David Youngberg
BSAD 210—Montgomery College

# LECTURE 06: DISPERSION

I. Range and Standard Deviation
   a. The most basic way to describe dispersion is the data's *range*, or the difference between the highest value and the lowest value.
      i. For example, the range of grade data is between an "A" (4) and an "F" (0), or 4.
      ii. This is obviously a very limited way to describe data—did a lot of people get the lowest grade or just one—so we turn to standard deviation.
      iii. Quartiles give you a little more information. A quartile represents one-fourth of the data.
   b. *Standard deviation*—expressed in the same units of data and describes the level of variation of the data.
      i. For samples, standard deviation is calculated as such:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

   c. There are other measures of dispersion—variance and standard error.
      *i.* *Variance*—the standard deviation squared; it is indicated as $s^2$.
      *ii.* *Standard error*—the standard deviation divided by the square root of the sample size.

II. Population
   a. The standard deviation of a population is similar:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

   b. Note the notation is different: the sample standard deviation uses "s" while the population standard deviation uses "σ" and "x-bar" is replaced with μ.
      i. Similarly, sample variance is "$s^2$" and population variance is "$\sigma^2$."

c. Note also this equation uses "*N*" rather than "*n* – 1." Why does the sample standard deviation use a different value? Suffice to say it's a quirk of the mathematics.

III. Quartiles, Min, Max Practice
   a. This video covers quartiles.
   b. Let's go back to **Data Set 1** and the stock market tab.
   c. Use the "=QUARTILE.INC" function to find a particular quartile.
      i. The QUARTILE.EXC function you can ignore for this class; it won't be able to find the $0^{th}$ and $4^{th}$ quartile.[1]
      ii. Once you type in the function, Excel will ask you for the array, or range of data. Highlight the appropriate area.
      iii. Press the comma "," key.
      iv. Now it will ask you the quartile you wish.
         1. 0 is the zeroth quartile, or minimum
         2. 1 is the first quartile
         3. 2 is the second quartile, or median
         4. 3 is the third quartile
         5. 4 is the fourth quartile, or maximum
      v. Press ENTER
      vi. For example "=QUARTILE.INC(B4:B63,1)" will tell you that the first quartile of stock prices is $93.10; about 25% of monthly stock prices are below $93.10.
   d. You can also use the min and max functions to find the largest and smallest values.
      i. "=MAX" finds the maximum value in an array.
      ii. "=MIN" finds the minimum value in an array.

IV. Dispersion Practice
   a. Excel doesn't have a range function because it's often more useful to report the maximum and minimum values.
      i. In A66, type "Max" and type "=MAX(B4:B63)" in B66. Repeat this for the other columns. (Don't forget to convert to percents for the growth rates.)
      ii. In A67, type "Min" and type "=MIN(B4:B63)" in B67. Repeat this for the other columns. (Don't forget to convert to percents for the growth rates.)
      iii. In A68, type "Range" and type "=B66-B67" in B68. Repeat.

[1] What's the point of this function, then? Some argue that it's better finding the "interquartile range" (the difference between the $3^{rd}$ quartile and the $1^{st}$ quartile) when there's an even number of observations.

b. Standard deviation is a really common function; the equation's built right in.
    i. In A69, type "Standard Deviation" and type "=STDEV.S(B4:B63)" in B69. Repeat for the other columns.
    ii. Note this calculates the standard deviation for a sample. Type "=STDEV.P" for the standard deviation of the population. But the population version is rarely used; use the sample version.

V. Coefficient of Variation

a. It's often useful to compare which sample has more variation. For example, which stock is more volatile? Which medical treatment results in a more consistent blood pressure? Which basketball player regularly makes free throws?
b. It's not simply a matter of which sample has a higher standard deviation.
    i. Consider two store locations: one with a lot of pedestrian traffic and one with a little. The high-traffic location probably has more sales because more people walk by.
    ii. Factors which affect traffic are magnified at the high-traffic location. If bad weather cuts the number of pedestrians in half, the high-traffic location will have a much larger drop in the raw number of pedestrians than the low-traffic location.
c. The *coefficient of variation* (CV) corrects this problem by adjusting standard deviation with mean. In general, higher means mean higher standard deviation. By adjusting for mean, you can compare two different samples or populations even if the means are very different.

$$CV = \frac{s}{\bar{x}}$$

    i. Note that CV is expressed as a percent.
d. CV has many applications. For example, it's used in lab testing to make sure a test generates consistent results. It was referenced in *Bad Blood: Secrets and Lies in a Silicon Valley*, by John Carreyrou, about the true story of Theranos, a company that pretended it could run thousands of blood tests with their "Edison" devices using only a single drop of blood. Founders Elizabeth Holmes would eventually go to prison for wire fraud.

> Soon, there were other things that began to trouble [future whistleblower] Tyler. One type of experiment he and [fellow analyst] Erika were tasked with doing involved retesting blood samples on the Edisons over and over to measure how much their results

> varied. The data collected were used to calculate each Edison blood test's coefficient of variation, or CV. A test [in biology] is generally considered precise if its CV is less than 10 percent. To Tyler's dismay, data runs that didn't achieve low enough CVs were simply discarded and the experiments repeated until the desired number was reached…Erika and Tyler might be young and inexperienced, but they both knew that cherry-picking data wasn't good science. (p186-187)

   i. Do ***not*** put emphasis on that 10 percent threshold mentioned here; that's a number specific to the situation. I'm just including this story as an example of how CV is used in the real world.

e. Consider the hypothetical weekly sales data of two different store locations (in thousands of dollars) below. Which location is more consistent?

| Week | High-Traffic | Low-Traffic |
|---|---|---|
| 1 | $80 | $9 |
| 2 | $60 | $6 |
| 3 | $40 | $3 |

   i. First, find the average: $60 for the first, $6 for the second.
   ii. Second, find the standard deviation of the samples: $20 for the first, $3 for the second.
   iii. Third, divide: 33.3% for the first, 50.0% for the second.
   iv. The high-traffic location has more consistent sales because its CV is lower.

VI. Coefficient of Variation Practice

a. Look back to Data Set 1. Which company has the most consistent stock price?
b. In A70 type "Mean" and then in B70 type "=AVERAGE(B4:B63)"
c. In A71 type "CV" and then type "=B69/B70" in B71. (Note this is the equation from the topic notes: standard deviation divided by mean.) Repeat for the sales of all divisions.
d. Note that at a standard deviation of $48.95, it looks like Berkshire-Hathaway has the least consistent stock price (because the standard deviation is the highest). But in the context of the average stock price, it's actually most consistent—its CV the lowest at just 20%.
e. Also note that Exxon-Mobil had the least consistent stock price (because its CV is the highest at 31%), even though its standard deviation was one of the lowest.