David Youngberg
BSAD 210—Montgomery College

LECTURE 05: LAW OF LARGE NUMBERS AND CENTRAL TENDENCY

I. Law of Large Numbers
   a. One of the basic rules of statistics is the *law of large numbers*, or as the number of observations increases, the empirical average will approach the theoretical average.
   b. Example: Coin flipping
      i. The theoretical probability of getting "heads" on a coin flip is 0.50.
      ii. If you flip a coin once, you'll get either heads or tails. That means the empirical probability of getting "heads" is either 1.00 or 0.00. That's way off!
      iii. Let's flip it twice. Here are the possible results:

Result	Chance of Heads
HH	1.00
HT	0.50

Result	Chance of Heads
TH	0.50
TT	0.00

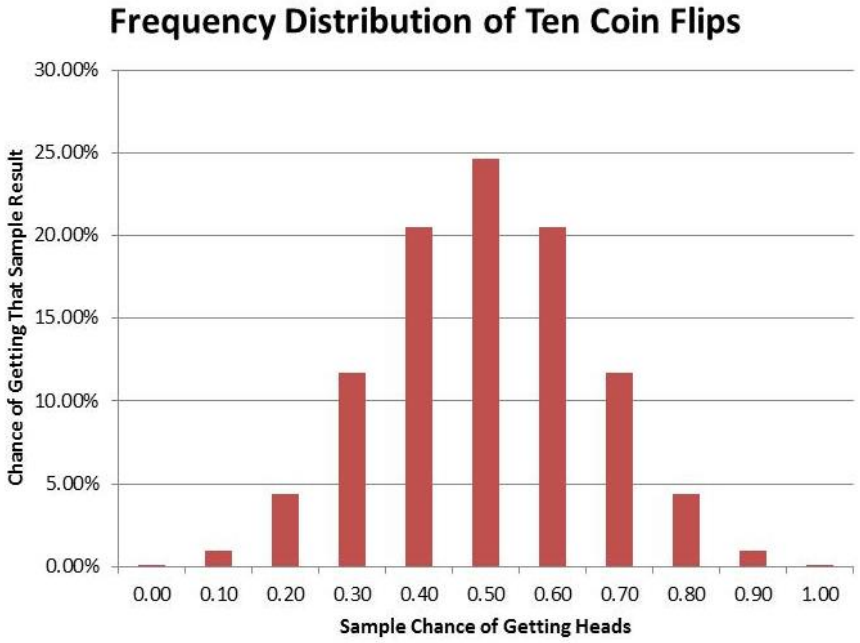
      iv. Now you have a 50% chance of getting the theoretical result and a 50% of getting an extreme result.
      v. Let's flip it four total times. Here are the possible results:

Result	Chance of Heads
HHHH	1.00
HHHT	0.75
HHTH	0.75
HTHH	0.75
THHH	0.75
HHTT	0.50
HTHT	0.50
THHT	0.50

Result	Chance of Heads
HTTH	0.50
THTH	0.50
TTHH	0.50
TTTH	0.25
TTHT	0.25
THTT	0.25
HTTT	0.25
TTTT	0.00

      vi. You may only have a 37.5% chance of getting the theoretical result, but you have only a 12.5% chance of getting one of the extreme results. With the mid-range results each at 25%, the theoretical result is the most likely result to get.

vii. And if you flipped the coin ten times…



II. Gambler's Fallacy

a. It's tempting to be fooled by the law of large numbers. If black comes up ten times in a row on a roulette wheel, people think that red must be "due." The thinking is that it must be more likely to come up in order to balance out the previous streak. Otherwise, how could we say increasing the sample size brings the sample mean closer to the theoretical mean?
    i. This is called the *gambler's fallacy*—believing **under-**represented results will be more likely to occur in future independent trials.
b. But look at our bar graph of ten coin flips: we get ten heads (or ten tails) 0.10% of the time. In other words, it's possible that a streak can continue.
c. The law of large numbers doesn't render independent trials dependent. The roulette wheel has the same chance of getting black if black came up ten times in a row or red came up ten times in a row.
d. On August 18, 1913, black came up twenty-six times in a row at the Monte Carlo casino. People bet (and lost) millions on the idea that red "was due" for a streak. But any particular sequence of red and black is just as likely as all black.
e. Note the gambler's fallacy does not apply to dependent events, such as card-counting.

III. Hot Hand Fallacy
   a. People sometimes succumb to the opposite of the gambler's fallacy, called the *hot hand fallacy*—believing **over**-represented results (particularly successes) will be more likely to occur in future independent trials.
   b. Success now does not mean success later. Just because a basketball player made three shots in a row does not mean they are suddenly more likely to make a fourth shot.
      i. This fallacy is a bit harder to detect when it comes up in games based at least partly on skill.
   c. Most (but not all) of the evidence analyzing basketball players' success at shooting suggests the result for any given shot (for a particular player) is random.
      i. These studies tend to focus on free-throws, where you can remove complexities like where the shot was taken from or what the other team is doing.
   d. If you're doing well at craps, that doesn't mean your next roll of the dice will be successful.
   e. If the slot machine you're using pays out, that doesn't mean it will continue to pay out.
   f. Just because a stock is doing well now doesn't mean it will continue to do well.
      i. The day-to-day and hour-to-hour movement of a particular stock is essentially random. At any particular price, it has an equal chance of going up and of going down.

IV. 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. <u>Problem</u>: Difficulty with continuous variables (e.g. income, though you can transform that data into a range).
      ii. <u>Problem</u>: May also mask important changes (e.g. many poor people enter country).
      iii. <u>Problem</u>: 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.

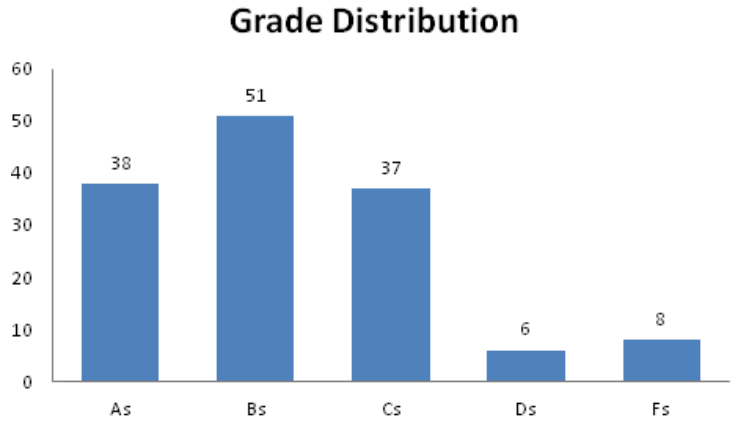
d. *Median* is the middle value. Half of the observations are below and half are above (if an even number of observations, take the mean of the two middle observations).

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.

V. Example: Grades

[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

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?

VI. 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.

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

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.