How likely is it that the next person you meet walking down a street has more than the average number of legs?

General Education Statistics
Class Notes
Measures of Central Tendency: Mean, Median, and Mode (Section 3.1)

Let's say we have a data set, like household incomes, ages, scores on an exam, or heights of giraffes on the Serengeti. How can we summarize them so we get the bigger picture?

We will see three ways to measure the "center" of this data. They are the **mean** (usually what is called the average but not always), the **median** (the middle number of the data), and the **mode** (the most common value).

Definition: The **arithmetic mean** of a variable is computed by adding all the values in the data set and dividing by how many numbers you had.

Because we usually have a population and a sample to concern ourselves with, we need two different symbols for the *population* mean (μ, pronounced "mew") and the *sample* mean ($\bar{x}$, pronounced "x bar").

Which is considered a parameter and which is a statistic?

Let's get a bit technical with this definition. We say we add up the values and divide by how many numbers we have, right? In math speak, for μ, that looks like

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\Sigma x_i}{N}$$

The Σ (Greek letter sigma) is shorthand for "add". The subscripts of "i" simply mean the 1st, 2nd, 3rd, etc. values in the list. Recall, we say there are N obervations in the population.

When we see this for $\bar{x}$, we will use n for the sample size. That will look like

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x_i}{n}$$

We will always round to one more decimal place than that in the raw data.

To find a mean, the data must be quantitative. Imagine trying to average the responses to the question "What is your favorite movie?"

expl 1: Let’s try this out. The following data represent the travel times to work (in minutes) for *all* seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

a.) Find the mean of these numbers. Should you label it μ or $\bar{x}$?

b.) Let’s say we sampled from this population and got the four numbers 23, 23, 5, and 43. Find the mean of these numbers. Should you label it μ or $\bar{x}$?

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.

expl 2: Line up the data values from example 1 in increasing order and find the middle value. Label it as M.

Think of the median as that on a highway, right down the middle.

When you are told that the **“average”** value is such-and-such, what does that mean? Sometimes this refers to the mean and sometimes this refers to the median. Often, more digging is required to see which is intended.

Like the mean, the data must be quantitative to find its median. Imagine trying to find the median of responses to the question “What is your favorite movie?”

expl 3: Let's change this example up a bit. What if an eighth employee joins this web development company? Find the median now.

23, 36, 23, 18, 5, 26, 43, 33

There is no middle number. So what do you think you should do?

Instructions for TI Calculators:

expl 4: A company pays its employees the following salaries.

\$25,000	\$26,000	\$26,000	\$27,000	\$28,000
\$30,000	\$30,000	\$35,000	\$36,000	\$200,000

a.) Find both the mean and median of this data. Do this on the calculator. Here's how.

Enter the data values in column **L1** in the **STAT** editor. We do this by pressing the **STAT** button and then selecting **EDIT > 1: Edit…** from the menu. If necessary, clear out any data in **L1** by arrowing up to the column heading and pressing **CLEAR**. When you arrow back down, any data should be gone. Enter the values of the salaries in **L1,** pressing **ENTER** after each one.

Then press the **STAT** button again. But this time, arrow over to select **CALC > 1: 1-Var Stats**. That will put this expression on the home screen. Press **ENTER** and the calculator will fill with many statistics. (Some newer calculators will have an intermediate screen, where you need to select **L1** for **List** and clear out any entry in the **FreqList:** row. Arrow down and select **Calculate**.)

Look for $\bar{x}$ and record it here. (The calculator will call this $\bar{x}$ even if you know the data is from a population.) Give it a dollar sign and comma.

Arrow down and you will see "Med=" which is the median. Record it here, with a dollar sign and comma.

expl 4b.) If you wanted to stress the company's great salaries to prospective employees, which "average" would you provide? Why?

4c.) Why do you think the mean and median are so different?

4d.) Give a likely explanation for the outlier salary of $200,000.

Definition: Outlier: A data value that is far from the other values.

Later, we will learn a way to determine, for sure, if a value is an outlier. For now, we will use this layman's definition.

Definition: Resistant: A numerical summary of data is said to be **resistant** if extreme values (very large or small) relative to the data do *not* affect its value substantially.

Considering the data above, would you say the mean or the median is resistant? Why?

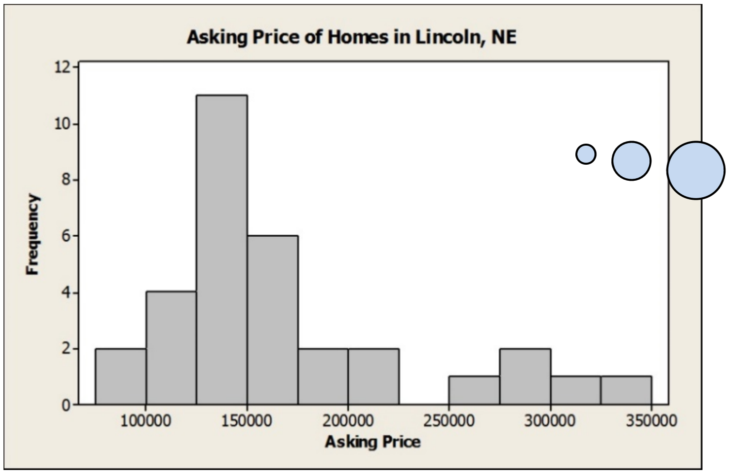
The Distribution of a Variable and Means versus Medians:

Let’s see how the mean and median are affected by a variable’s distribution.

expl 5: The data below represent the asking price of homes (in dollars) for sale in Lincoln, NE.

Asking Prices of Homes in Lincoln, Nebraska (dollars)			
79,995	128,950	149,900	189,900
99,899	130,950	151,350	203,950
105,200	131,800	154,900	217,500
111,000	132,300	159,900	260,000
120,000	134,950	163,300	284,900
121,700	135,500	165,000	299,900
125,950	138,500	174,850	309,900
126,900	147,500	180,000	349,900

The mean is $168,320.
The median is $148,700.



Here is the histogram of this data. It uses a class width of $25,000. Which classes hold most of the data?

How would you describe the distribution? In other words, is it symmetric, skewed right, or skewed left?

Since the mean is less resistant than the median, it is pulled up by the large amounts seen on the right of the histogram. Draw vertical lines on the histogram to mark the mean and median.

The mean is the point where the histogram is perfectly balanced.

expl 6: Let’s return to a riddle asked at the very beginning of these Notes. How likely is it that the next person you meet walking down a street has more than the average number of legs? It is common to imagine the “average” to be two legs. Would that number be the median or the mean? In that case, it would be *very unlikely* to come upon a person with more than this “average” of two legs.

But if we use the mean for “average”, what would you guess its value to be? Make an educated guess and write an approximate number. What, now, is the likelihood of coming upon someone with more legs than average?

When an average is presented to you, always ask if is it the mean or the median.

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

A set of data can have no mode, one mode, two modes (**bimodal**) or more than two modes (**multimodal**). *If no observation occurs more than once*, we say the data have **no mode**.

Modes, unlike means and medians, can be found for qualitative data. You could find the mode of responses to the question “What is your favorite movie?”

expl 7: Mr. Kramer gets a yearly evaluation from his students. Using a scale of [strongly agree, agree, neutral, disagree, strongly disagree] students were asked which most fits their level of agreement to the statement “The teacher is fair.” The replies are listed below. Make a frequency chart and determine the mode.

strongly disagree	disagree	neutral	strongly agree
agree	strongly disagree	disagree	disagree
agree	strongly disagree	strongly agree	strongly agree
agree	disagree	strongly agree	strongly agree

expl 8: I have thirteen frogs whose weights I measured. The mean weight was 0.80 ounces. I then noticed a runaway frog and so scooped him up and measured him at 0.65 ounces. Find the mean weight of all fourteen frogs. Round to the nearest hundredths place.

expl 9: I had ten salamanders whose lengths (in centimeters) I measured. The data lined up like the following.

10 10 11 12 13 --- 16 16 19 20

a.) The problem is that I know the median is 14 but the missing length was unfortunately obscured by salamander doo-doo. What is the missing measurement?

b.) What is the mode of all ten measurements?

c.) Draw a picture of what you think salamander doo-doo looks like.

Instructions for STATCRUNCH:
Within MSL problems, you will see a little icon that looks like overlapping rectangles next to the data. Click on it and select “Open in StatCrunch”. This will open StatCrunch and import the data. Alternatively, if you have your own data to enter, open StatCrunch from the left-hand MSL menu and make your way to the spreadsheet. Enter the data in column 1 and label it if you want.

Select **Stats** > **Summary Stats** > **Columns**. You will need to tell it where the data is (“Select column(s)” at top). By default, it will calculate lots of stuff including stuff we have not covered yet. You can select more to display under “Statistics”. If you just need mean, median, and mode, select “Mean” and scroll down to Control-click “Median” and “Mode”. You will see those selected items appear to the right of the selection list. Press the “Compute!” button and it will output a little window with the results.