INFO 1301

Wednesday, September 14, 2016

Topics to be covered

- Last class we spoke about the middle (central tendency) of a data set. Today we talk about variability in a data set.
- Variability of Data
  - Variance
  - Standard Deviation
- The math to calculate measures of variability are more tedious than calculating measures of central tendency, so after doing a bit of introductory material by hand, we will rely on tools such as Minitab Express.

Variance

- Intuitively, variance is the distance data points vary from the mean for a data set.
- Consider two data sets.
  - A is 4,5,6,7,8
  - B is 2,4,6,8,10
- The mean is the same in both cases. (What is the mean?)
- But you can intuitively say that data set B varies more from its mean that data set B.
  - So, however we define variance, it should be higher for data set B than for data set A.

More about variance

- We know how to measure the distance between two points; it is a subtraction.
- So, in the case of data set A, the distances of the various points are:

4-6 = -2

5-6 = -1

6-6 = 0

7-6 = 1

8-6 = 2

The ones to the left of the mean yield negative numbers, the ones to the right yield positive numbers.

- But we are interested in measuring the overall distance of the data set from the mean, not just the distance of each individual data point from the mean.
- So we must do something with these individual distances to calculate a single number that represents the variance of the entire data set.

Even More About Variance

- We calculated the mean by adding up the values of the data points and dividing by the size of the data set. This won't work with variance. (Why?)
- Do you remember from high school how to calculate the distance between two points in the x-y plane? (Hint: use the Pythagorean Theorem)
- Example: The distance from (1,1) to (4,5). (Hint: Draw the right triangle and calculate.)
- Use this same idea to calculate the variance (from the mean).

Your lucky day – even more about variance!

- Process for calculating variance and standard deviation
  1. Find the mean
  2. Find the difference of the mean from each data point
  3. Square each of those differences
  4. Add them together
  5. Divide by n-1 where n is the number of data points in the data set This is known as the variance and is designated by σ^2. [*You would think you would divide by n. However, for reasons that will be explained in a later course, having to do with the difference between samples and populations, you divide by n-1. You will just have to live with this seeming anomaly for now.*]
  6. Take the square root = σ = standard deviation

Variance of Data Set A

1. mean = (4+5+6+7+8)/5 = 30/5 = 6
2. Difference of mean from each data point: 4-6=-2, 5-6=-1, 6-6=0, 7-6=1, 8-6=2
3. Square of those differences: $(-2)^2=4$, $(-1)^2=1$, $0^2=0$, $1^2=1$, $2^2=4$
4. Add them together: 4+1+0+1+4=10
5. Divide by n-1: 10/(5-1)=2.5 = σ^2 (variance)
6. Square root: $(2.5)^{.5}$= approximately 1.6 = σ (standard deviation)

Your turn: calculate the variance and standard deviation for data set B.

How does it compare to the variance for data set A?

Standard Deviation

- The standard deviation is a measure of how close the data is to the mean
- Often – but not always! – 70% of the data will be within one standard deviation of the mean.
- In Data set A, the mean is 6 and the standard deviation is approximately 1.6. So, data points 5,6,7 are within the standard deviation; data points 4,8 are not. So, 60% within 1 σ.
- Often – but not always! – 95% of the data points within 2 σ.
- In Data set A, the mean is 6 and 2 σ = 3.2. All 5 data points – or 100% - within 2 σ.

In-class Exercise

How do the Rockies compare? Assign each person a team.

- Go to the mlb.com site, choose statistics, choose which team, click on AB, record the top 15 batting averages in your spreadsheet in Minitab Express.
- In Minitab Express, click summary along the top, then click descriptive statistics, then click statistics.
- Make sure that the only ones that have check marks beside them are: mean, standard deviation, variance, minimum, median, maximum, mode, N
- Use Minitab Express to calculate these descriptive statistics.
- Compare with your neighbors. What are the Rockies prospects?
- **Keep your data set! You are going to use it again next class.**