David Youngberg
BSAD 210—Montgomery College

LECTURE 04: CENTRAL TENDENCY

I. Data Descriptions

a. When examining data, one of your first steps should be to familiarize yourself with its statistics. First among these statistics is the data's *central tendency*—a single value which describes the center point of the data set. It can be described in three different ways: mean, median, and mode. But all three of them have issues.

b. *Mode* is the most common value. It's often used for data organized into discrete categories with few alternatives; this is also called *categorical data*.

i. Problem: Difficulty with continuous variables (e.g. income, though you can transform that data into a range).

ii. Problem: May also mask important changes (e.g. many poor people enter country).

iii. Problem: There may be more than one mode.

iv. The mode, it seems, is rarely used because it has so many problems. But in fact modes are used whenever you examine a pie chart or a bar graph.

c. *Mean* (or the arithmetic mean) is the average. Sum all the values and divide by the number of observations.

d. *Median* is the middle value. Half of the observations are below and half are above (if an even number of observations, take the mean of the two middle observations).

e. Mean vs. median

i. The mean and the median are good at different things.

ii. Using the average can give you a distorted understanding of the typical observation thanks to *outliers* (unusually high or low observations). The median can be better because it treats outliers as the same as non-outliers; observations are just high or low.

1. For example, the average student loan debt in 2010 was $17,916. It's so high because it includes graduate students like doctors and lawyers. While they're a relatively small segment of the borrowing population, they take out huge amounts—often over $100,000—and

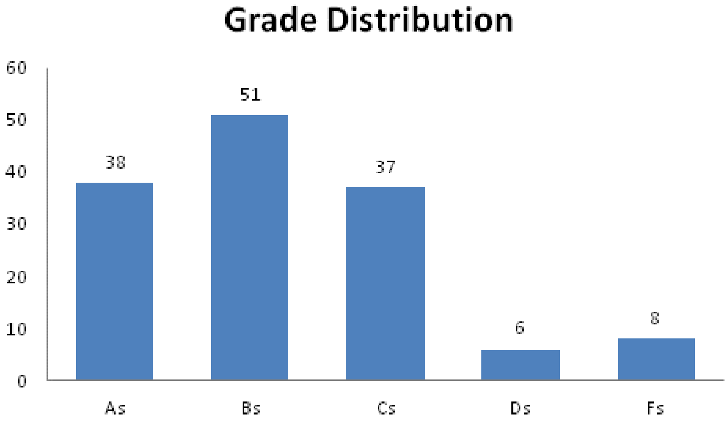
that throws off the average. Median student debt is much lower: $8,500 in 2010.[1]

iii. But precisely because the median treats a very high value and a somewhat high value as the same (both are in the upper half of distributions of observations), it can be deceptive.

1. If wealthy people are getting wealthier but no one else is, median wealth wouldn't change but mean wealth would increase.
2. You'd have a much better idea how your store is doing if you know the mean amount of money customers spend rather than knowing the median amount.
3. Most Americans don't smoke; the median number of cigarettes per week is zero. You wouldn't be able to distinguish this society from one where literally no one smokes. But you could if you used the mean.

II. Example: Grades

a. Below is a graph of all the grades I assigned in the spring of 2014. If we assign a value of "4" to each A, "3" to each B, etc, what is the mean, median, and mode of this data?



b. The mode is an easy one: the most common value here is 3, or a B.

c. The median is a little harder: since there are 140 grades here, the 70th grade (counting from the highest down or the lowest up) is 3, or a B.

d. The mean takes a few steps:

i. First, we must multiply the number in each grade by the value:

1. 38 x 4 = 152
2. 51 x 3 = 153
3. 37 x 2 = 74
4. 6 x 1 = 6

[1] http://www.brookings.edu/~/media/research/files/reports/2014/06/24-student-loan-crisis-akers-chingos/is-a-student-loan-crisis-on-the-horizon.pdf

5. 8 x 0 = 0

ii. Second, we add them together: 152 + 153 + 74 + 6 + 0 = 385.

iii. Third, we divide: 385 / 140 = 2.75

e. Which central tendency is most useful here? Why do you think it turned out that way?

III. Example: U.S. Income

a. The mean household income in the United States in 2012 was $71,274. (For individuals, it's $40,563.)

b. The median household income in the United States in 2012 was $51,017. (For individuals, it's $26,989.)[2]

c. Why this gap?

i. Below is the distribution of U.S. household income by income bracket. For example, 13.0% of Americans in 2012 had an income below $15,000.

Under $15,000	$15,000 to $24,999	$25,000 to $34,999	$35,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 to $149,999	$150,000 to $199,999	$200,000 and over
13.0	11.7	10.7	13.6	17.5	11.7	12.5	5.0	4.5

d. What's a more useful way of determining the central tendency? It really depends on what you want.

i. Median is better for describing what's "typical."

ii. Mean is a better summary of the central tendency when each observations' exact value is important, rather than just knowing what's high or low.

IV. Geometric Mean

a. When taking the average of growth rates, it's helpful to calculate the average differently. To understand why, consider the annual sales growth rate of a company. Last year it was 1% and the year before that it was 9%. If sales started at $100,000, what are the sales now?

i. A 9% increase in sales means sales grew by $9,000; it became $109,000.

ii. A 1% increase in sales means sales grew by $1,090; it became $110,090.

iii. In other words, sales went from $100,000 to $110,090. Or:

($100,000)(1.09)(1.01) = $110,090

[2] http://finance.townhall.com/columnists/politicalcalculations/2013/09/29/what-is-your-us-income-percentile-ranking-n1712430/page/full

Note the use of adding “1” to the growth rate. That way we not only include what’s being added but also what we started with.

b. We can simplify the approach with this equation:

$$New\ result = Starting\ amount \times \prod_{i=1}^{n}(1 + x_i)$$

   i. The giant pi symbol means multiply;
   ii. The “x’s” are the growth rates, expressed as a decimal;
   iii. The “i” means you’re considering the ith rate;
   iv. The “N” means there are that many rates to consider.
   v. In our example, N was two, x_1 was 0.09 and x_2 was 0.01.

c. Now suppose we claimed the average growth rate was 5%. That means if the growth was five each year, we should get the same total sales. But we don’t.
   i. ($100,000)(1.05)(1.05) = $110,250.
   ii. We got a higher number than before. It may seem close enough, but keep in mind it should be *exactly the same* and we were only using two years. If you repeated this example using ten or twenty years of data, we’d be way off.
   iii. Using the “arithmetic mean” on growth rates results in overstating the average growth rate. We have to use the geometric mean.

d. Here’s the equation for the geometric mean:

$$Geometric\ Mean = \sqrt[n]{\prod_{i=1}^{n}(1 + x_i) - 1}$$

   i. Rather than adding all the observations up and dividing by the number of observations, we’re multiplying all the observations together and then taking the Nth root. Note how similar this is
   ii. So our growth rate is:

$$Geometric\ Mean = \sqrt[2]{(1.09)(1.01)} = \sqrt[2]{(1.1009)} \cong 1.0492 - 1 = 0.492$$

   iii. A more accurate growth rate would be just over 4.92%.

V. Weighted Average

a. Sometimes some observations matter more than others and you want to give more emphasis to those when finding an average.

b. That additional emphasis is called a "weight." Using a normal average, all observations have equal weight. With a weighted average, each observation has a weight assigned to it. That observation's value is multiplied by the weight before being added. You then divide not by the total number of observations but by the sum of the weights:

$$Weighted\ Average = \frac{\sum_i^n (w_i x_i)}{\sum_i^n w_i}$$

i. Where x_i is the ith observation; and

ii. w_i is the weight for that ith observation.

c. Examples:

i. Your grade. Each assignment may have the same possible points but each one is weighed as indicated in the syllabus. I use the equation above to calculate your final grade.

ii. Stock markets. Some stock market indicators, like the S&P 500, weigh the price of each company's stock by how many stocks of that company exist.

iii. Slugging average. This baseball statistic measures how many bases a single player gets to when the batter hits the ball. The number of times a player can pass three bases has more weight than the number of times a player can make it to only one base.

d. Typically weights are less than one and the sum of all weights equal one (thus dividing is not necessary because you'll just be dividing by one). But not always.

i. Sometimes while the weights equal one, not all the information is there and you end up dividing by the sum of all the weights for the information you have. I do this when a student wants to know his/her current grade before every assignment is complete.

ii. Other times the weights are quantities. The example from II was technically a weighted average. I multiplied each value by the weight and then divided the whole thing by the sum of the weights.

VI. Mean, Median, Mode

a. This video covers mean, median, mode, and standard deviation.

b. Excel has the equations for mean, median, and mode built into it.

   i. For mean, type “=AVERAGE” and press the TAB key. Then highlight the cells you wish to find the average of and press ENTER.

   ii. For median, type “=MEDIAN” and press the TAB key. Then highlight the cells you wish to find the median of and press ENTER.

   iii. For mode, type “=MODE” and press the TAB key. Then highlight the cells you wish to find the mode of and press ENTER.

c. Open Dataset 2 which contains historic data on Disney’s total sales and sales by division. (Under Disney Sales sheet.)

d. Which division has the most revenue growth for a particular year? That’s indicated in Column B and to answer that question we need to find the mode.

   i. The mode is a mathematical operation so Excel needs to analyze numbers. I assigned each division a number and indicated the appropriate number in Column C.

   ii. In A22, type “Mode” and then type “=MODE(C5:C21)” in C22. You should get 3, or Parks and Resorts.

e. In A23, type “Mean” and find the average sales of the five different divisions plus the total sales average.

   i. For example, in D23 type “=AVERAGE(D4:D21)”

   ii. Note once typed, you can select and copy the cell and then paste it in the appropriate places; the references will update automatically.

f. In A24, type “Median” and find the median sales of the division plus the total.

   i. For example, in D24 type “=MEDIAN(D4:D21)”

VII. Geometric Mean

a. Here’s a video of this section (the technique is slightly different).

b. Suppose we’re also curious about the company’s average growth rate. To find the geometric mean, recall the best way is to add 1 to all observations, take the geometric mean, and then subtract one.

   i. In P5, type “1+O5”. Because all the growth rates are displayed as percents, you should get 1.02. You can increase the decimal places displayed with the button in the Number section. But you don’t have to do this; Excel knows those values are there even if they are not shown.

ii. Now double-click the square in the lower-right corner of the selected box.

iii. In P23, type “=GEOMEAN(P5:P21)-1”. You may want to click the % button to the left of the decimal button and increase the decimal. You should get about 4.67%.

iv. Note if we took the arithmetic mean, you’d get about 4.75%. That doesn’t sound like much of a difference, but if you assumed 4.75% growth every year starting in 1998, you’d get about $600 million more in sales than you’d actually have. But with 4.67%, you are exactly correct.

VIII. Weighted Average

a. In the next sheet, under Performance Review, there is a hypothetical Disney employee named Andy. Suppose Disney determines if people get a raise based on three criteria: how often they take sick days (Attendance), what customers say about them (Customer Reviews), and their sales (Sales).

b. Each criterion has weights, as indicated, and Andy gets the scores as indicated. Each is out of 100.

i. Andy is a great salesperson but he’s also very rude to a few people; he has a fair number of complaints.

c. Suppose Disney requires a score of 85 to qualify for a raise. To determine if Andy gets it:

ii. In D3, type “=B3*C3”. Note we are multiplying the weight by the score.

iii. Copy this cell in D4 and D5.

iv. In D6, type “=SUM(D3:D5)”. You should get a result of 87. Andy (barely) gets his raise.

d. You can also divide the result in D6 by the sum of all the weights. This isn’t important do to explicitly here as the sum of the weights equal 1, but it’s what you do when they don’t equal 1.