David Youngberg
BSAD 210—Montgomery College

LECTURE 04: 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 tends to 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.

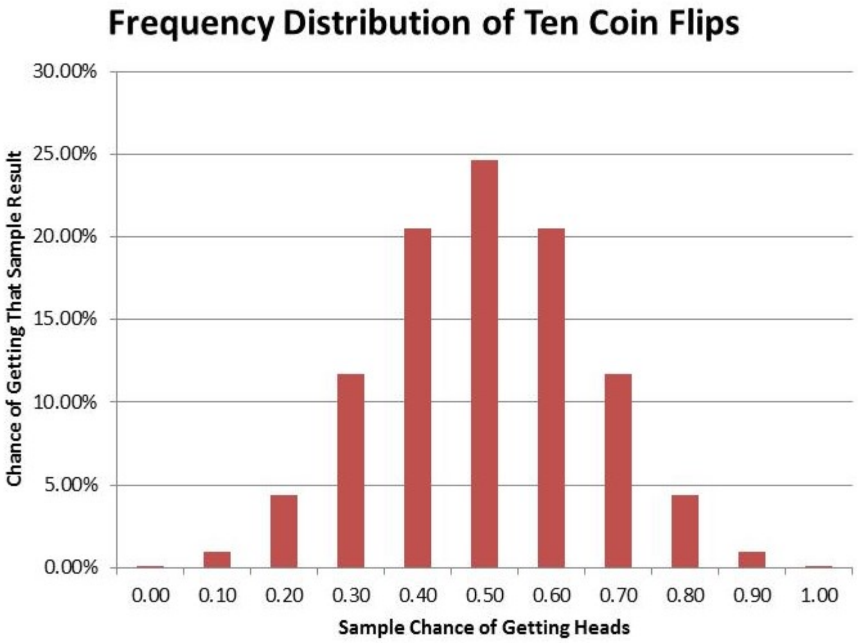
v. Let's flip it four total times. Here are the possible results:

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…



c. Suppose you had a jar of 10 squares of paper. On 8 eight squares was the number 100. One 1 square was the number 0. On the last square was the number 200. Note on average, the value you should get is 100.
   i. Suppose you draw a square of paper and it's 200. You record it and replace the square.
   ii. Now you draw another square of paper. What's the chance that it's 200 again? 10 percent. What's the chance that it's less than 200? 90 percent.
   iii. Precisely because the 200-valued square is so unusual, it is unlikely to happen again. Strange values (either very small or very large) happen because of unusual circumstances and, therefore, they are unlikely to be repeated.
   iv. Note that this means the average from the first pull was 200. One observation at 200. The average from the second pull is likely to be less. Suppose it's 100. That means the average of two pulls is 150. We're getting closer to the theoretical mean of 100.
   v. As we pull more and more squares, pulling the unusual values (both high and low ones) will continue to be unlikely and we will arrive at the theoretical average.

II. 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).

III. Mean vs. median

a. Thanks to *outliers* (unusually high or low observations), the mean and the median are good at different things.

i. The median is best when you're interested in what's "typical." For example, if you become a civil engineer the median gives you a good idea of what salary you'd make.

ii. The mean is best when you're interested in the "big picture" and you want to include outliers. For example, knowing mean bill for each table in a restaurant is much more helpful than knowing the median bills. You *want* to include the full effect of outliers to account for the occasional big spender. It gives you a better idea of how much money your restaurant is making.

b. The median can be better because it treats outliers as the same as non-outliers; observations are just high or low. Since you're unlikely to be an outlier (by definition), having their influence reduce can sometimes be an advantage.

i. For example, the mean individual income in the United States in 2021 was $57,143. In contrast, the median individual income in the United States in 2021 was just $37,522.[1] Again, the massive difference between these values is due to outliners.

c. 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. Sometimes you want the outliers.
    i. If high-income people are earning more money but no one else is, median income wouldn't change but mean income would increase.
    ii. 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.
    iii. 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.[2]

ON AVERAGE, PEOPLE SWALLOW ABOUT 8 SPIDERS A YEAR
WELL ACTUALLY, MOST PEOPLE SWALLOW ZERO
@MYGUMSAREBLEEDING
BUT THAT GUY EATS LIKE 40 SPIDERS A DAY
HE'S THROWING THE WHOLE AVERAGE OFF

d. In other words, including the value of outliers is both good and bad; it depends on what you're interested in.
    i. Median is better for describing what's "typical."
    ii. Mean is a better when you want to know the "big picture."

IV. <u>Example</u>: Restaurant bills

a. Restaurant customers tend to care about the median bill. They want to know what they can expect to spend, so they want to know what's typical, so they want a measure that diminishes the impact of outliers—the occasional big group or celebrating table or customer that just always gets a meal with appetizers, wine, and dessert.

[1] Source: U.S. Census, retrieved by St. Louis Federal Reserve Economic Database

[2] I found the spider comic on Reddit. The "8 spiders a year" statistic is made up. You'd also have to eat way more than 40 spiders a day to bring the average up to 8 a year, unless the "people" mentioned in the first panel only refers to a subset of the population. Fun math question: how many people would have to be in that subset in order for one person to throw off the average as the comic describes?

b. Restaurant owners tend to care about the average bill. They want the central tendency to fully reflect the impact of the rare, but very lucrative, bill. Given how tight restaurant profits are, the difference between the mean and median is likely the difference between making money and losing money.
c. Statisticians focus on the mean—it'll be in our equations, rather than the median—because we focus on the big picture of what the data are telling us.