David Youngberg
BSAD 210—Montgomery College

LECTURE 08: DISPERSION

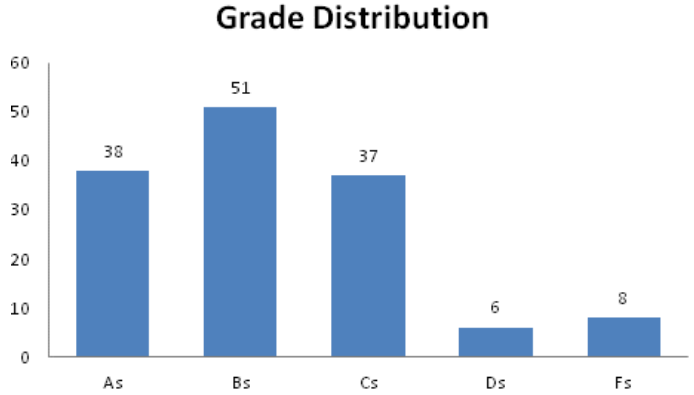
I. Range and Standard Deviation

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. You might wonder why you should learn the calculation for standard deviation if computers can do it for you. Sometimes you get summary data and can't get a computer to do it all for you. But we can use the equation to determine standard deviation.



   i. Let As=4, Bs=3, Cs=2, Ds=1, Fs=0
   ii. N = 38+51+37+6+8 = 140
   iii. The sum is 38(4) + 51(3) + 37(2) + 6(1) + 8(0) = 385

iv. The average, as before, is 385/140 = 2.75
v. Now we calculate: $38*(4 - 2.75)^2 + 51*(3 - 2.75)^2 + 37*(2 - 2.75)^2 + 6*(1 - 2.75)^2 + 8*(0 - 2.75)^2 = 59.375 + 3.1875 + 20.8125 + 18.375 + 60.5 = 162.25$
vi. Now we divide the result by 139 (one minus the sample size) and take the square root: $\sqrt{162.25/139} \cong 1.08$

d. 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 "σ^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 2 and the Disney tab.
c. Use the "=QUARTILE" function to find a particular quartile.
   i. Once you type in the function, Excel will ask you for the array, or range of data. Highlight the appropriate area.
   ii. Press the comma "," key.
   iii. 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
   iv. Press ENTER

v. For example "=QUARTILE(N4:N21,1)" will tell you that the first quartile of sales data is $25,499 million; about 25% of sales are below $25,499 million.

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 A25, type "Max" and type "=MAX(D4:D21)" in D25. Repeat this for the other columns. (Don't forget to convert to percents for the growth rates.)
   ii. In A26, type "Min" and type "=MIN(D4:D21)" in D26. Repeat this for the other columns. (Don't forget to convert to percents for the growth rates.)
   iii. In A27, type "Range" and type "=D25-D26" in D27. Repeat.

b. Standard deviation is a really common function; the equation's built right in.
   i. In A28, type "Standard Deviation" and type "=STDEV.S(D4:D21)" in D28. 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 by default.

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 CVS locations: one with a lot of pedestrian traffic and one with a low amount of pedestrian traffic. 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.

$$CV_{sample} = \frac{s}{\bar{x}}(100) \qquad CV_{population} = \frac{\sigma}{\mu}(100)$$

   i. Because we multiply by 100, the result will be a percent.

d. Consider the hypothetical weekly sales data of two different CVS 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:
      1. (\$80 + \$60 + \$40) / 3 = \$60
      2. (\$9 + \$6 + \$3) / 3 = \$6
   ii. Second, find the standard deviation of the samples:
      1. $\sqrt{}(0.5)[(\$80 - \$60)^2 + (\$60 - \$60)^2 + (\$40 - \$60)^2] = \$20$
      2. $\sqrt{}(0.5)[(\$9 - \$6)^2 + (\$6 - \$6)^2 + (\$3 - \$6)^2] = \$3$
   iii. Third, divide and then multiply:
      1. \$20 / \$60 * 100 = 33.3%
      2. \$3 / \$6 * 100 = 50.0%
   iv. The high-traffic location has more consistent sales.

VI. Coefficient of Variation Practice

a. Which division has the most consistent sales? It looks like it should be Interactive, with a standard deviation of just \$375.7 million. But it also has the lowest revenue. So we turn to the coefficient of variation.
b. In A23 type "Mean" and then in D23 type "=AVERAGE(D4:D21)"
c. In A29 type "CV" and then type "=D28/D23*100" in D29. (Note this is the equation from the topic notes: standard deviation divided by mean.) Repeat for the sales of all divisions.
d. Interactive has the least consistent sales; it's Studio that's most consistent. The standard deviation is only about 10% of the mean.