Unit 3: Describing Data with Numbers

When analyzing quantitative and categorical data our first step is to look at the picture. Our second step with quantitative data is to calculate numerical values that will allow us to better describe the center and spread of a distribution. The two major ways of describing data numerically is with the 5-number summary or with the mean and standard deviation. Which of these you use will depend on the data, we will discuss this at the end of this unit.

5-number summary

_____
_____

Median (I am assuming you know how to find a median)

_____
_____

Quartiles and Interquartile Range (IQR)
To calculate the quartiles:

1. _____
2. _____
3. _____
4. _____

Notes:
The interquartile range can be used as another measure of spread when describing the spread of a data set.
When there is an odd number of data some statisticians will use the median when finding both Q_1 and Q_3, others will ignore it when finding the quartiles. I am going to ignore it, the decision is yours to make.

Example:
Suppose we have the following data set: 14, 3, 25, 2, -17, 13, 45

Boxplots
A Boxplot is a way to display our 5-number summary. Using your data from above, construct a boxplot

Outliers
We classify a data member as an outlier if falls outside the following boundaries.

Lower boundary: _____

Upper boundary: _____

Use your data above to calculate the boundaries for outliers. _____

(**Fun Fact:** You are probably wondering why 1.5 times the IQR? The statistician, Sir Ronald Fischer, who created this said "1 is not enough and 2 is two many". (This is a true quote!!!))

All of the above is easily done with the aid of technology. We are less concerned with being able to calculate the 5-number summary and more concerned with what it says about our data. (FYI: Graphing calculators do not calculate the boundaries for outliers, you need to know how to do this.)

Example
Here are the calorie counts for a single serving of 23 Kellogg's brand cereals.

50 70 90 90 100 100 100 110 110 110 110 110 110 110 110 110 110 120 120 120 140 140 160

Find the 5-number summary, sketch a boxplot and find the boundaries for outliers

Mean and Standard Deviation

I am going to assume that you understand that the mean is the numerical average of the data and how to find it.

Standard deviation is a measure of spread or variability. In nonmathematical terms it is the average distance that the data members are away from the mean. The formula for standard deviation is:

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

(Note: if we square both sides $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$, this is called the variance)

Example:

Suppose we asked 9 elementary school children how many pets they had and got the following data.

1 3 4 4 4 5 7 8 9

Thank goodness for technology and that we never have to calculate standard deviation by hand!!!!

Some fun facts about the standard deviation.
* s measures spread about the mean, $\bar{x}$. Use standard deviation to describe the spread of a data set only when you use $\bar{x}$ to describe the center.

* s = 0 only when there is no variability. This happens only when all observations have the same value. So standard deviation zero means no spread at all. Otherwise, $s > 0$. As observations become more spread out about their mean, s gets larger.

What numerical description should you use for center and spread, median and IQR or mean and standard deviation?

Start by making a picture of your data and describe the shape of your distribution.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Let us use our Presidents Data again.

Ages of Presidents at Death

Washington	67	Filmore	74	Roosevelt	60	Ford	93
Adams	90	Pierce	64	Taft	72	Reagan	93
Jefferson	83	Buchanan	77	Wilson	67		
Madison	85	Lincoln	56	Harding	57		
Monroe	73	Johnson	66	Coolidge	60		
Adams	80	Grant	63	Hoover	90		
Jackson	78	Hayes	70	Roosevelt	63		
Van Buren	79	Garfield	49	Truman	88		
Harrison	68	Arthur	57	Eisenhower	78		
Tyler	71	Cleveland	71	Kennedy	46		
Polk	53	Harrison	67	Johnson	64		
Taylor	65	McKinley	58	Nixon	81		

Using your calculator find the 5-number summary, mean, standard deviation, boundaries for outliers, create a boxplot and decide which measure of center and spread you should use.