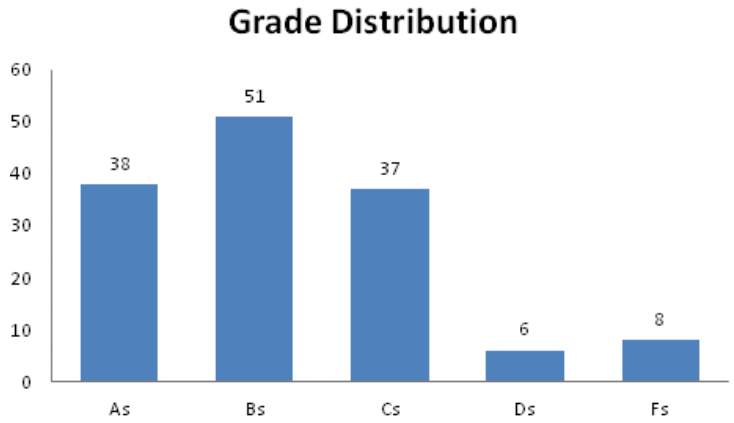
David Youngberg
BSAD 210—Montgomery College

LECTURE 03: 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.
      i. *Mean* (or the arithmetic mean) is the average. Sum all the values and divide by the number of observations.
         1. Problem: Influenced by unusually high or low data points (e.g. high incomes). These unusual data points are called *outliers*.
      ii. *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).
         1. Problem: Not influenced by some important changes (e.g. the very poor lose half their income).
      iii. *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*.
         1. Problem: Difficulty with continuous variables (e.g. income, though you can transform that data into a range).
         2. Problem: May also mask important changes (e.g. many poor people enter country).
         3. Problem: There may be more than one mode.
   b. 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.

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
        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?

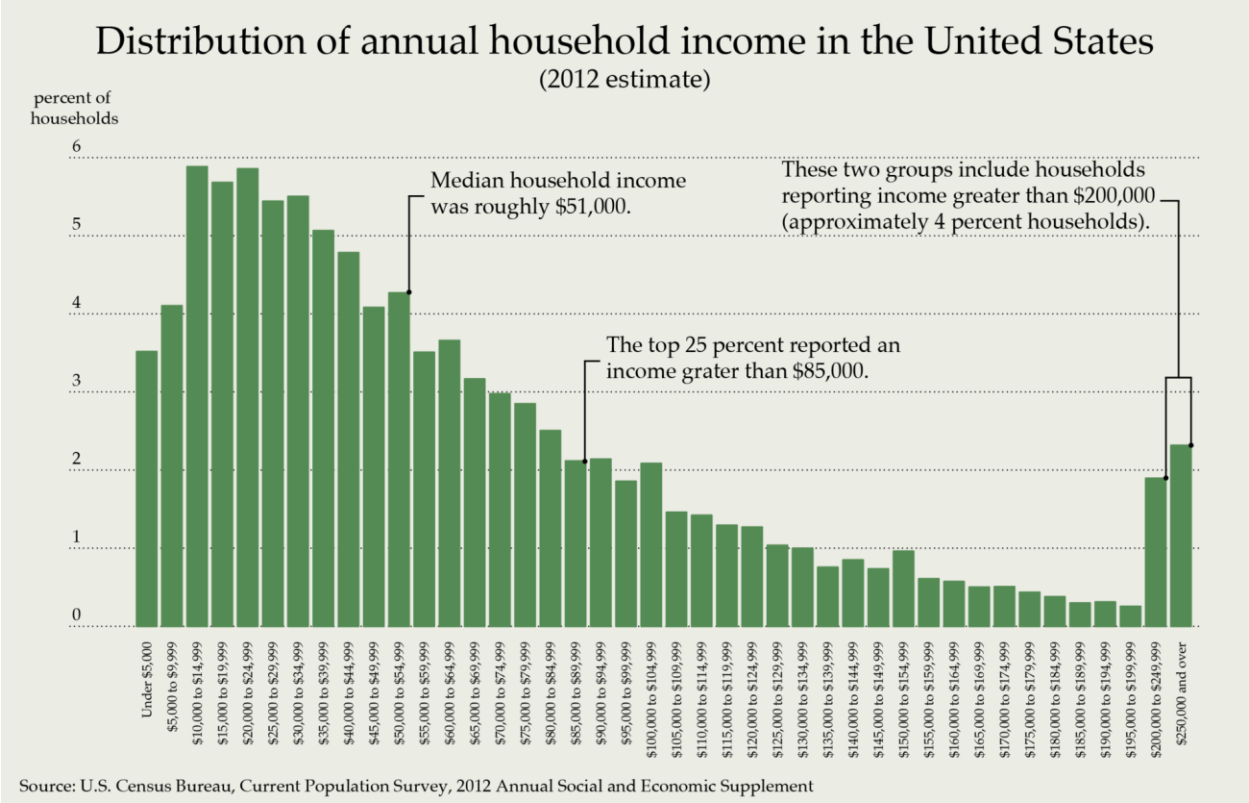
III. Example: U.S. Income

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.)[1]
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

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

d. What's a more useful way of determining the central tendency? Why?
    i. This distribution of U.S. income might be helpful in answering this question.



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

    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 (you might remember a similar symbol that looks like an "E" to mean add);
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 year. If you repeated this example using ten or twenty years of data, we'd be way wrong.
   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%.