How are the values related to each other? Are there values that are far away from the others?

General Education Statistics
Class Notes
Measures of Position and Outliers: Z-scores, Percentiles, Quartiles, and Interquartile Range (Section 3.4)

We have talked about the center and the spread of a set of data values. Now, we look at how the values are positioned about each other.

Using z-scores to compare data:

Consider the Los Angeles Angels, a baseball team in the American League. During the 2014 season, the Angels scored 773 runs. During that same season, the Colorado Rockies, who play in the National League, scored 755 runs.

It seems as though the Angels outperformed the Rockies. But did they really?

The National League and American League have an important difference. The National League makes their pitchers hit (and they are rather notorious for doing it poorly). The American League, on the other hand, uses designated hitters to replace the pitchers when it is their time at bat.

The National League averages 640 runs with a standard deviation of 55.9 runs. The American League averages 677.4 runs with a standard deviation of 51.7 runs.

The idea of **z-scores** will allow us to compare each team, not directly with each other, but with their respective leagues, and then against each other.

Definition: The **z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population z-score and a sample z-score.

$$z = \frac{x-\mu}{\sigma} \quad \text{or} \quad z = \frac{x-\bar{x}}{s}$$

The z-score has *no* units (like feet or seconds). They have a mean of 0 and a standard deviation of 1.

expl 1: Use the population information given for each league to find the *z*-scores for the Angels and the Rockies. Compare them.

Angels (2014): 773 runs
The American League
averages 677.4 runs with
a standard deviation of
51.7 runs.

$$z = \frac{x - \mu}{\sigma}$$

Rockies (2014): 755 runs
The National League
averages 640 runs with a
standard deviation of
55.9 runs.

So, who did better? Explain.

expl 2: What would cause a *z*-score to be negative versus being positive?

expl 3: Bob and Maggie ran a marathon. The mean time to complete the marathon for men was 242 minutes (with a standard deviation of 57 minutes). The mean time for women was 273 minutes (with a standard deviation of 52 minutes). Bob's *z*-score is –0.51 and Maggie's *z*-score is –0.62. Who did better?

Consider what a
better score means?

Definition: Percentiles: The ***k*th percentile**, denoted, P_k, of a data set is a value such that k percent of the observations are *less than or equal to* the value.

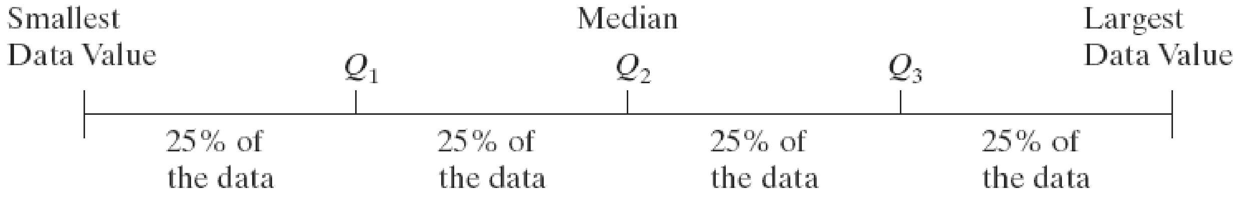
You may have gotten SAT or ACT results back and learned that you scored in the 74th percentile. What does that mean, with respect to the other test-takers?

Percentiles break the data up into 100 parts, essentially. We could divide the data up into just four parts. This is called **quartiles**.

Definition: Quartiles divide data sets into fourths, or four equal parts.

- The 1st quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the 1st quartile is equivalent to the 25th percentile.
- The 2nd quartile, denoted Q_2, divides the bottom 50% of the data from the top 50% of the data. The 2nd quartile is equivalent to the 50th percentile, which is, in fact, the **median**.
- The 3rd quartile, denoted Q_3, divides the bottom 75% of the data from the top 25% of the data. The 3rd quartile is equivalent to the 75th percentile.

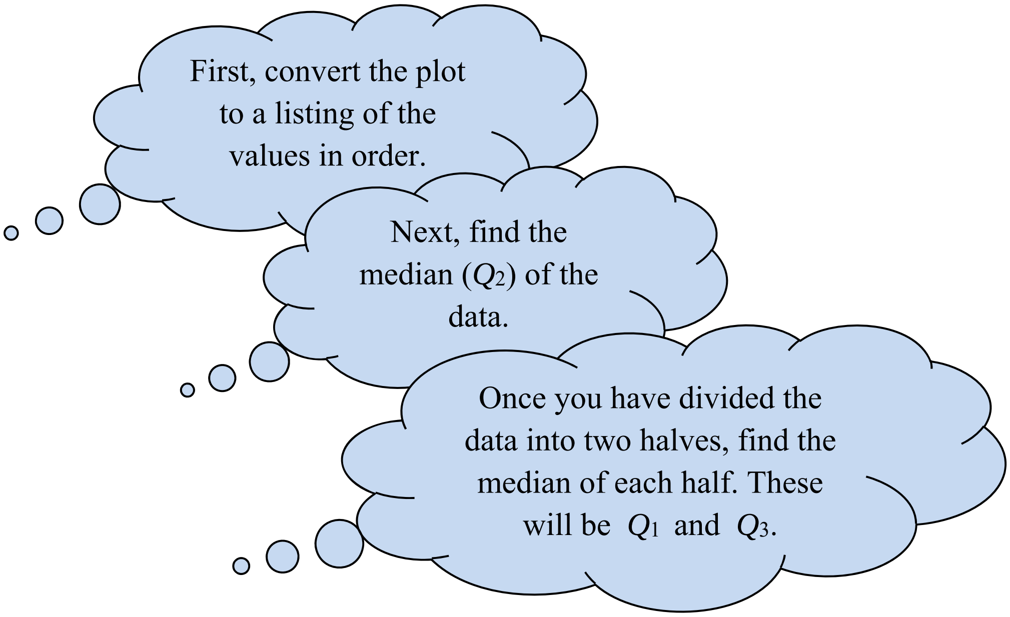
Here is a nice picture of how these quartiles break up the data.



expl 4: Find the median (Q_2) and then Q_1 and Q_3 for the following data.

This is a stem-and-leaf plot for 17 states detailing the percentage of people who are aged 65 or older. In this graphic, the entry 10 | 6 means 10.6 %. (Source: Statistical Abstract of US, 1995)

4	6
5	
6	
7	
8	
9	
10	1 1 6
11	6
12	1 6 7 7 8 9
13	4 9
14	2 8
15	4
16	
17	
18	4



First, convert the plot to a listing of the values in order.

Next, find the median (Q_2) of the data.

Once you have divided the data into two halves, find the median of each half. These will be Q_1 and Q_3.

These quartiles are part of the calculator output when you perform **1:1-Var Stats** in the **STAT** > **CALC** menu.

Definition: The **interquartile range**, **IQR**, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

$$\text{IQR} = Q_3 - Q_1$$

Return to page 3 and label this on the graphic.

Quartiles and IQR are resistant. They are *not* much affected by outliers.

expl 5: Find the interquartile range of the data in example 4.

Use the same units as the data.

expl 6: Let's use the results from examples 4 and 5 to answer some questions.

a.) What percentage of the data has a value that is less than or equal to 11.1? Write your answer in a sentence that explains the full meaning of the data.

b.) What percentage of the data has a value that is greater than 12.7? Write your answer in a sentence that explains the full meaning of the data.

c.) Between which two values do the middle 50% of the data lie? Write your answer in a sentence that explains the full meaning of the data.

The IQR is a good measure of the dispersion of the data.

For the remainder of the semester, when asked to find the distribution of a data set, we will describe its shape (symmetric, skewed left, or skewed right), its center (mean or median), and its spread (standard deviation or interquartile range). Use medians and interquartile ranges if you have skewed data.

Checking Data for Outliers:

To this point, we have described outliers to be values that appear way larger or smaller than the other values. However, there is a more defined statistical method we see now.

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. **Fences** serve as cutoff points for determining outliers.

Lower Fence = $Q_1 - 1.5(\text{IQR})$

Upper Fence = $Q_3 + 1.5(\text{IQR})$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an **outlier**.

expl 7: Refer back to the data in example 4 concerning the percentage of states with populations aged 65 or older.

a.) From examples 4 and 5, record the first and third quartiles as well as the interquartile range.

b.) Follow step 3 above to find the lower and upper fences. Follow step 4 above to determine if the data set has any outliers. What are the outlier(s)?