Section 1G – Quantitative Data Analysis for Non-Normal Data and Summary Statistics

Vocabulary

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Normal Data: Data that is bell shaped, symmetric and unimodal.

Skewed Right Data: Also called positively skewed. Data where the center is on the far left and has a long tail to the right.

Skewed Left Data: Also called negatively skewed. Data where the center is on the far right and has a long tail to the left.

Sample Size: Also called the total frequency. The number of values are in a data set.

Median Average: The center of the data when the numbers are put in order. Also called the "50th Percentile" (P_{50}). Since about 50% of the numbers in the data set are less than the median. It is also called the "Second Quartile" (Q_2). The average for a data set that is not normal.

First Quartile (Q_1): The number that about 25% of the data values are less than. Used for typical values for data that is not normal.

Third Quartile (Q_3): The number that about 75% of the data values are less than. Used for typical values for data that is not normal.

Interquartile Range (IQR): The distance between the middle 50% of the numbers in a data set. Calculated by subtracting the 1st and 3rd quartiles. The measure of typical spread for a data set that is not normal.

Maximum: The largest number in a data set.

Minimum: The smallest number in a data set.

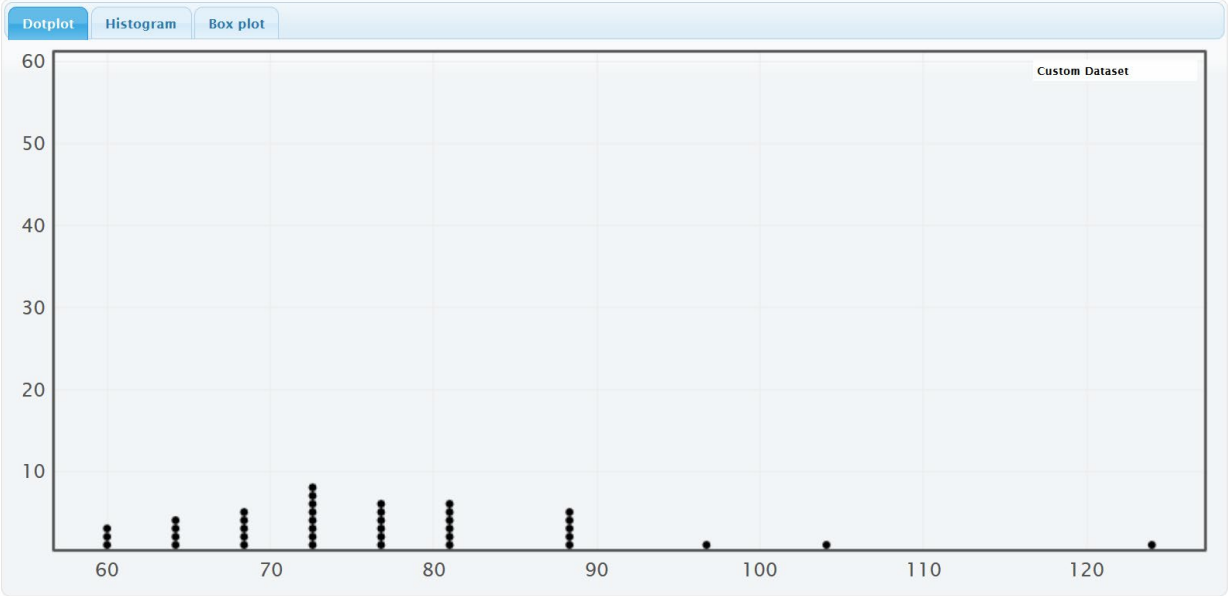
Range: A quick measure of total spread. Calculated by subtracting the minimum and maximum values in a data set.

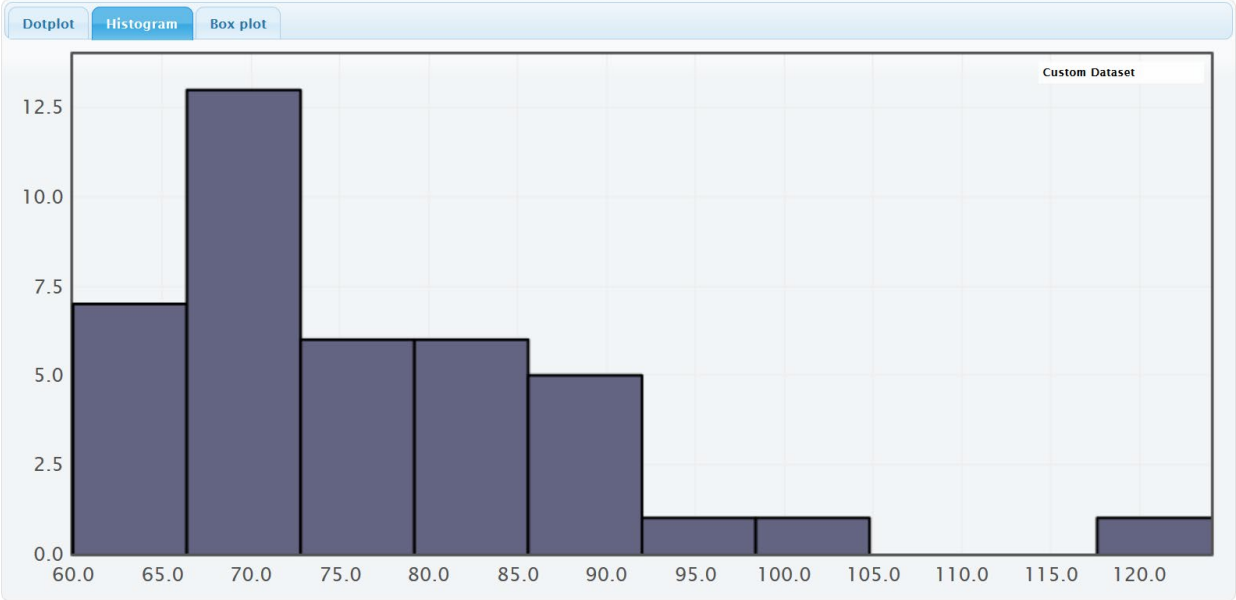
Outliers: Unusual values in the data set.

Introduction

When a data set is normal (or bell-shaped), we use the mean as our average and the standard deviation as our measure of typical spread. Not all data sets are normal though. Let us explore some data that is not normally distributed.

Let us look at another example from the health data. This time we will look at women's pulse rates in beats per minute (BPM). Go to www.matt-teachout.org and click on the "Statistics" tab and then the "Data Sets" tab. Open the health data in Excel. Copy the women's pulse rate data. Now go to www.lock5stat.com and click on "StatKey". Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable". Under "Edit Data", paste the women's pulse rate data into StatKey. Uncheck the box that says, "First column is identifier". Check the box that says, "Data has header row". Push "OK". Here are the graphs and summary statistics.





Notice first that this is not normal data. The highest bar (center) is on the far left. The graph has a short tail to the left of the highest bar and a long tail to the right of the highest bar. This shape is called “skewed right” or “positively skewed”. We can adjust the number of bars (buckets) by using the slider on the right of the graph.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	76.300
Standard Deviation	12.499
Minimum	60
Q_1	68.000
Median	74.000
Q_3	80.000
Maximum	124

Remember the mean and standard deviation are only accurate if the data is normal. Therefore, for this data set, we should not use the mean as the average and we should not use the standard deviation as our typical spread.

So what statistics should we use? Here is the general rule for skewed data or any data that is not normal.

Summary statistics for non-normal data

Average: Median

Typical Spread: Interquartile Range (IQR)

Typical Values: Between the first quartile (Q_1) and the third quartile (Q_3)

Outliers: Boxplot will indicate if there are outliers.

Quartiles are based on the numbers in order, so are much more accurate for data that is not normally distributed. The median is also called the 2nd quartile or the 50th percentile. It is the center of the data when the numbers are in order. About 50% of the numbers will be less than the median and about 50% of the numbers will be greater than the median. When a data set is not normally distributed, we use the median as our average. It is much closer to the center. Look at the histogram above. The summary statistics provided by StatKey show us that the mean was 76.3 beats per minute (bpm) and the median was 74 bpm. Notice 74 is closer to the highest bar in the data set. In other words, the median is closer to the center and a more accurate average than the mean. Mean averages are based on distances so will be pulled off the center in the direction of the skew.

The median is calculated by first putting the numbers in order from smallest to largest. If there is one number in the middle (sample size n is odd), then that is the median. If there are two numbers in the middle (sample size n is even), then the median will be half way between the two numbers in the middle.

The first quartile (Q_1) is also called the 25th percentile and is the number that about 25% of the data is less than. The third quartile (Q_3) is also called the 75th percentile and is the number that about 75% of the data is less than. The first and third quartiles are markers that mark the middle 50% of the data when it is in order. The middle 50% is considered "typical" in a data set that is not normally distributed. For normal data, we want the middle 68% (empirical rule) because there is more data in the middle.

The distance between the first and third quartiles is called the interquartile range (IQR). This is the best measure of typical spread for data that is not normally distributed. StatKey does not list the IQR in its summary statistics, but we can calculate it with the following formula.

IQR = $Q_3 - Q_1$

Since our women's pulse rate data was skewed right, we would use the following statistics.

Variable and Units: Women's pulse rates in beats per minute (bpm)

Minimum: The lowest pulse rate for these women was 60 bpm.

Maximum: The highest pulse rate for these women was 124 bpm.

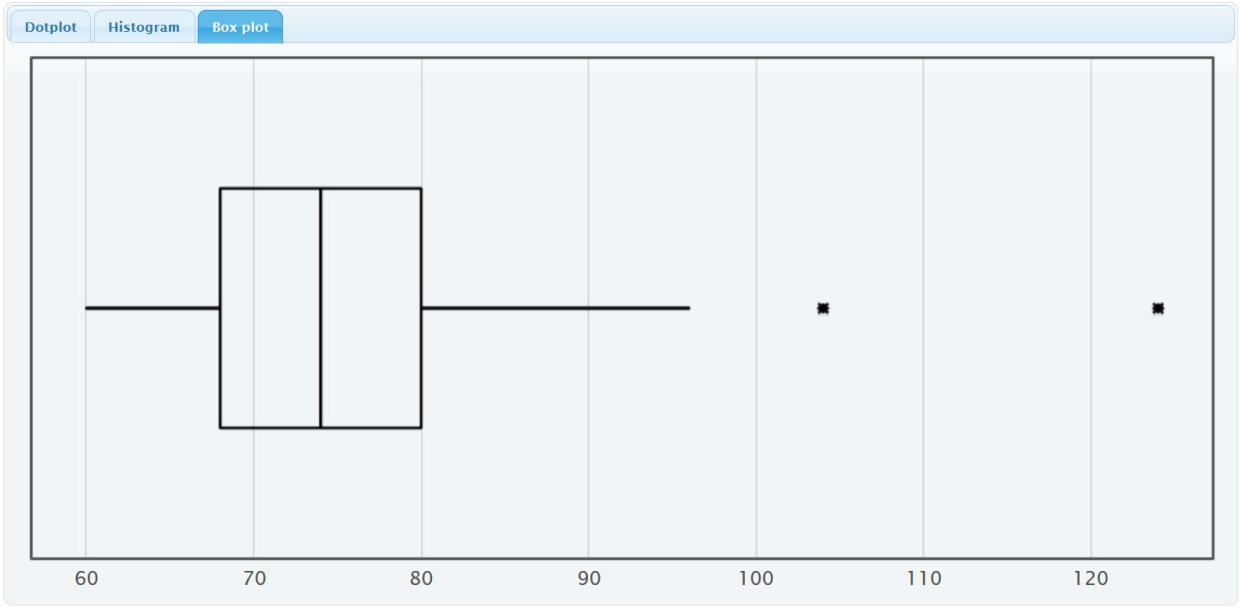
Average: The average pulse rate for these women is 74 bpm (median).

Typical spread: IQR = $Q_3 - Q_1$ = 80 – 68 = 12 bpm
Typical women in the data set had a pulse rate within 12 bpm of each other.

Typical Values: Typical pulse rates are between 68 bpm (Q_1) and 80 bpm (Q_3).

Finding outliers for non-normal data

To find outliers for data sets that are not normally distributed, we will introduce another graph. The graph is called a "box and whisker plot" or "box plot" for short.



A box plot is a graph of the first quartile, median, third quartile and outliers. It is the perfect graph to look at when a data set is not normal. The left of the box is Q_1(68 bpm) and far right of the box is Q_3(80 bpm). So the box represents the typical values (middle 50%). The line inside the box is the median average of 74 bpm. The lines that go to the left and right of the box are called whiskers. The whiskers go to the lowest and highest numbers in the data set that are not unusual (not outliers). The outliers are usually denoted by stars in StatKey and circles and triangles in Statcato. See the two stars the far right. Those are both outliers. There are two unusually high pulse rates in the data set. In StatKey, you can hold your curser over the stars and they will tell you what the numbers are. In this case, the two high outliers are at 104 bpm and 124 pbm. There are no unusually low values since we do not see any stars on the left of the graph.

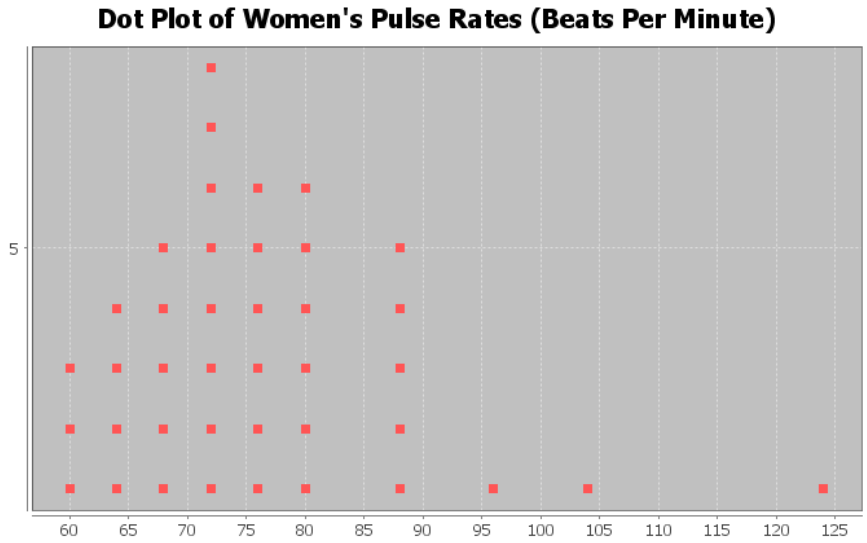
In case you are wondering, here are the formulas used by computer programs to determine outliers in a box plot. You do not need to calculate this yourself. The computer has already found your unusual values.

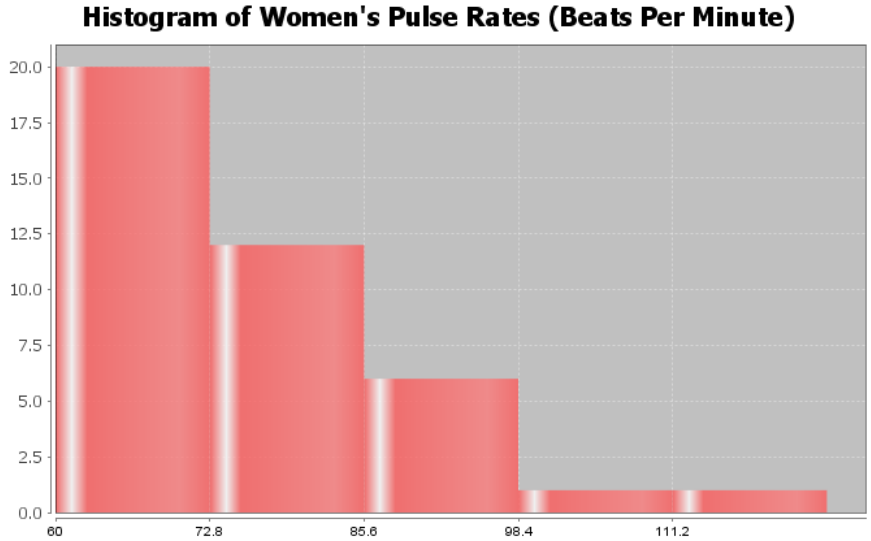
Unusual high (high outlier) cutoff: $Q_3 + (1.5IQR)$

Unusual low (low outlier) cutoff: $Q_1 - (1.5IQR)$

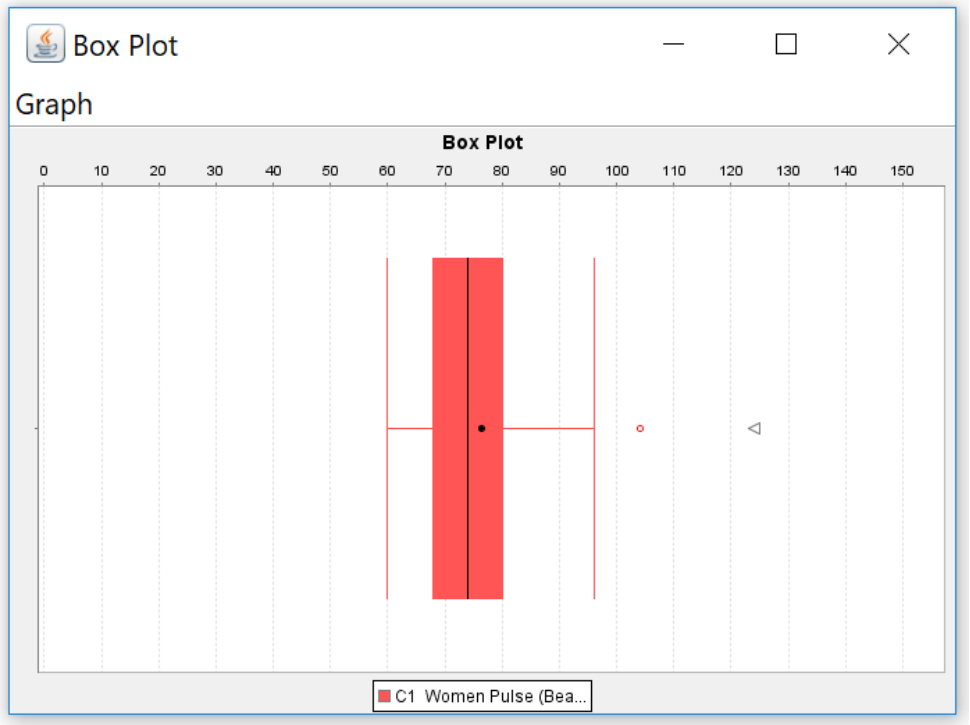
Note about box plots and normal data: Remember, a box plot is a graph of the quartiles and the median. They work really well for data that is not normal. However, they do not show the mean or standard deviation, so it is important to be careful how you interpret box plots for normal data. Normal data has different characteristics than those shown on a box plot. For example, typical values for normal data are <u>not</u> between Q_1 and Q_3. In addition, the outlier cutoffs are different for normal data so there may be differences in what is considered an outlier.

In the last section, we saw that we could also calculate dot plots, histograms, box plots and summary statistics with Statcato. Copy and paste the data into a column of Statcato. Then go to the graph menu and click on "dot plot", "histogram" or "box plot".





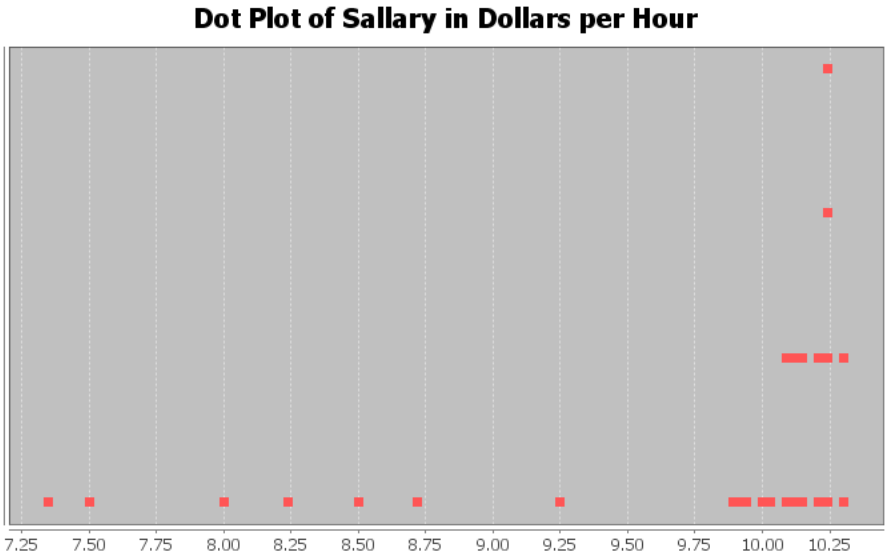
Notice that something is wrong with the Statcato box plot. The outliers have been left off. This is a common problem. To fix this, right click on the box-plot. Click on "zoom out" and "range axis". You may have to do this multiple times. You want to be able to see the minimum value (60 bpm) and maximum value (124 bpm) on the scale of the graph. Here is the correct box plot.



Notice Statcato designated 104 with a circle (regular outlier) and 124 with a triangle (far out outlier). The dot in the middle of the box plot is the mean. Most box plots do not have the mean, but Statcato puts it in so that you can compare it to the median.

Let us look at some other examples.

Here is some salary data from a small company with 26 employees. The salaries are given in dollars per hour. We created a dot plot and histogram for this data.

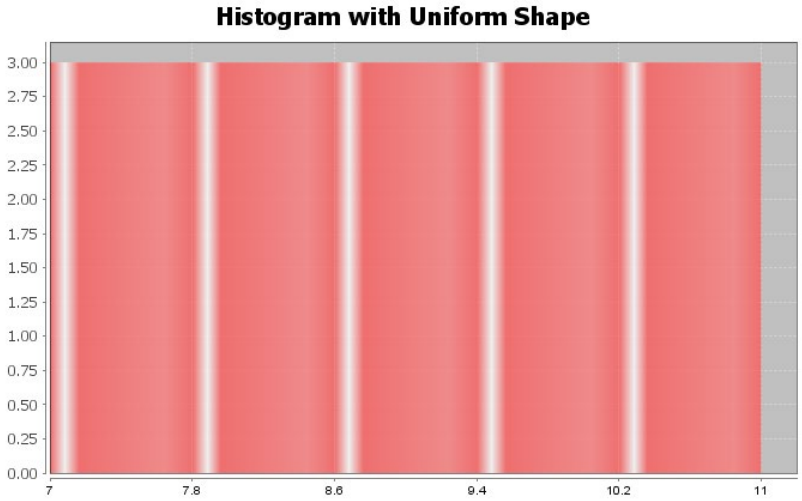


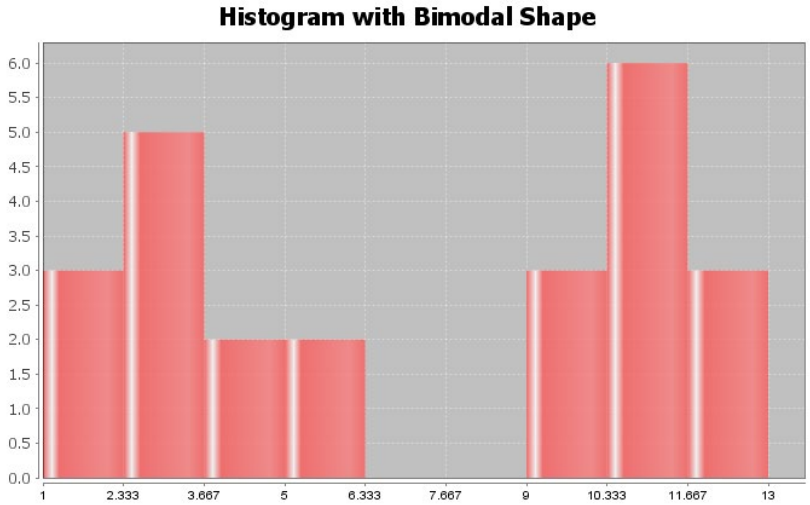
Notice the highest bar and most dots are on the far right, while there is a long tail to the left. Therefore, this is called "skewed left" or "negatively skewed".

Note: *Real data rarely has a perfect shape. Most data has a shape somewhere in between bell shaped and skewed, and you will need to make a decision. Look for a significant difference in the length of the tail to classify something as skewed. If my highest hill is toward the middle and I had two bars to the right and three bars to the left of the highest bar, I would still classify that bell shaped or normal. Some say that is "nearly normal". If the highest hill is on the far right and I have two bars to the right of the highest hill and seven bars to the left of the highest hill, I would classify that as skewed left. Some call this "negatively skewed" since negative numbers are to the left on the number line.*

Here are a couple unusual shapes that sometimes appear.

A graph that looks like a rectangle is called "uniform". A graph with two distinct high bars is called "bimodal".





Summary Statistics: Measures of Center, Spread and Position

Though the mean, median, standard deviation and IQR are used most often in data analysis, there are many different types of statistics that can be used to dig deeper into the data. We will not be covering these statistics in depth, but it is good to at least have an idea of what they measure.

Measures of Center

Mean Average: The balancing point in terms of distances. The measure of center or average used when a data set is bell shaped (normal).

Median Average: The center of the data in terms of order. Also called the second quartile (Q2) or the 50th percentile. Approximately 50% of the data will be less than the median and 50% will be above the median. This is the measure of center or average used when a data set is skewed (not bell shaped).

Mode: The number that occurs most often in a data set. Data sets may have no mode, one mode, or multiple modes. It is also sometimes used in bimodal or multimodal data.

Midrange: A quick measure of center that is usually not very accurate, but can be calculated quickly without a computer. (Max + Min) / 2

Measures of Spread

Standard Deviation: How far typical values are from the mean in a bell shaped data set. It is the most accurate measure of spread for bell shaped data. If you add and subtract the mean and standard deviation, you get two numbers that typical values in a bell shaped data set fall in between. It can also be used to find unusual values in bell shaped data. Should not be used unless the data is bell shaped.

Variance: The standard deviation squared. A measure of spread used in ANOVA testing. Only accurate when the data is bell shaped.

Range: A quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical values in the data set. It can be calculated quickly without a computer. (Max – Min)

Interquartile range (IQR): How far typical values are from each other in a skewed data set. Measures the length of the middle 50% of the data. It is the most accurate measure of spread for skewed data sets. Should not be used when data is bell shaped. (Q3-Q1)

Measures of Position

Minimum: The smallest number in the data set. Is sometimes classified as an unusual value (outlier).

Maximum: The largest number in the data set. Is sometimes classified as an unusual value (outlier).

First Quartile (Q1): The number that approximately 25% of the data is less than and 75% of the data is greater than. Used for finding typical values for skewed data sets.

Third Quartile (Q3): The number that approximately 75% of the data is less than and 25% of the data is greater than. Used for finding typical values for skewed data sets.

Frequency or Sample Size (n)

The frequency or sample size of a data set (n) is not a measure of center, spread or position, but is important bit of information. It tells us how many numbers are in the data set.
