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
The five-number summary of a distribution consists of the smallest observation, the first quartile, the median (second quartile), the third quartile , and the largest observation, written in order from smallest to largest.

Min Q_1 M (Q_2) ($\tilde{x}$) Q_3 Max

Median (I am assuming you know how to find a median)
The median (M, Q_2, $\tilde{x}$) is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger.

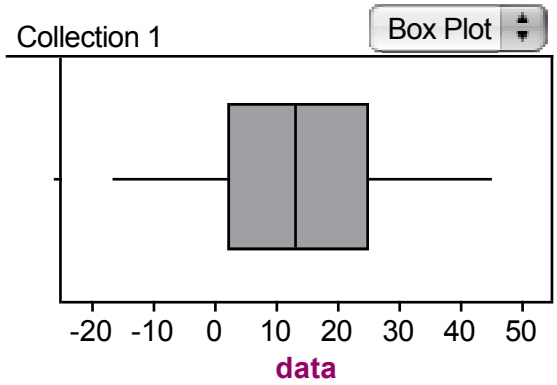
Quartiles and Interquartile Range (IQR)
To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median in the list of observations
2. The first quartile Q_1 is the median of the lower half of the observations.
3. The third quartile Q_3 is the median of the upper half of the observations.
4. The interquartile range is defined as IQR = $Q_3 - Q_1$

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.

Suppose we have the following data set: 14, 3, 25, 2, -17, 13, 45
First we order it: -17, 2, 3, 13, 14, 25, 45
Since there are 7 data members the median is 13
Ignoring 13 I find the middle of the lower three numbers – 2, this is Q_1
Ignoring 13 again I find the middle of the upper three numbers – 25, this is Q_3
My 5-number summary is: -17, 2, 13, 25, 45

Boxplots
A Boxplot is a way to display our 5-number summary. Below you will see a boxplot of the above data.



A central box is drawn from the first quartile to the third quartile. A segment in the box marks the median. Lines extend from the box out to the smallest and largest observations that are not outliers.

We will describe this boxplot that same as before: shape, center, spread and outliers. For our boxplot, it is symmetric, the center is 13 (we now have a calculated value, not an approximation) and our spread is 62 or 23 depending on whether you want to use the range or IQR. There are no outliers. If there were outliers I would have definitely used the IQR to describe the spread.

Is it an outlier or not?
So far determining whether a data point is an outlier has been a matter of opinion. We will now discuss a more concrete method of determining outliers. Sometimes we will still have to have an educated opinion on outliers but if we have the entire data set we can use this method

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

Lower boundary: Q1 – 1.5 IQR (any number less than this is considered an outlier)
Upper boundary: Q3 – 1.5 IQR (any number more than this is considered an outlier)

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!!!)

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

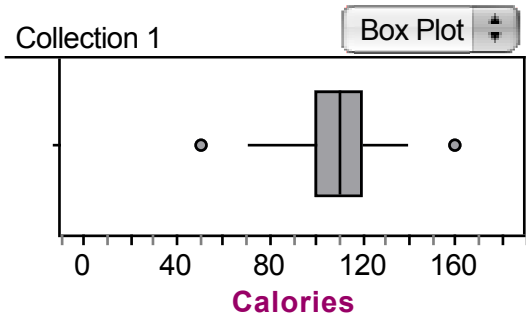
50 70 90 90 100 100 100 110 110 110 110 110 110 110 110 110 110 120 120 120 140 140 160

The 5-number summary is:

Min	50
Q1	100
M	110
Q3	120
Max	160

IQR = 120 – 100 = 20
Boundaries for outliers: 100 – 1.5(20) = 70 and 120 + 1.5(20) = 150

This tells us that 50 and 160 are considered outliers. The boxplot below shows how we represent data with outliers.



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.)

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. Since the mean is a numerical average, anytime you see the word average (in these pages) it relates to the mean. This is one way that the media tries to trick you, when they use the word average what are they using, the mean or median? (Something to think about) Some books will tell you that if you are given a histogram the mean will be the balancing point of that histogram. This is true, that is why up until this point when we were trying to find the center of a histogram I kept telling you to look for the balancing point.

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. Unfortunately we cannot just take the distance each data member is from the mean and average those distances together. I'll show you why shortly!! 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)

It looks pretty complicated but if you do it step-by-step you will see how easy it is!!

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

The first thing we would do is calculate the mean. $\frac{1+3+4+4+4+5+7+8+9}{9} = 5$ (this is the $\bar{x}$ in the formula.

Next we will subtract each data member from the mean and square this distance. This is the $(x - \bar{x})^2$ in the formula.

$(1 - 5)^2 = (-4)^2 = 16$
$(3 - 5)^2 = (-2)^2 = 4$
$(4 - 5)^2 = (-1)^2 = 1$
$(4 - 5)^2 = (-1)^2 = 1$
$(4 - 5)^2 = (-1)^2 = 1$
$(5 - 5)^2 = (0)^2 = 0$
$(7 - 5)^2 = (2)^2 = 4$
$(8 - 5)^2 = (3)^2 = 9$
$(9 - 5)^2 = (4)^2 = 16$

Note: Remember how I said it would be great if we could just add up the distances each data is from the mean and average these distances but we could not, here is why. Add up all the distances (before you square them) from the mean:

$-4 + -2 + -1 + -1 + -1 + 0 + 2 + 3 + 4 = ?$ **It equals 0!!!**

This will happen with every data set because the mean is the center of the data set. This is why we have the formula for standard deviation.

Continuing with the formula, the $\sum$ tells us to add up all the distances from the mean that we squared.

$16 + 4 + 1 + 1 + 1 + 0 + 4 + 9 + 16 = 52$

Now we divide by n – 1 (n is the number of data members).

$52 \div (9 - 1) = 6.5$
(since we have not taken a square root yet this is the variance, $s^2 = 6.5$)

Now we take a square root: $\sqrt{6.5} = 2.55$. This is your standard deviation, s = 2.55. This is the average difference that each data member is away from the mean. See, the formula is not that hard … except if you have a large data set. Because of this we will be using technology to find our standard deviation. (Note: On your graphing calculators we will be using the S_x for standard deviation because this is the standard deviation for a sample. Remember sample, the subset of the population. The σ_x on your calculator is the standard deviation for the population, we will address this later in the course.)

Some fun facts about the standard deviation.
* s measures spread about the mean, $\bar{x}$. Use s 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. If this shape is skewed use the median and IQR to describe the center and spread of your distribution. Why?

When a data set has very high or low values (skewed picture) this affects the mean and standard deviation and doesn't affect the median and IQR. The mean and standard deviation are calculations that use ever data member their formulas while the median and IQR are positions in the data set. A data set that is skewed will have a mean and median that are not close to each other. The mean gets pulled toward the skew. If the data is skewed right, the mean will be greater than the median. If the data is skewed left, the mean will be less than the median.

If the shape of your distribution is symmetric use the mean and standard deviation to describe the center and spread. Because the data is symmetric the mean and median should be almost equivalent. In fact the mean and median are exactly equal if the distribution is exactly symmetric.

Comparing Distributions

Some of the most interesting statistics questions involve comparing two or more distributions. To make such comparisons, start with a graph. If you are studying a categorical variable, use a bar graph or pie chart. For quantitative variables, make a dotplot, stemplot, histogram, or boxplot, then add numerical summaries, comparing their shape, center and spread. Below you will see two boxplots comparing the grams of fat vs. grams of fiber in various cereals.