David Youngberg
BSAD 210—Montgomery College

LECTURE 04: DISPERSIONS & COMPARISONS

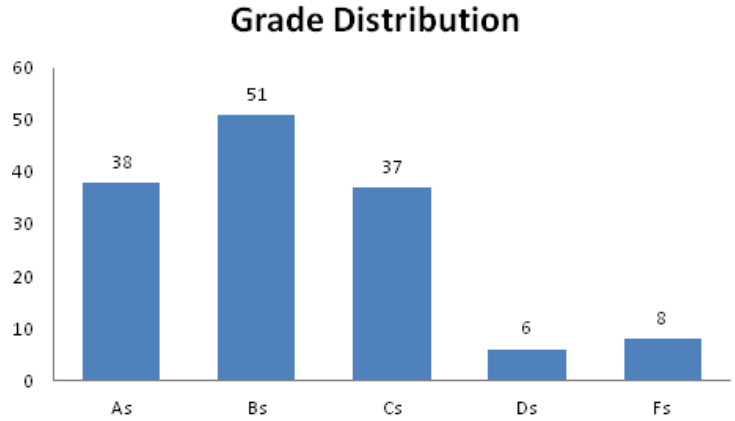
I. Range and Standard Deviation

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.

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 (as in Lecture 5)
   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."

III. 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. Your book has an excellent example demonstrating why "*n* – 1" generates a better result on page 99.Overview

a. Comparing samples is one of the basic things you can do with statistics. It's fundamentally useful, but it can be a little deceptive.

IV. 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.

V. Simpson Paradox

a. Statistics are really useful, but they can also be deceptive since how raw numbers are presented can radically change the result. Consider a suit filed against UC Berkley in 1973, claiming gender bias.
   i. The plaintiffs argued Berkley was biased against women for graduate school admissions. If men are significantly more likely to be admitted than women, that is evidence of gender bias.
   ii. The investigation looked at not just admittance rates for all programs, but each program individually. For simplicity, we

will use made up numbers but these numbers will emulate the paradox investigators found.

	Men				**Women**		
	Applied	*Accepted*	*%*		*Applied*	*Accepted*	*%*
Program A	900	450			100	80	
Program B	100	10			900	180	
Total							

iii. Disaggregating these numbers (describing acceptance rates by major) suggests there's an anti-men bias. Aggregating the numbers suggests there's an anti-women bias.

b. Thus the *Simpson Paradox*—when a correlation present while separated into different groups is reversed when groups are combined.
   i. The paradox arises because we subconsciously assume assignment is random: men and women are equally likely to apply to either program.
   ii. But look again: that's not the case. Nine times as many men as women apply for Program A, which is a tougher program to get into. Nine times as many women as men apply for Program B which is an easier program to get into.

c. By adjusting for population, we can fix this problem. This is called a *weighted average*.
   i. 90% of men applied to Program A, 50% were accepted.
   = 0.90*0.50 = 0.45
   ii. 10% of men applied to Program B, 10% were accepted.
   = 0.10*0.10 = 0.01
   iii. Before, you might have been intuitively adding 0.50 + 0.10 = 0.6; you weren't adjusting for population.
   = 0.45 + 0.01 = 0.46
   iv. Which is what we got when we did the total, above.

d. So the Simpson Paradox isn't "really" a paradox. The strange conclusion just comes from being mathematically incomplete. But since it's such as easy mistake to make, it's worth highlighting.