# 5 – Position & Data Analysis

Measures of position are numerical values that represent a data point's relative position in a dataset. There are two such measures you should know:

(1) A data point's standard score (aka, z-score) quantifies the number of standard deviations that point is from the mean.
(2) A data point's percentile is the portion of data that is less than that data point.

Here is how each of these statistics can be found:

**Z-score**: A data point's *z*-score is computed by first subtracting the mean and then dividing the result by the standard deviation. If the data point is *x*, then $z = \frac{x-\bar{X}}{s}$ for sample data or $z = \frac{x-\mu}{\sigma}$ for population data.

**Percentile**: To compute this statistic, it is usually easiest to rank the data set and then directly compute the portion of other data points below the one in question. See the example below.

Example: Consider the dataset shown here, the number of text messages received by Math 175 students in the four hours prior to class one day, tabulated over several consecutive semesters.

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 2 | 5 | 10 | 23 | 49 |
| 0 | 0 | 3 | 6 | 11 | 23 | 49 |
| 0 | 0 | 3 | 6 | 11 | 24 | 51 |
| 0 | 1 | 3 | 6 | 12 | 24 | 52 |
| 0 | 1 | 3 | 7 | 12 | 30 | 59 |
| 0 | 1 | 3 | 7 | 15 | 35 | 75 |
| 0 | 1 | 3 | 10 | 15 | 35 | 100 |
| 0 | 1 | 4 | 10 | 16 | 38 | 150 |
| 0 | 1 | 5 | 10 | 20 | 38 | 170 |
| 0 | 2 | 5 | 10 | 20 | **46** | 250 |

Mean: 22.60

Standard Deviation: 41.85

Choosing the student who responded with the value of 46 text messages, we can compute the *z*-score: $z = \frac{46-22.6}{41.85} = 0.56$. This tells us that the value of 46 is 56% of a standard deviation above the mean (positive ⇔ above the mean; negative ⇔ below)

Separately, if we wish to compute the percentile, we count 70 values in the dataset, and the value 46 is larger than 59 of those values. The percentile can thus be computed as $\frac{59}{70} \approx 84\%$. We say that 46 is the 84th percentile.

As a second problem, since this is a sample of Point Park students and we assume that it is a representative sample, what number of text messages would represent the 70th percentile?

That value, which may be a theoretical number, can be found by finding 70% of 70, which is 0.7×70 = 49. That means that the 49th number in the dataset, or 20, is the 70th percentile.

* Note: The textbook provides several procedures to computing percentiles. Please know that in my class I only require that you use the definition of percentile (portion that are lower) and will be lenient when grading computations of percentiles.

Just like central tendency and spread are characteristics that represent a dataset, the Five Number Summary is a way of summarizing a data set with only five numbers. Those five numbers are: {minimum, first quartile, second quartile, third quartile, and maximum}. Finding these numbers is simple once the data is ordered from lowest to highest. The minimum and maximum values are obvious. Use this procedure to find the other three numbers:

(1) Find the median. This is the second quartile, or $Q_2$.
(2) Find the median of the half of the dataset below the median. This is the first quartile, or $Q_1$.
(3) Find the median of the half of the dataset above the median. This is the third quartile, or $Q_3$.

Note: When computing $Q_1$ and $Q_3$, do not include $Q_2$ in the portion of the dataset you're working with.

In the above example, the minimum and maximum values are 0 and 250. Since $N = 70$, the median is the number halfway between the 35th and 36th value (which are both 7), so $Q_2$ is 7.

$Q_1$ is the median of the first 35 numbers, or the 18th, which is 1, and $Q_3$ is the median of the top 35 numbers, or the one 18th from the last, which is 24.

The Five Number Summary for this dataset is {0, 1, 7, 24, 250}.

From the identification of $Q_1$ and $Q_3$ comes a procedure to identify outliers. An outlier is an extreme value that is either much lower than most of the other values in a dataset, or much larger. Outliers are identified as data points that are more than 1.5 times the Inter-Quartile Range below or above $Q_1$ and $Q_3$. This is explained in more detail below using the example dataset.

First, the Inter-Quartile Range (IQR) is just the difference between $Q_3$ and $Q_1$, or 24 – 1 = 23 in the example dataset.

1.5 times the IRQ is 1.5×23 = 34.5. So, we look for values in our dataset that are more than 34.5 below $Q_1$ or more than 34.5 above $Q_3$. The interval we're using goes from 1 – 34.5 = – 33.5 to 24 + 34.5 = 58.5.

Therefore, this dataset has six outliers: 59, 75, 100, 150, 170, and 250.

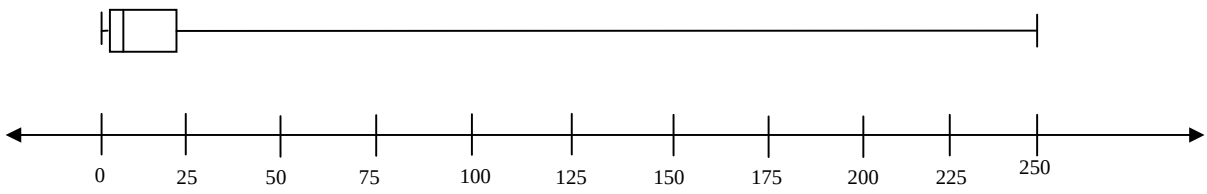
In addition to their usefulness in identifying outliers, the quartiles can be helpful in another visual display of a dataset. The Five Number Summary can be used to draw a display with no vertical axis called a boxplot.

To draw a boxplot for a dataset, we must first find the Five Number Summary. Then, a horizontal axis is drawn that will contain all five of these values. It is important that the axis is drawn to scale *before* the boxplot is drawn.

In the example above, this means that we need to make room on a horizontal axis for the numbers from 0 to 250.

Then, a rectangle is drawn horizontally between $Q_1$ and $Q_3$ with a vertical divider at $Q_2$.

Lastly, "whiskers" are drawn from $Q_1$ to the minimum and from $Q_3$ to the maximum.




Variations of the boxplot remove the outliers. In this case, if we choose to, we can replace 250 (which is an outlier) with 52(which is the largest value that is not an outlier) in the Five Number Summary. The boxplot could be modified to look this way, with a symbol representing an outlier that is outside of the scale shown.