Math 247: Summaries for Symmetric Distributions (Section 3.1)

Measuring Center:

Mean: The *mean* of a data set is just the *average* of the data values. It's the "balancing point" of the data.

Notation: The data values are listed as $x_1, x_2, x_3, x_4, \ldots, x_n$

$\sum x$ represents the sum of these numbers; i.e., $\sum x = x_1 + x_2 + x_3 + x_4 + \ldots + x_n$

Sample Mean: If the data values come from a sample, then we denote the mean as $\overline{x}$

Formula for the Sample Mean: $\overline{x} = \dfrac{\sum x}{n}$

Measuring Variation:

Deviation/Error: How far a data value is from the mean.

Data value = x

Sample mean = $\overline{x}$

Error = $x - \overline{x}$

Squared Error = $\left(x - \overline{x}\right)^2$

Sum of the Squared Error (SSE) = $\sum\left(x - \overline{x}\right)^2$

Standard Deviation: This is a way of combining all the deviations into one number which gives us a concrete measurement of the spread.

(Sample) Standard Deviation: $s = \sqrt{\dfrac{\sum\left(x - \overline{x}\right)^2}{n-1}}$

Variance: The "variance" (formal definition) of a data set is the square of the standard deviation.

Variance = s^2

(Variance doesn't have its own symbol which is a bit confusing! Also, "variance" sounds like "variation". *Variance* is precisely defined, whereas *variation* or *variability* is a way to loosely describe how spread out the data is.)

Example: Research question: How distracted are drivers? Suppose you poll some people to find out how many texts they sent on their commute to school this morning. The following data sets are their answers.

Construct a dot plot for the data set. Label the axis.

Find each mean (use the proper notation). Show the mean on the dot plot as a triangle.

Next, use a table to find the standard deviation of the data by hand.

a) First poll results: 0, 1, 2, 2, 3

b) Second poll results: 0, 1, 2, 2, 3, 28

Does each mean represent a “typical value” for the data set? Explain.

Which data set has the larger standard deviation? Why was one standard deviation larger?

Which data value(s) contributed the most to each standard deviation?

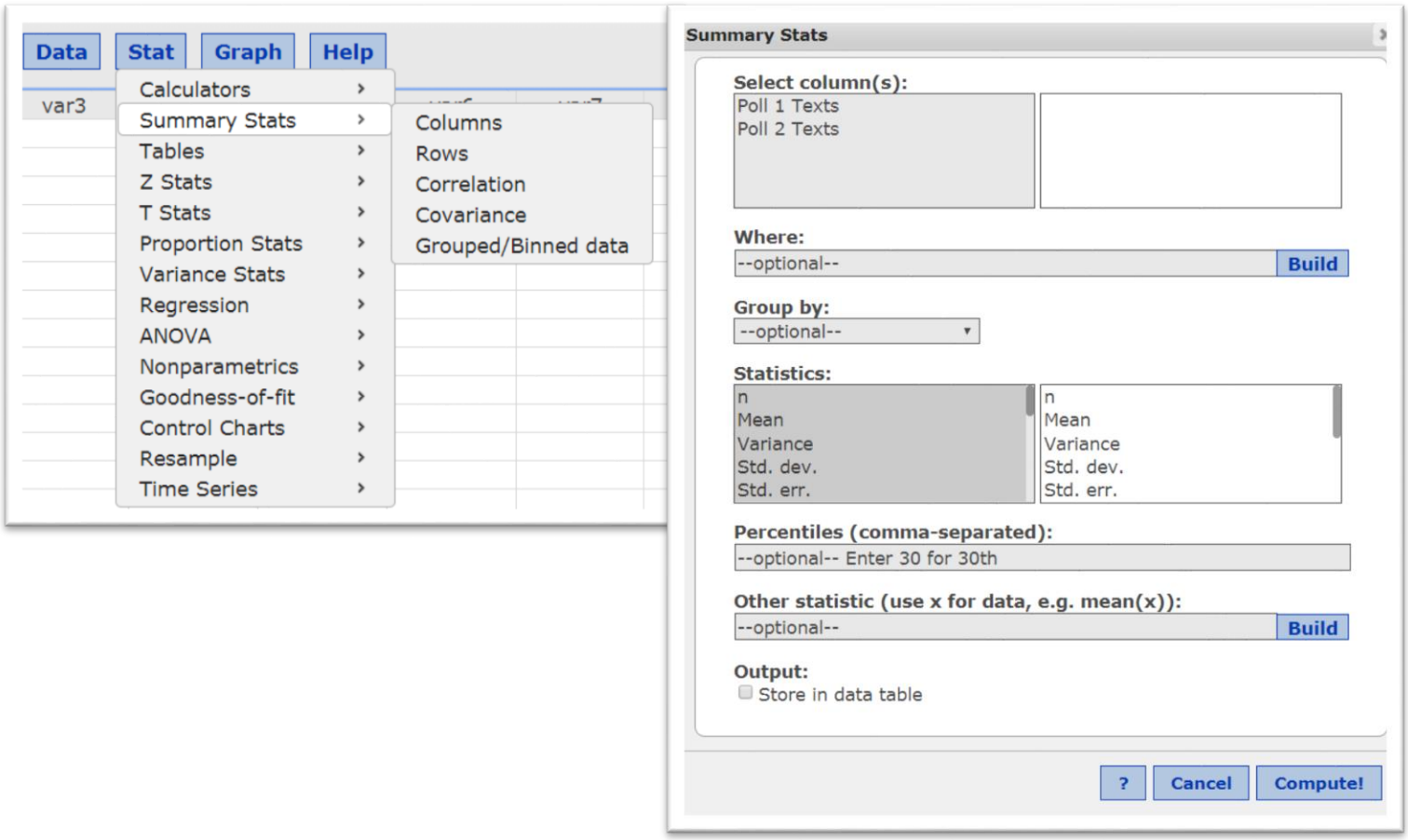
Would you have known, just by looking at the dotplot, which data set would have a larger standard deviation? Explain.

Finding Descriptive Statistics using StatCrunch:

(1) Type the data in by hand into a blank data table. Label the columns Poll 1 Texts, Poll 2 Texts

(2) Next, find the "Summary Stats"
Click the "Stat" box,
then "Summary Stats",
then "Columns".

In the dialogue box, select the column(s) of interest (we'll select both).



This is what you should end up with:

Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Poll 1	5	1.6	1.3	1.1401754	0.50990195	2	3	0	3	1	2
Poll 2	6	6	117.2	10.825895	4.4196531	2	28	0	28	1	3