# Chapter 6: Central Tendency

**Objectives for This Chapter**

- Develop an understanding of central tendency.
- Make friends with Big Sigma.
- Define, discuss and compute the mean, median, and mode.
- Enroll in the prestigious Khan Academy.
- Understand how the differences between the mean, median and mode are affected by the shape of a distribution of scores.
- Identify to the most appropriate measure to describe the center of symmetric and skewed distributions.
- Figure out which measure of central tendency would win in a fight.

**Stuck in the middle? Where is the Middle?**

What is "central tendency," and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz.

Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is "3/5." How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, "Whad'ja get?" and then ask the instructor, "How did the class do?" In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a

| Student | Dataset A | Dataset B | Dataset C |
|---|---|---|---|

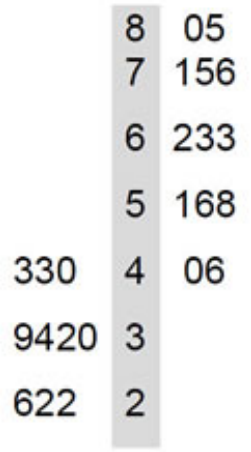
distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in this table. They are labeled "Dataset A," "Dataset B," and "Dataset C." Which of the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

| | | | |
|---|---|---|---|
| You | 3 | 3 | 3 |
| John's | 3 | 4 | 2 |
| Maria's | 3 | 4 | 2 |
| Shareecia's | 3 | 4 | 2 |
| Luther's | 3 | 5 | 1 |

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset 1. The problem is that the other four students had higher grades, putting yours below the center of the distribution.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.




Now let's change the example in order to develop more insight into the center of a distribution. This figure shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions.

The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.The maximum possible score was 89.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

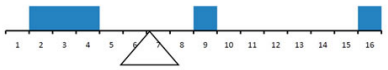
**Can we define the Middle with more precision?**

NO!! What kind of class do you think this is? Math?? This is statistics. We *apply* math. We never prove anything and we never get really precise. *Central tendency* is just "middleness." It's about as precise as Middle Earth (Gandalf Lives!) or middle management or middle school. I think that the following joke makes my position clear:

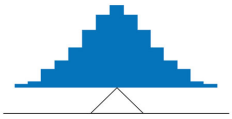
*A historian, an engineer and a statistician go duck hunting. a duck rises from the lake. the historian fires first, and shoots 10' over the duck. Then the engineer shoulders the shotgun and shoots 10' under the duck. Suddenly, the statistician jumps up and shouts excitedly, "We got him! We got him!"*

Not laughing? No LOL? Means that there is still hope for you. In order to understand the humor of this joke, you have to pass into the *first dimension* of statistics. This is the dimension that most people understand--the

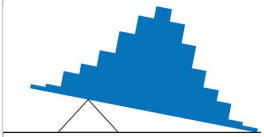
dimension of middleness. As I suggested in the previous section, whenever there is a group of numbers of interest, we're quite adept at inquiring about the middle of the distribution of scores. We like middleness. It makes us feel like we have some control over our lives. It helps us feel balance. In with the good air, out with the bad air. Let's talk about balance.



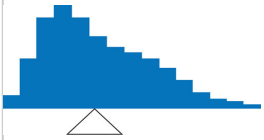
(../../images/central_tendency/balance1.jpg)A terrific metaphor for conceptualizing central tendency is ***balance***. When talking about numbers, it is the point at which the distribution is in balance. The figure here, shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and isplaced at its position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8. **Click Images to enlarge.**



(../../images/central_tendency/balance2.jpg)This image illustrates the balance that we typically think of. An equal distribution on both sides of the fulcrum. This is the type of balance that we desire most when we collect data. Few scores in the extreme high and low regions and increasingly more scores towrd the middle of the distribution.



**(../../images/central_tendency/unbalance.jpg)**This figure illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center. In other words, if I want to talk about middleness for this distribution, I better not veer to far from center.



(../../images/central_tendency/asymmetric.jpg)This last figure shows an asymmetric distribution. This distribution is skewed badly to the right, indicating that most scores (they could be annual income scores of people in (../../images/central_tendency/herman.jpg)Chicago, IL) are near the bottom of the distribution. There are a few scores that are really high. To balance it, we cannot put the fulcrum halfway between the lowest and highest values as we did with the first distribution. Placing the fulcrum at the "half way" point would cause it to tip towards the left. In this case, middleness is no longer in the middle, but with the fulcrum placed where it is, we have a balanced distribution.


"How do you expect me to average 55 miles an hour if I don't speed?"

Let's explore this from another angle. With the fulcrum, we want to place it in the right location to achieve balance in the distribution. If we put the fulcrum in the wrong place, then the physical object it is intended to balance will lean to one side or the other. We have a similar concern when choosing a measure of central tendency to describe an entire distribution. We strive to identify the best measure of central tendency to provide a summary of a group of scores. When this group of scores is normally distributed, then it's easy. The mean, median and mode fall at the same place. But if the distribution isn't symmetric, the mean, median and mode will be located in different places. Picking the wrong one to explain your research can lead others to an *unbalanced* or incorrect understanding of your work.

Wow! That metaphor almost worked. To get a more "hands-on" feeling for this idea, go to the Balance Scale Simulation (http://onlinestatbook.com/2/summarizing_distributions/balance.html) and select "Run The Simulation."

**So what is big sigma?**

What is Big Sigma? Seriously, you don't know? Big sigma is the thing that will turn you in to a bonafide, genuine, data crunching statistician.

| Grapes | X |
|---|---|

Many statistical formulas involve summing numbers. Fortunately there is a convenient

| | |
|---|---|
| 1 | 4.6 |
| 2 | 5.1 |
| 3 | 4.9 |
| 4 | 4.4 |

notation for expressing summation. This section covers the basics of this **summation notation**, which I also affectionately refer to as *Big Sigma*. The capital Greek letter sigma (Σ) indicates summation. It really is an instruction to *add everything that comes after it.* Usually, what comes after it is a variable or sequence of variables.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes. The data are shown in this table. We label Grape 1's weight X1, Grape 2's weight X2, etc. The following formula instructs us to sum up the weights of the four grapes. Simple...right? In some ways, yes. But this is not the only "look" that Big Sigma will give us. In a math class, Big Sigma would appear with extra instructions on its top and bottom. These extra instructions are not necessary for us to perform our calculations in this class. So, they are not shown.

$$\sum X$$

Aside from notation on the top and bottom of Big Sigma, this symbol will be expressed in a few other ways in this course. Let's take a look at a couple of looks that Big Sigma will take in an upcoming chapter. Many formulas involve *squaring numbers before they are summed*. This is indicated as

$$\Sigma X^2 = 4.62 + 5.12 + 4.92 + 4.42 = 21.16 + 26.01 + 24.01 + 19.36 = 90.54$$

Notice that...

$$\left(\sum X\right)^2 \neq \sum X^2$$

...because the expression on the left means to sum up all the X values **AND THEN** square the sum ($19^2 = 361$) whereas the expression on the right means to **SQUARE THE NUMBERS FIRST** and then sum thos squares (90.54, as shown). We'll worry more about these Big Sigma looks in the next chapter. For now, let's just focus on ΣX.

**How do we calculate mean, media and mode?**

This section defines the three most common measures of central tendency: the *mean*, the *median*, and the *mode*.

## Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the *mean of a population*. The symbols "M" or an X with a bar over the top (hereafter referred to as *X-bar*) are used for the mean of a sample.

*Computation Of The Mean*

$\mu = \Sigma X/N$ where ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for the sample mean (M or X-bar) is essentially identical:

$M = \Sigma X/N$ where ΣX is the sum of all the numbers in the sample and N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

The table below shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League (NFL) in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20 20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

$$\mu = \Sigma X/N$$
$$= 634/31$$
$$= 20.4516$$

| *X* | *f* |
|---|---|
| 37 | 1 |
| 33 | 2 |
| 32 | 1 |
| 29 | 1 |
| 28 | 2 |
| 23 | 1 |
| 22 | 3 |
| 21 | 3 |
| 20 | 2 |
| 19 | 2 |
| 18 | 4 |
| 16 | 1 |
| 15 | 1 |
| 14 | 3 |
| 12 | 2 |
| 9 | 1 |
| 6 | 1 |

What if this distribution was arranged as a *frequency distribution* and you didn't have all of the raw scores to calculate the mean? You guessed it. First thing that we would do is figure out the sample size, N, by adding up all of the numbers in the *f* column. In this case, this gives us 31. Then, we would multiply each each pair of numbers, add those products and divide by the total number of scores. Remember that the values in the left column represent all of the scores that comprise the distribution. The most TD passes thrown by a team was 37 and the fewest was 6 (sorry Cincinnati). So, our average will be somewhere in between. Calculation would look like the following:

$$\mu = \Sigma fX/N$$
$$= 634/31$$
$$= 20.4516$$

Notice that in the above formula, we have inserted an f, between sigma and N. This instructs us to multiply each score by its corresponding frequency before adding. Another way to think about this is that Σ*f*X is basically telling us to create a third column--an *f*X column--which will then be summed and divided by the total number of scores.

Statistics: The ave...

The mean is the *most useful* of the three measures of central tendency because many important statistical procedures are based on it. Because the mean takes into account *all of the data* in a distribution, it is the most reliable or stable of the three measures of central tendency. Specifically, if several samples are taken from a known population and each of the three measures of central tendency is computed for each sample, the mean will vary least of the three measures, followed by the median and the mode, in that order. (http://www.khanacademy.org)

Before we talk at length about the mean, median and mode, It's time for you to take a giant step in your academics. It's time to enroll in the Khan Academy. Enrollment is easy--just bookmark www.khanacademy.org (http://www.khanacademy.org) in your web browser(s). That's it! You're enrolled.

Here is a video of Salman Khan discussing measures of central tendency. Salman is really an interesting and very nice guy with a wonderful mission--to teach math and science topics to the masses. This website will help you in so many classes. It's not just a statistics website. In fact, the Khan Academy didn't have any statistics videos, a few years ago. I spoke with Salman a few times about this, and at some point, I noticed that more and more statistics instruction was becoming a part of the Academy's curriculum. I like to think that my nagging was part of the reason for this. I'm really good at nagging. Anyhoo...you owe it to yourself to check out everything that the Khan Academy (http://www.khanacademy.org) has to offer. You won't regret your visit. We'll visit the Khan Academy several times during the semester.

## Median

This median (symbolized as *Md*)is also a frequently used measure of central tendency. The *median is the midpoint* of a distribution: the same number of scores are above the median as below it. The median is also referred to as

the 50th percentile. Before you can find the midpoint of a distribution, you must order the distribution of scores in sequential order--lowest to highest or highest to lowest. Just as we practiced in a previous chapter. You can order the raw data or create a frequency distribution with X and f columns.

For the data in this table, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

*Computation of the Median*

When there is an **odd number** of scores in a distribution, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an **even number** of scores, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5.

Scores: 1,1,2,2,2,2,2,3,3,3,3,4,4,5

| Score | Frequency | Cumulative Frequency |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 5 | 7 |
| 3 | 4 | 11 |
| 4 | 2 | 13 |
| 5 | 1 | 14 |

What if you had to identify the median from a *cumulative frequency* (Cum *f*) distribution? First thing that we would need to do is figure out the total number of scores in the distribution. In this table, we see that the total number of scores in the distribution is 14. Okay. So, we know that there are an even number of scores in this distribution. This means that the median will be the average of the two middle numbers. Now, the trick is figuring out which two numbers are in the middle. To do this, you will first look at the Cum *f* column. Let's think about it. If there are 14 scores in the distribution, then the two middle scores are the seventh and eighth scores. If we start from the top of the Cum *f* column, we see that the seventh score is 2. The eighth score is 3. We add these two scores together, divide by 2 and come up with 2.5. This is the median of the distribution. We would go through a similar procedure if we were figuring Md from a cumulative percent column.

**REMEMBER!** The median is an X value. It's a score (e.g., number of TD passes) from the distribution or the average of two scores in the middle of the distribution. The median is NOT the *location* or *position* of that score. So, even though we know that the seventh and eighth positions in this distribution will figure into our calculation of the median, the median itself is NOT 7.5. It's the scores associated with the seventh and eighth positions that are added and divided by two. In this case, these are the scores of 2 and 3.

## Mode

The mode (symbolized as *Mo*) is the most frequently occurring value. For the data in the table above, the mode is 2 because that score has a greater frequency than any of the other scores in the distribution. If you go up a little further on this page and look at the table depicting touchdown passes per team in the NFL's 2000 season, you see that 18 is the mode. That is, more teams had 18 touchdown passes than any other number of touchdown passes during the 2000 season. It's really quite an easy call to make when the data are *discrete*.

| Range | Frequency |
|---|---|
| 500-600 | 3 |
| 600-700 | 6 |
| 700-800 | 5 |
| 800-900 | 5 |
| 900-1000 | 0 |
| 1000-1100 | 1 |

With *continuous data*, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see previous discussion of *continuous variables*). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. This table shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

The mode is the *least stable* of the three measures. For example, suppose that we send a group of children on two scavenger hunts. The distributions below represent the number of items that each child finds in each scavenger hunt:

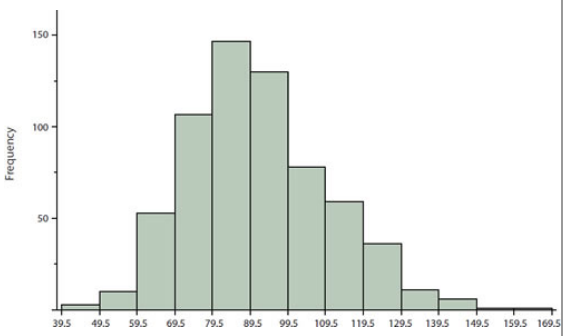
First Hunt: 3, 5, 6, 7, 7, **7**, 8, 11, 11, 14, 20

Second Hunt: 3, 5, 6, 7, 7, **11**, 8, 11, 11, 14, 20

The only difference between the two hunts is that the child who found 7 items in the first hunt, found 11 items in the second hunt. except that one child who previously found 7 items now finds 11. In the first scavenger hunt, the mode was 7. What is the new mode? It is 11, because there are now three 11s and only two 7s. This is a very big change in the mode, considering that most of the scores in the two hunts were the same, and the mean and median would not have changed by very much.

**How does the shape of a distribution figure into all of this?**

## How Shape Affects Mean, Median and Mode




(../../images/central_tendency/psych_test.jpg)How do the various measures of central tendency compare with each other? For *symmetric distributions* (i.e., bell-shaped), the mean, median, and mode are equal. Differences among the measures occur with skewed distributions. When a distribution is skewed, the mean will get pulled farthest toward the tail, the mode will be at the top of the curve and the median will be in the middle. Click any of the images to enlarge.

The figure here shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

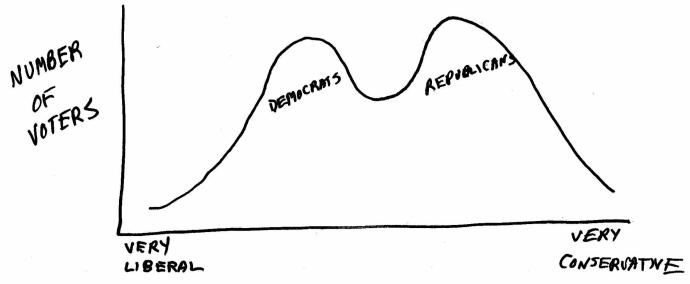
| Measure | Value |
|---|---|
| Mode | 84.00 |
| Median | 90.00 |
| Mean | 91.58 |

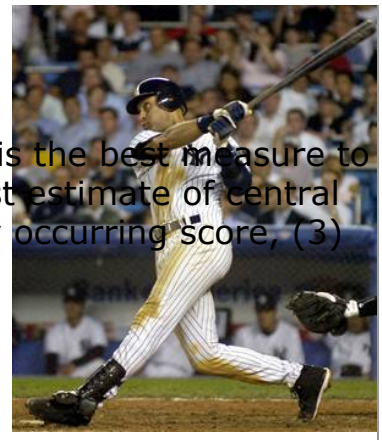
Measures of central tendency for these data are listed in this table. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions.

When we do run into a *bimodal distribution*, it becomes critical for the researcher to step back and re-examine his/her data. The thing is, a bimodal distribution underlying (../../images/central_tendency/bimodal.jpg)a single set of data actually suggests that the single set comprises two different populations. Another way to say it is that a bimodal distribution is really two overlapping distributions from two different populations. And at that point, trying to use single measures of central tendency is a bit inappropriate and misleading.

Suppose that we went out to see what voter turnout was like for a particular election, and we came out of it with a bimodal distribution. If we tried to describe the numbers with a group average or median, that numeric descriptor would fall in the valley--between the two humps. That hardly provides an accurate description of what's really going on. In this type of research, it would be better to analyze the data separately since the shape violates fundamental principles of central tendency. If we had conducted an experiment and came out of it with such data, we would really need to think about what two populations we might be dealing with and whether we needed to start all over again. :-(

## Which one to use and When?

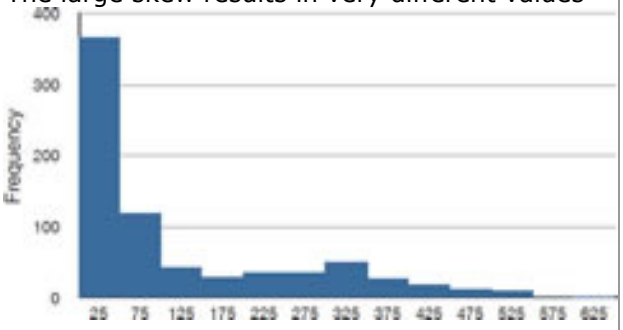




- (../../images/central_tendency/derek-jeter.jpg) The *mode* is the best measure to describe in three instances: (1) when you need the quickest estimate of central tendency, (2) when you want to report the most frequently occurring score, (3) when a rough estimate will do, or (4) when you have nominal scale data.
- The *median* is preferred when (1) you have a small, badly skewed distribution, or (2) there are missing or arbitrarily determined scores.
- The *mean* is the most useful of the measures of central tendency because many important statistical procedures depend on it. Also, the mean is the most stable of the measures from sample to sample. The sample mean is an *unbiased estimate* of the population mean.
- When in doubt...report all three and others, as well.

Here is an example of a badly skewed distribution. And yes, I understand that this has a lot to do with the New York Yankees' player salaries. Guess what...I don't care. Go Yankees! The distribution of baseball salaries (from the mid 1990s) shown in this other figure has a much more pronounced skew than the distribution in the first figure above. This histogram shows the salaries of major league baseball players (in tens of thousands of dollars: 25 equals $250,000.

This table shows the measures of central tendency for these data. The large skew results in very different values for (../../images/central_tendency/MLB_salaries.jpg)these measures. No single measure of central tendency is sufficient to summarize data such as these.




| Measure | Value |
|---|---|
| Mode | 250 |
| Median | 500 |
| Mean | 1,183 |

If you were asked the very general question: "So, what do baseball players make?" and answered with the *mean* of $1,183,000, you would have not told the whole story since only about one third of baseball players make that much. If you answered with the *mode* of $250,000 or the *median* of $500,000, you would not be giving any indication that some players make many millions of dollars.

Fortunately, there is no need to summarize a distribution with a single number. Don't feel like you have to. Tell the story of the data as completely as you can. That's your job! When the various measures differ, you should **report the mean, median, and mode**. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

**So which measure of central tendency would win in a fight?**

Well, now that you know everything there is to know about the different measures of central tendency, you should be able to provide a coherent rationale as to who would win in a fight...*M, Md* or *Mo*. Believe me, it can get nasty. Here is a picture of me after I tried to break up a small disagreement. I guess that if you're really going to have a good go at this, you will need a little more info. Here are the rules:


**1st RULE:** Bimodal distributions cannot fight.
**2nd RULE:** M, Md and Mo from N<15 are not allowed to fight.
**3rd RULE:** Only two measures to a fight.
**4th RULE:** No shirt, no shoes.
**5th RULE:** Md must be exactly at the 50th percentile.
**6th RULE:** M must be from a representative sample.
**7th RULE:** Interquartile range cannot fight.
**8th RULE:** When Big Sigma calls the fight, it's over.

Other than the above, anything goes--hair pulling, wedgies, prison rules. So, who would win? C'mon, it's a perfectly legitimate scientific question.

**Self Test**

- Self-test for chapter (self_test_central.pdf)
- Answers to self-test (answers_central.pdf)

This bulletin board is for you to communicate questions, concerns and even good vibes to other students as you read through the web book. The bulletin board is live. So you can post text, images, video, etc. When you do post text (on a sticky note), make sure that you choose a smaller font for your text. That will allow for more posts in the visible space, without having to click and drag the board.

Powered by lino

## Central Tendency

https://www.youtube.com/watch?v=uRGi4Bb9RYQ

Central Tendency Video

back to top

(index.html)

Some content adapted from other's work. See home page for specifics.

LAST UPDATED: 2014-10-13 1:29 PM

Mesa Community College | 1833 W. Southern Ave. Mesa, AZ 85202 | E-mail Address: dborman@mesacc.edu | Phone: (480) 461-7181 |


DEREK BORMAN: **PSYCHOLOGICAL SCIENCE**
MCC **PSYCHOLOGICAL SCIENCE** HOMEPAGE