and see if the results are similar.																				

Oops! AVOIDING COMMON ERRORS

- **Generalizing results from convenient or non-representative sample**

Example: Eight friends buy lottery tickets and on their \$2 ticket, 3 individuals didn't win anything, 2 people won \$2, the other 3 people won \$5, \$50, and \$1,000. I can expect to average about \$132 for every ticket I buy.

Why? If your sample is fairly large and it comes from a random sample of the population, it is reasonable to assume that the overall central tendency will match the central tendency of your sample. If you use a convenient sample or a non-representative sample, do not generalize to a larger group.

- **Constructing means and medians for nominal data**

Example: A survey asks, "What is your favorite ice cream?" We pick 1: Vanilla, 2: Chocolate, 3: Strawberry, 4: Pistachio, 5: Rocky Road, and 6: Other. We find the average is skewed because of the popularity of vanilla and chocolate. The mean is 2.2.

Why? If you are given categorical data, do not attempt to compute the mean or median. The only measure that is reasonable is the mode.

- **Emphasizing the mean for highly-skewed data**

Example: The average GPA of students at Phillips Exeter Academy is 8.5. But half of the students have a GPA of at least 9.5 because the maximum GPA is 11.

Why? Real-world data is often skewed. This is especially true with financial data such as salaries or selling prices. The mean of skewed data is of limited value if you want to know a typical value.

Are You Ready?

Before continuing, you should be able to ...

I can...

- compute the mean for a set of data
- compute the median for sets of odd and even numbers of elements
- determine the mode of a data set
- determine which measure of central tendency is best for a given data set
- predict the relationship among the mean, median, and mode of a skewed data set.

OR Here's my question...

A Successful Performance Successful application of your learning looks like this

As you begin to apply what you've learned, you should have a good idea of what success looks like.

A SUCCESSFUL PERFORMANCE

I interpret and evaluate a data analysis presentation based on its central tendency. I...

- Select the most effective measures
- Perform high-quality analysis
- Produce meaning and critical implications

I present an analysis of a set data showing the central tendency. I...

- Use graphics effectively
- Use the measure of central tendency appropriately
- Explain graphs and statistics effectively

Demonstrate Your Understanding Apply it and show you know in context!

1. Create a set of data where the mean is much lower than the median. What does this tell you about your distribution?
2. Examine the data available at online and find the mean and median heights for right and left-handed people. Are the results as you expected?
3. Using the data available online, construct a bar graph for nominal data and define the relevant mode.
4. Imagine you have a set of college grade point averages for students graduating from an honors program. What would your predicted values for the mean, median, and mode be?

Hardest Problem How hard **can** it be? Can you still use what you've learned?

Based on the Models, the Methodologies, and the Demonstrate Your Understanding (DYU) problems in this activity, create the **hardest** problem you can. Start with the hardest DYU problem in this experience and by contrasting and comparing it with the other DYU problems, play “What if” with the different conditions and parameters in the various problems.

Can you still solve the problem? If so, solve it. If not, explain why not.

What are the conditions and parameters that make a problem where you must identify and measure central tendency a difficult problem to solve?

Troubleshooting Find the error and correct it!

What is the normal length of time it takes for a woman to conceive? The length of time for a woman to conceive (if that is her intention) is different than for a woman who is trying **not** to conceive. The average (mean) amount of time is 7.5 months for women who want to conceive. But 50% of women trying to conceive will be pregnant within 4 months (this is the median). Therefore, we can expect that a woman will have to wait for almost 8 months to determine if she can conceive.

