3.3 Measures of Relative Standing and Boxplots

Standardized Exam A: $50 \le x \le 250$ $\mu_1 = 187.4$ $\sigma_1 = 9.7$

Standardized Exam B: $0 \le x \le 80$ $\mu_2 = 63.3$ $\sigma_2 = 3.8$

Two students each take one of the standardized exams with summary stats above. Alice scores a 205 on Exam A, and Bob scores a 69 on Exam B. If only one scholarship is given to the highest score, which student is awarded the scholarship?

So, how can we determine who did better? Maybe by percentage:
$A\% = \frac{205}{250} \approx 82\%$ and $B\% = \frac{69}{80} \approx 86.25\%$

Appears Bob did better. BUT!!!! it looks like test A is more difficult with a lower average:

$\overline{A} = \frac{187.4}{250} \approx 75\%$, while $\overline{B} = \frac{63.3}{80} \approx 79\%$. So, maybe an 82% on a more difficult test is better than an 86% on an easier test.

It also appears they both did about 7% higher than their respective test means. Maybe they're equal.

But, how does the fact that test A starts at 50 affect the overall scores. And, does the fact that test A has a greater standard deviation affect the comparison.

The best way to compare two values from different distributions is to compare how many standard deviations from their respective means each score is. This is done using the **z-score.**

To convert values to a standardized scale, we calculate how many standard deviations from the mean a particular value is. This value is called the **z-score** (or standardized value).

$z = \frac{x-\overline{x}}{s}$ using sample statistics or $z = \frac{x-\mu}{\sigma}$ using population parameters

The numerator, $x - \overline{x}$ (or $x - \mu$) calculates the distance each value is from the mean. But to see how many **standard deviations** the values are away from the mean, divide by the respective standard deviation.

The greater the z-score (both positive or negative) the more "unusual" or rare the value actually is.

Compare the *z*-scores for both Alice and Bob above and determine the better score. (Round all *z*-scores to two decimal places.)

Alice: $z = \frac{205-187.4}{9.7} \approx 1.81$ Bob: $z = \frac{69-63.3}{3.8} \approx 1.50$
Since Alice is 1.81 standard deviation above the mean compared to only 1.5 for Bob, Alice definitely had the better score. Give her the scholarship.

Later in the quarter we'll calculate the percentage of scores that her score is better than. Alice actually scored better than 96.5% of all test takers, and Bob only scored better than 93.3%. Fairly close actually.

(**Note:** usual values will have z-scores $-2 \le z \le 2$, and unusual values have $z < -2$ or $z > 2$.)

This means it's rare that a data value is more than 2 standard deviations from the mean. Remember the Empirical (68-95-99.7) Rule. Values outside 2 standard deviations (i.e., $z-\text{score} > 2$ or $z-\text{score} < -2$) make up 5% of the data, so chance that a value is **greater** than 2 st.dev is only 2.5%, very rare.

Another way of comparing sets of data is by way of the **five number summary** and **boxplots**, or side-by-side boxplots.

BOXPLOTS AND THE FIVE NUMBER SUMMARY

The *Five Number Summary* divides an entire data set into four equal groups, where each group contains roughly 25% of the data values. These are called quartiles.

Five Number summary:

1. Minimum Value
2. First Quartile, Q_1 (median of the lower half)
3. Second Quartile, Q_2 (or the median of all the data)
4. Third Quartile, Q_3 (median of the upper half)
5. Maximum Value

Use the five-number summary to compare home runs of Hank Aaron and Babe Ruth

Aaron: 11 12 13 20 24 26 27 29 30 32 34 34 38 39 39 40 40 44 44 44 44 45 47

Ruth: 0 2 3 4 6 11 22 25 29 34 35 41 41 46 46 46 47 49 54 54 59 60

The first step in finding the five number summary is to write the data in increasing order, which is already done for us here. Use the SortA() command on the TI84 to get it sorted quickly.

Here's how to find Aaron's five number summary:
1) Obviously the max and min are easy to find: min = 11, max = 47.
2) Q_2 is the median, Since $n = 23$, the 12^{th} value will split the bottom 11 values from the top 11 values

11 12 13 20 24 26 27 29 30 32 34 **34** 38 39 39 40 40 44 44 44 44 45 47

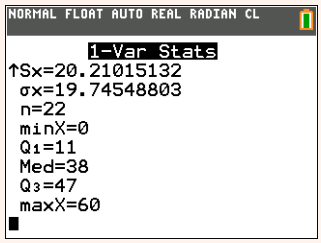
3) Q_1 is the **median** of the lower half, and Q_3 is the median of the upper half:

11 12 13 20 24 **26** 27 29 30 32 34 **34** 38 39 39 40 40 **44** 44 44 44 45 47

The Five number summary for Aaron is:

min = 11
Q_1 = 26
Q_2 = 34
Q_3 = 44
max = 47

Evaluating 1-Var Stats on the TI84 also gives the five number summary. Here is Ruth's FNS:



We can also represent the five number summary graphically using boxplots.

Boxplots are a graphical representation of the five-number-summary:

Boxplots come in a variety of styles. Mine is a little different than our text. What they all show is:
- The range of the smallest 25% of the data is between min and Q_1.
- The next 25% is between Q_1 and Q_2 (or the median).
- 75% of the data lies below Q_3.

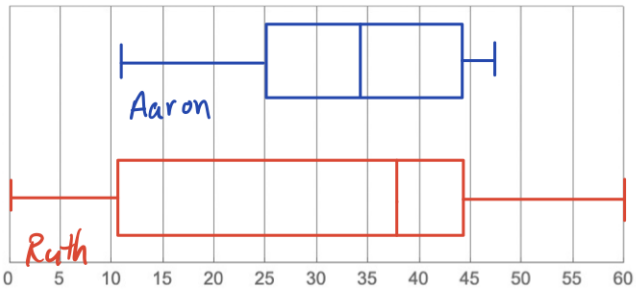
This is why these values are called **quartiles.**

Also, the amount of data between Q_1 and Q_3 is 50%. This is called the **interquartile range,** and gives us another way of looking at unusual values, or more specifically, outliers. These are usually noted with some kind of symbol, e.g., *.

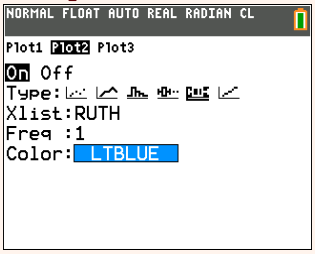
Define the **Interquartile Range** as **IQR** = $Q_3 - Q_1$. An outlier is any value higher than Q_3 by 1.5 × IQR, or below Q_1 by 1.5 × IQR. Outliers are indicated by a symbol.

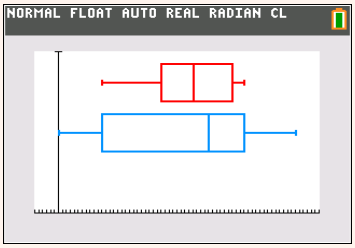
Use your calculator to create side by side boxplots for the Aaron-Ruth data. Which is the better hitter; more consistent hitter?

Since we have the five number summaries for Aaron and Ruth above, you can do this without the calculator.



However, with the two lists entered in your calculator, we can create two side-by-side boxplots for a nice comparison of the data (shown is the setup for Plot2 for Ruth):





PERCENTILES

Quartiles are a special case of quantiles or **percentiles**, e.g., $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$.

Example: The value that separates the bottom 30% from the top 70% is the 30 percentile or P_{30}.

1. Sort the data.
2. Calculate $L = n \cdot \frac{30}{100}$, where n is the number of values:
   a) if L is an integer, P_{30} is the mean of L^{th} value and $L + 1$ value.
   b) if L is not an integer, round up and use that L^{th} value as P_{30}.

I think these are fairly straight forward. The percentiles (and quartiles) are a measure of position, meaning, the 80^{th} percentile, or P_{80} is a value that separates the bottom 80% from the top 20%. Going back to Alice and Bob, we can say that Alice is in about the 95^{th} percentile, while Bob is in the 93^{rd} percentile.

Find the 65th percentile for Ruth.

There are 22 values in Ruth's data, so

$L = 22\ (0.65) = 14.3$
Since this isn't an integer, we round this up to 15. We need to find the 15^{th} data value in Ruth's data, which I find to be 46. Therefore, $P_{65} = 46$, meaning Ruth hit at least 46 home runs per year in 65% of the years he played.

One thing I forgot to mention was if you have skewed data (data that looks fairly normal, but might appear to be stretched to the right). How does this affect the location of the mean when plotted on a box plot? With perfectly normal data the mean and median are right on top of each other. For skewed data, what happens to the mean?