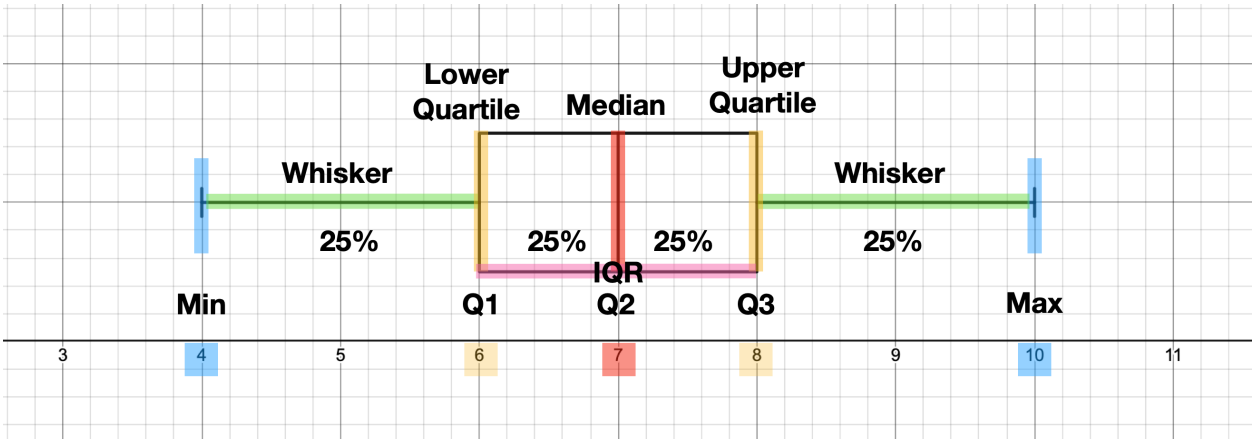
Box Plots and Distribution Skewness

Left Skew, Right Skew, Normal, Inconclusive in DESMOS

Box Plot Definitions and the 5-Number Summary

Box plots are useful to provide a visual summary of the data enabling researchers to quickly identify summary values, dispersion of the data set, and signs of skewness.



Minimum Score- The lowest Score, excluding outliers, shown at the end of the left whisker.

Lower Quartile- 25% of the data falls below the lower Quartile value. This is known as the first quartile.

Median- The median marks the midpoint of the data set and is shown by the line that divides the box into two parts. This is also known as the second quartile. 50% of the data is greater than or equal to this value and 50% are less than or equal to this value.

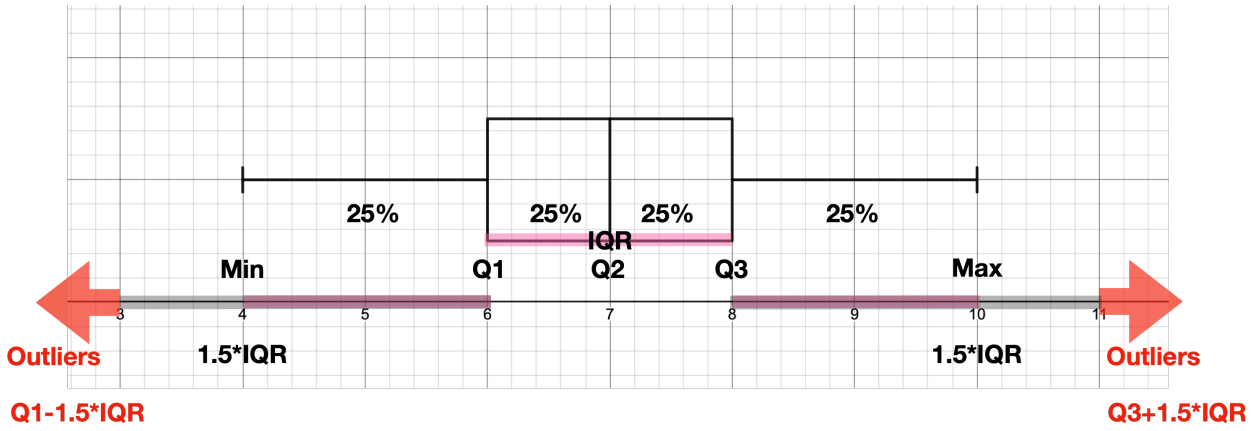
Upper quartile- 75% of the data falls below the upper quartile value and this is also known as the third quartile. This also means 25% of the data is above this value.

Maximum Score- The highest score, excluding outliers, shown at the end of the right whisker.

Outliers

Box plots are useful as they can show outliers from a distribution. An **outlier** is a data value that is numerically distant from the rest of the data set. It is a data value that is outside the whiskers of a box plot. It is **outside** $\mathbf{1.5} * \boldsymbol{IQR}$ both below the lower quartile and above the upper quartile. It is a data value outside the following interval.

$$(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$$

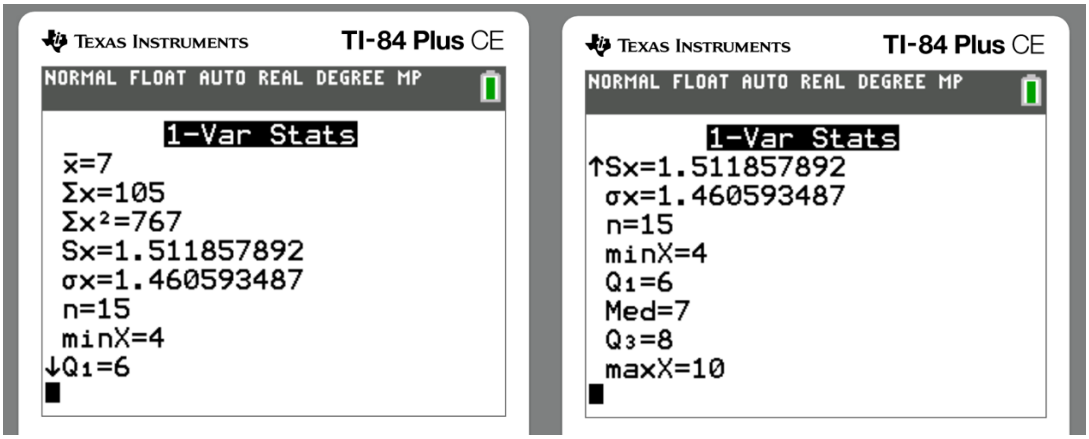


Q: How much sleep (hours) did you get last night?

The following sample data was collected in hours.

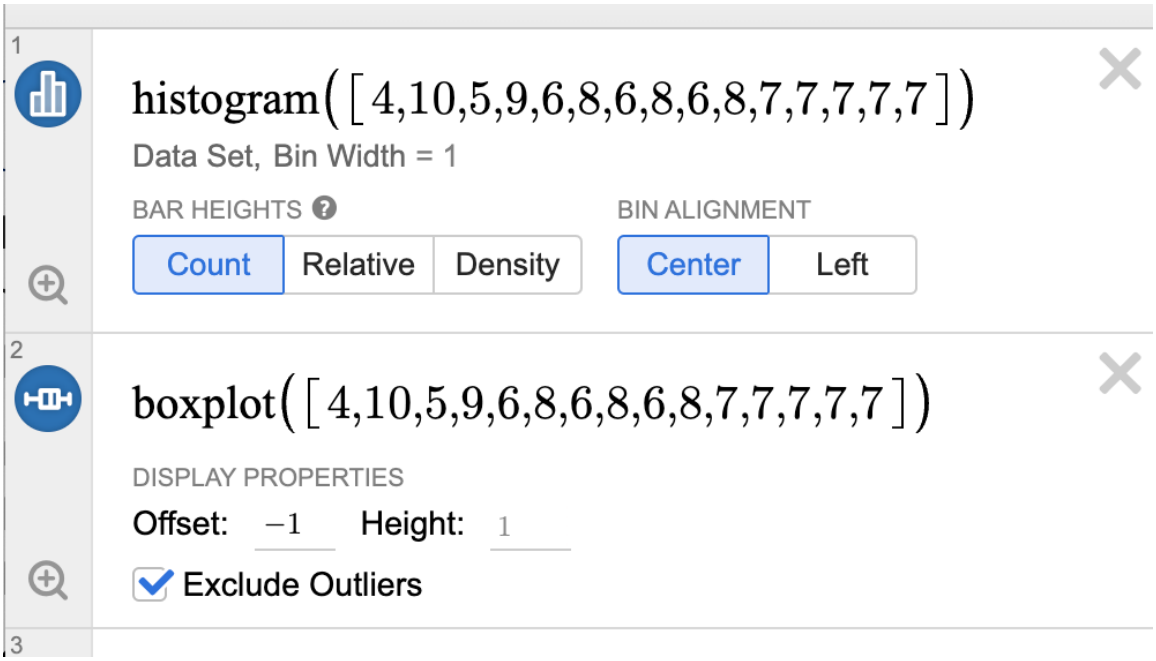
4,10,5,9,6,8,6,8,6,8,7,7,7,7,7,7

We can use the **TI-84 Plus CE Calculator** to determine the **5-number summary**.

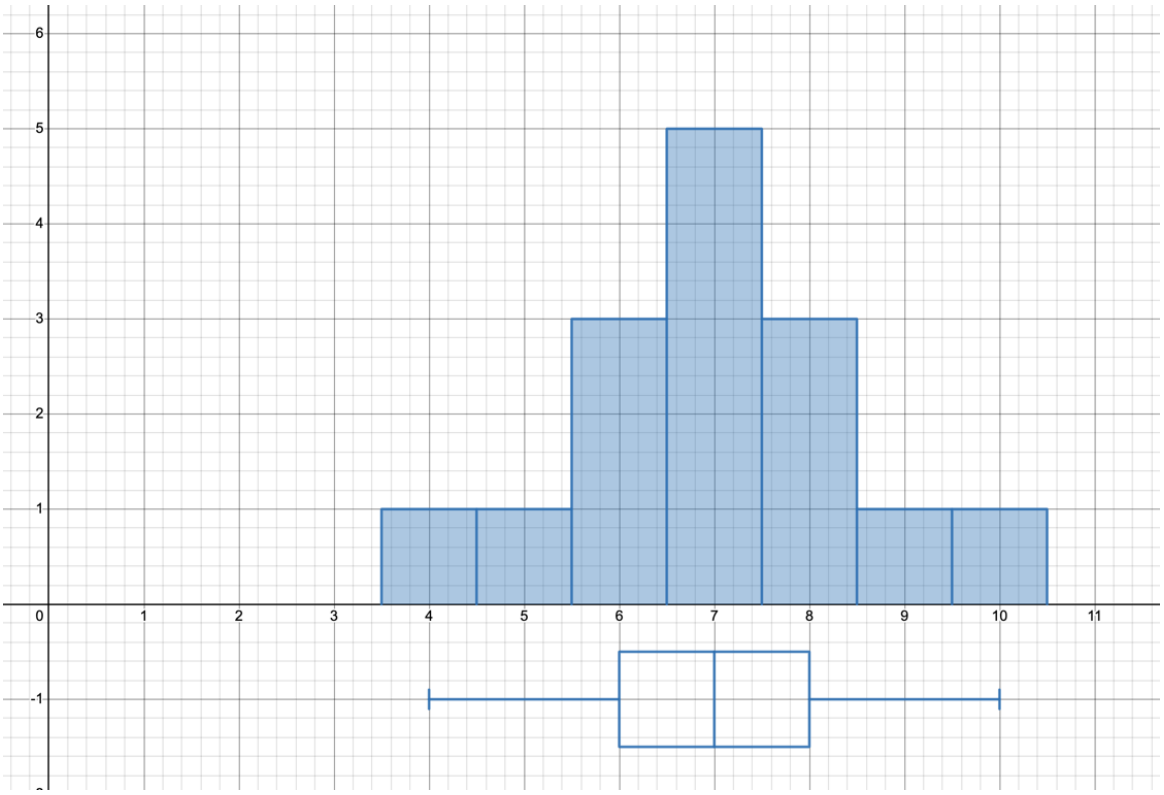


5-Number Summary: 4, 6, 7, 8, 10

We can also use DESMOS to draw a picture a **Histogram** and a **Box Plot**.



Normal (Most of the Data is in the Center)



Observations

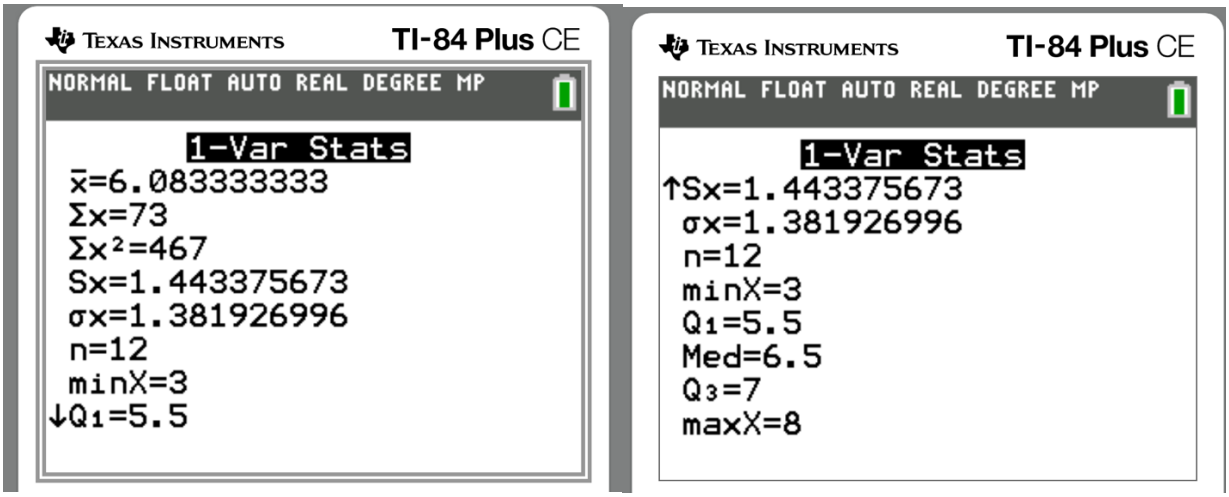
- Most of the data is in the center of your Histogram and the data is considered symmetrically distributed.
- The left whisker and left box are the **same lengths** as the right whisker and right box; therefore, the distribution is symmetric, which also known as **Normal.**
- **Mean = Median=Mode** that is **7=7=7**

Q: How much sleep (hours) did you get last night?

The following sample data was collected in hours.

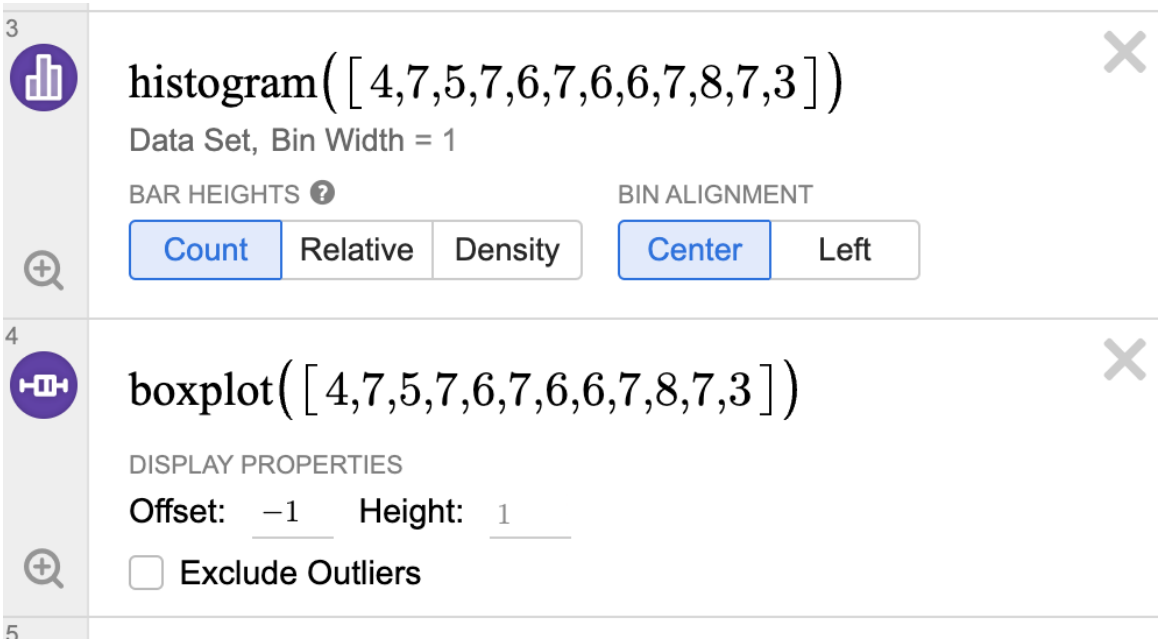
4,7,5,7,6,7,6,6,7,8,7,3

We can use the **TI-84 Plus CE Calculator** to determine the **5-number summary**.

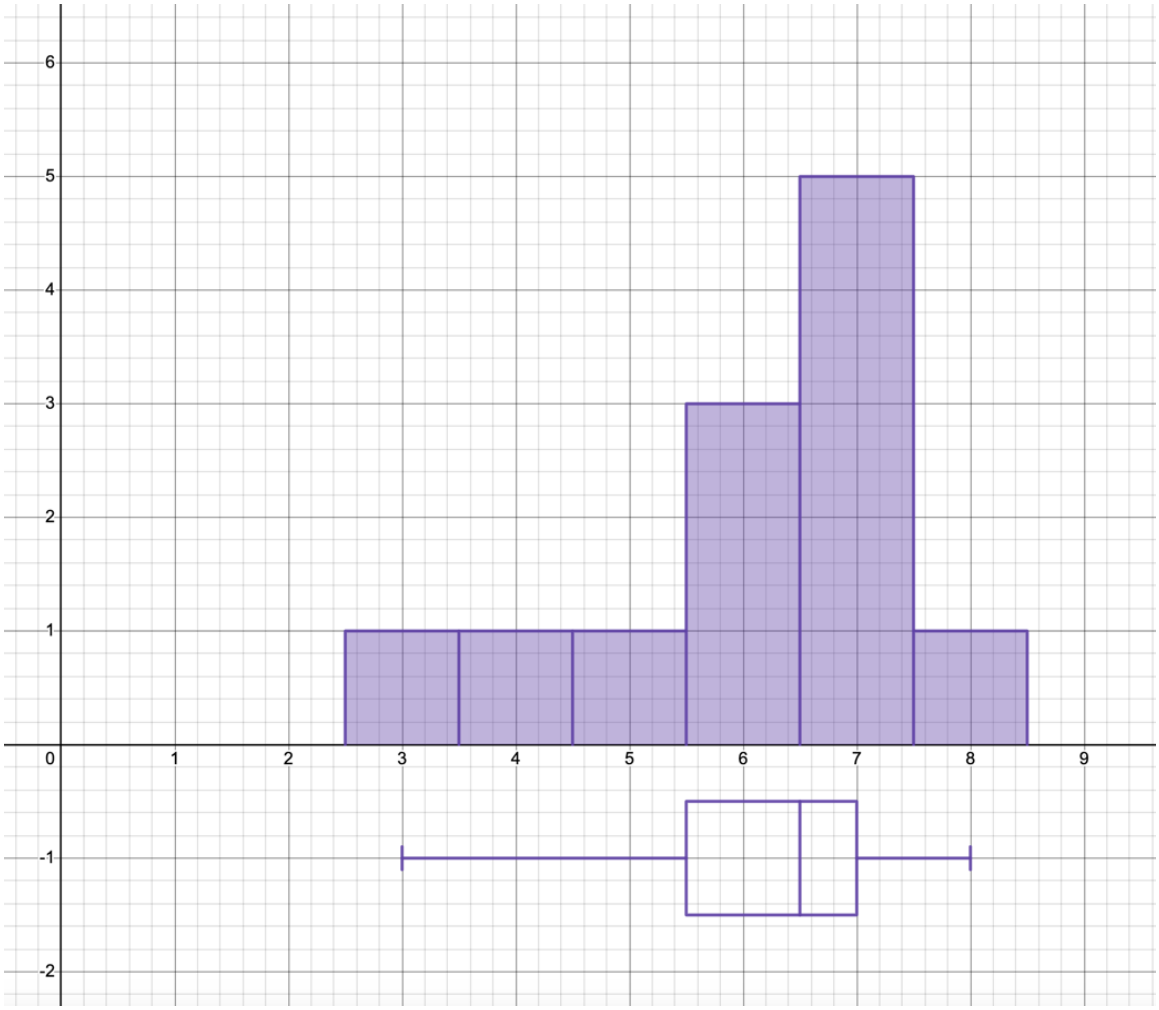


5-Number Summary: 3, 5.5, 6.5, 7, 8

We can also use DESMOS to draw a picture a **Histogram** and a **Box Plot**.



Left Skew (Most of the Data is on the Right)



Observations

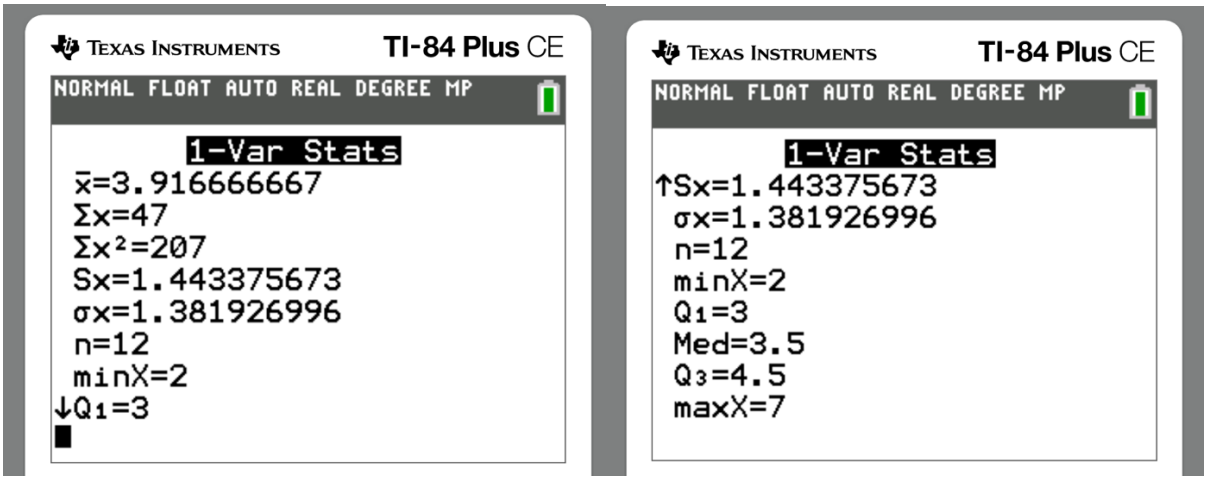
- Most of the data is to the right in your **Histogram** and the data is considered non-symmetrically distributed. It looks like the data is pulled to the left; therefore, it is **Left Skewed.**
- The left whisker and left box are longer than the right whisker and right box in the **box plot**. This indicates a **Left Skewed distribution**.
- **Mean < Median<Mode** that is **6.1<6.5<7**

Q: How much sleep (hours) did you get last night?

The following sample data was collected in hours.

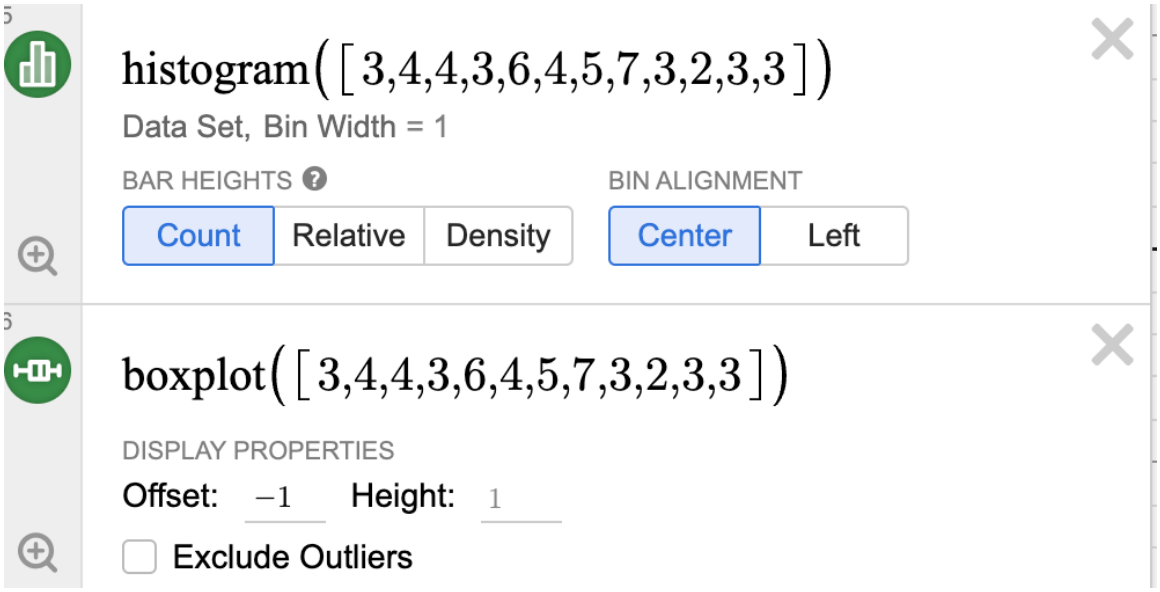
3, 4, 4,3,6,4,5,7,6,2,3,3

We can use the **TI-84 Plus CE Calculator** to determine the **5-number summary**.

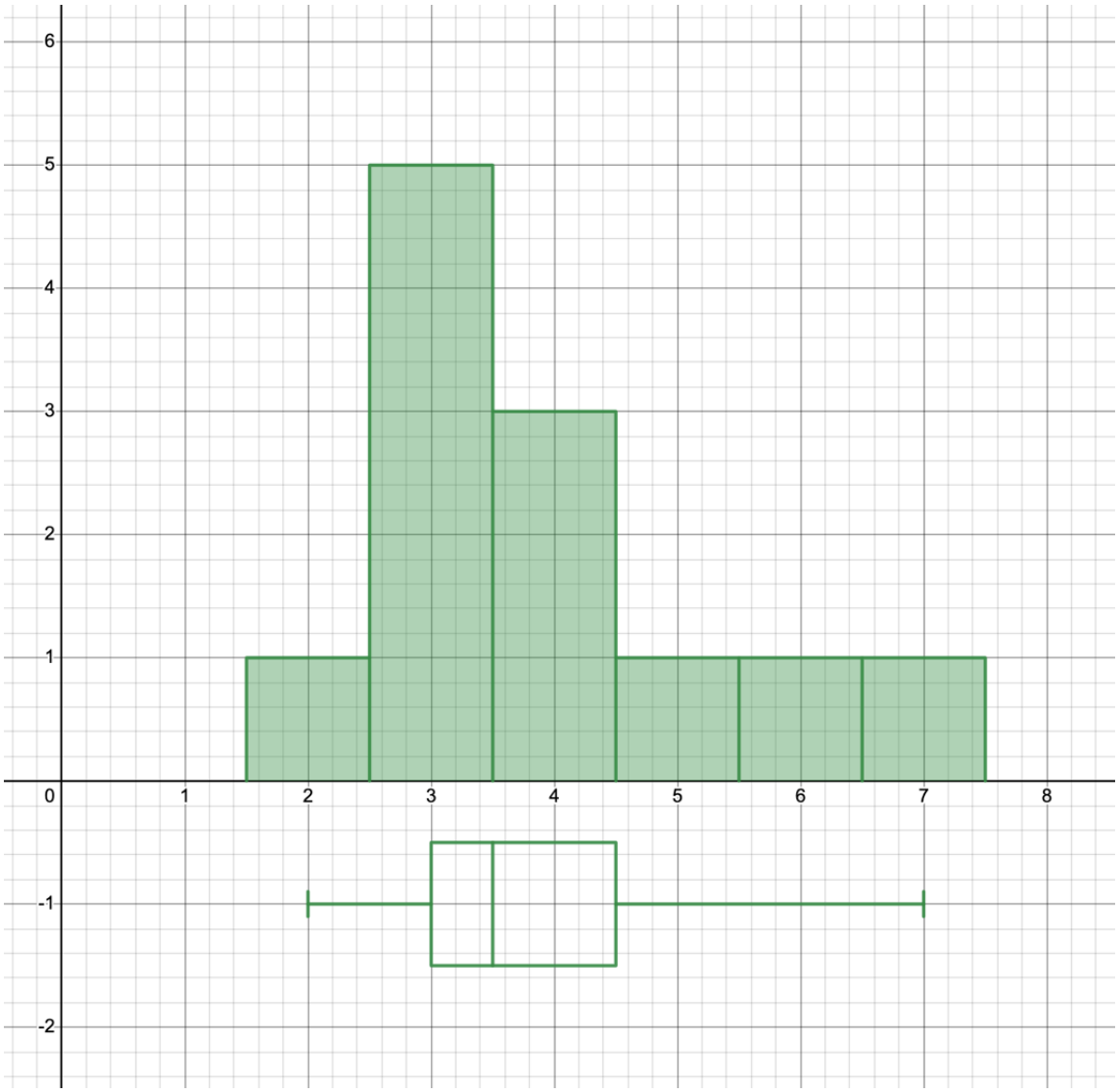


5-Number Summary: 2, 3, 3.5, 4.5, 7

We can also use DESMOS to draw a picture a **Histogram** and a **Boxplot**.



Right Skew (Most of the Data is on the Left)



Observations

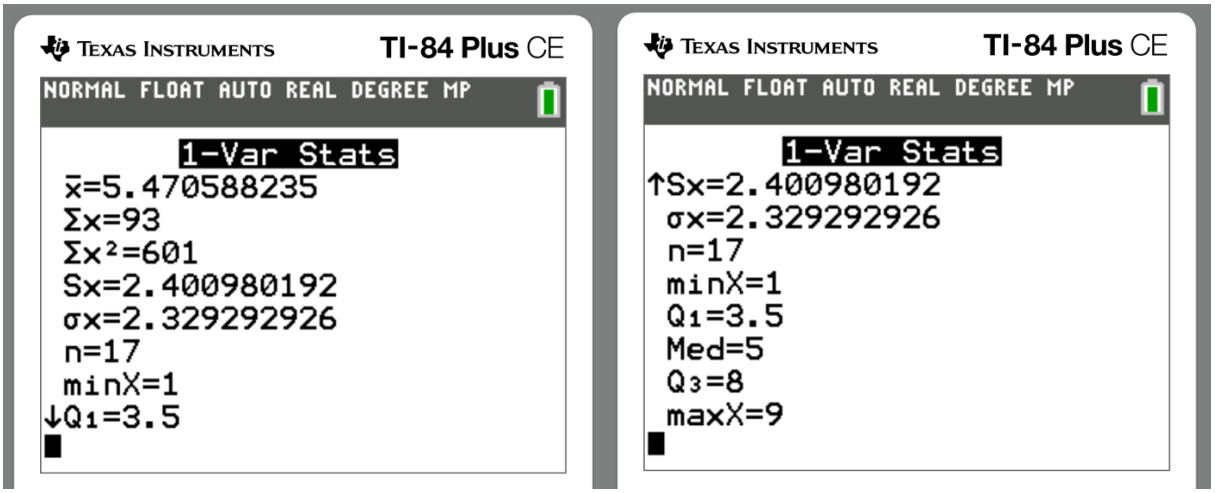
- Most of the data is to the right in your **Histogram** and the data is considered non-symmetrically distributed. It looks like the data is pulled to the right; therefore, it is **Right Skewed.**
- The Right whisker and right box are longer than the left whisker and left box in the **box plot**. This indicates a **Right Skewed distribution**.
- **Mean >Median>Mode** that is **3.9>3.5>3**

Q: How much sleep (hours) did you get last night?

The following sample data was collected in hours.

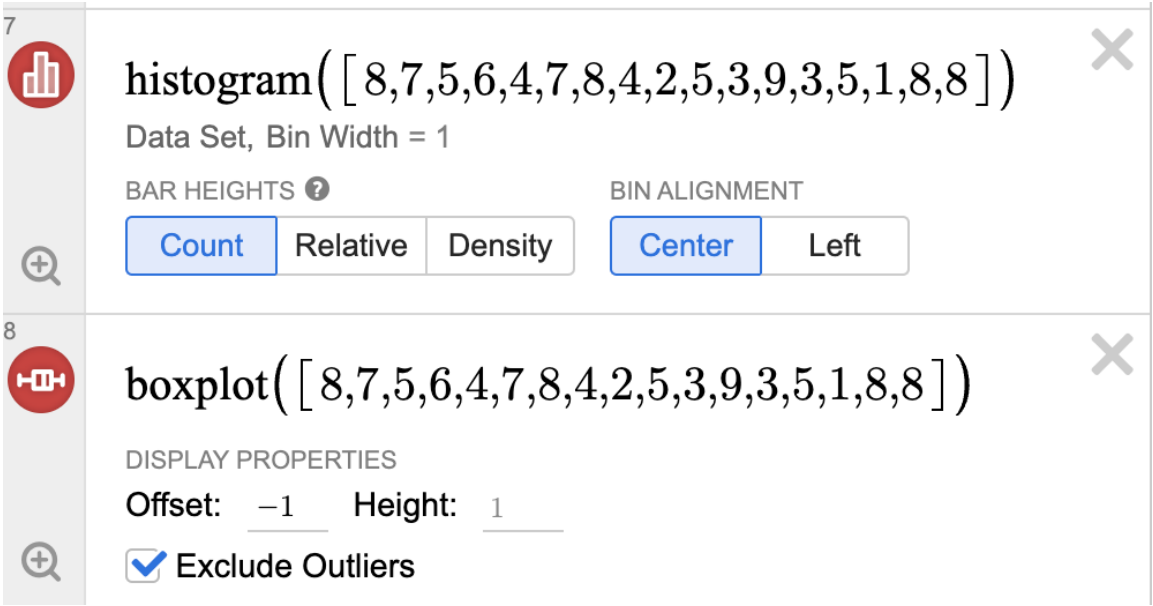
8,7,5,6,4,7,8,4,2,5,3,9,3,5,1,8,8

We can use the **TI-84 Plus CE Calculator** to determine the **5-number summary**.

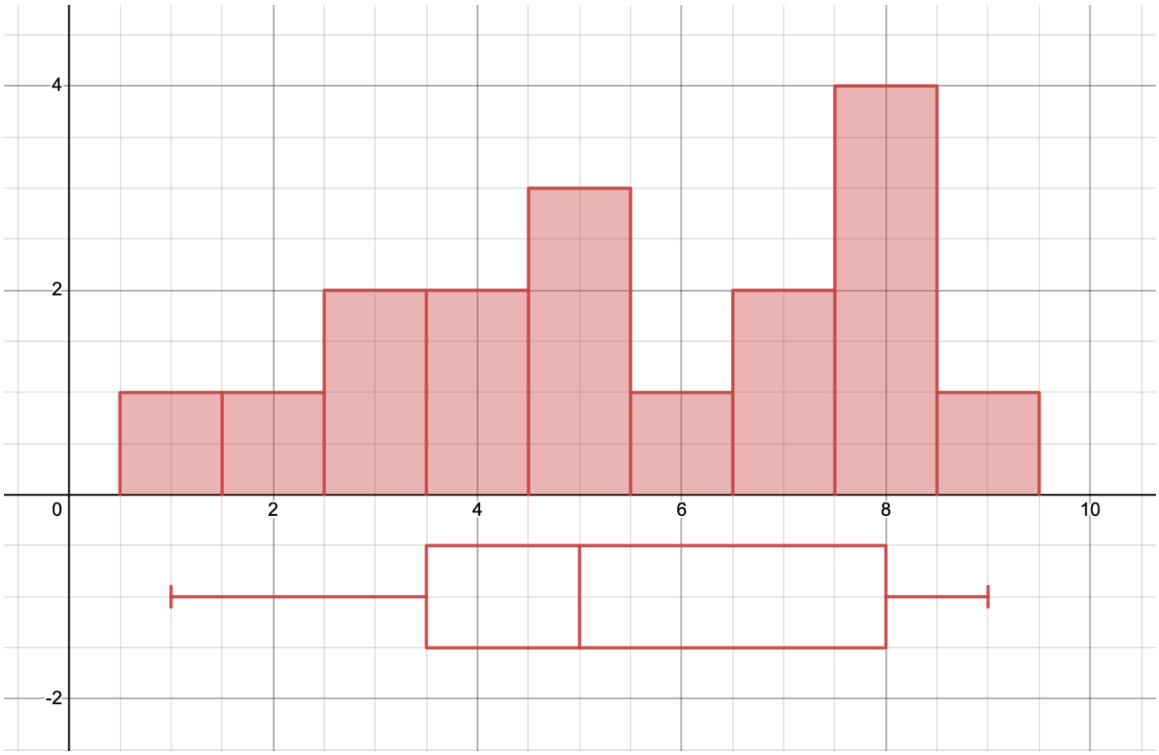


5-Number Summary: 1, 3.5, 5, 8, 9

We can also use DESMOS to draw a picture a **Histogram** and a **Box plot**.



Inconclusive



Observations

- The data almost looks uniform in your **Histogram** and no symmetry can be observed.
- The left whisker is longer than the right whisker, but the right box is shorter than the left box in the **box plot**. This indicates a contradiction, and we cannot conclude any symmetric or non-symmetry aspects to the distribution. Therefore, we have an inconclusive graphical description of this distribution.
- **Mean>Median<Mode** that is **5.5>5<8**

Comparing Boxplots

Family Size and Mother's Education
Keyfitz, 1953, American Journal of Sociology

The following two data sets represents the number of children for two separate groups of mothers. Those who are considered not educated (no more than 6 years of education) and those who are considered educated (at least 7 years of education).

Mother Educated for 6 years or less:

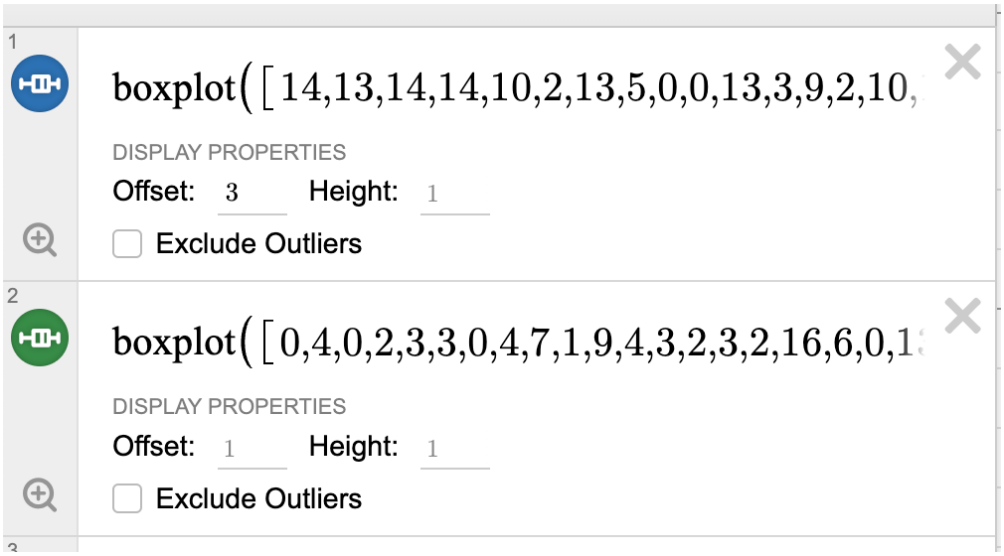
14,13,14,14,10,2,13,5,0,0,13,3,9,2,10,11,13,5,14

Mother educated for at least 7 years:

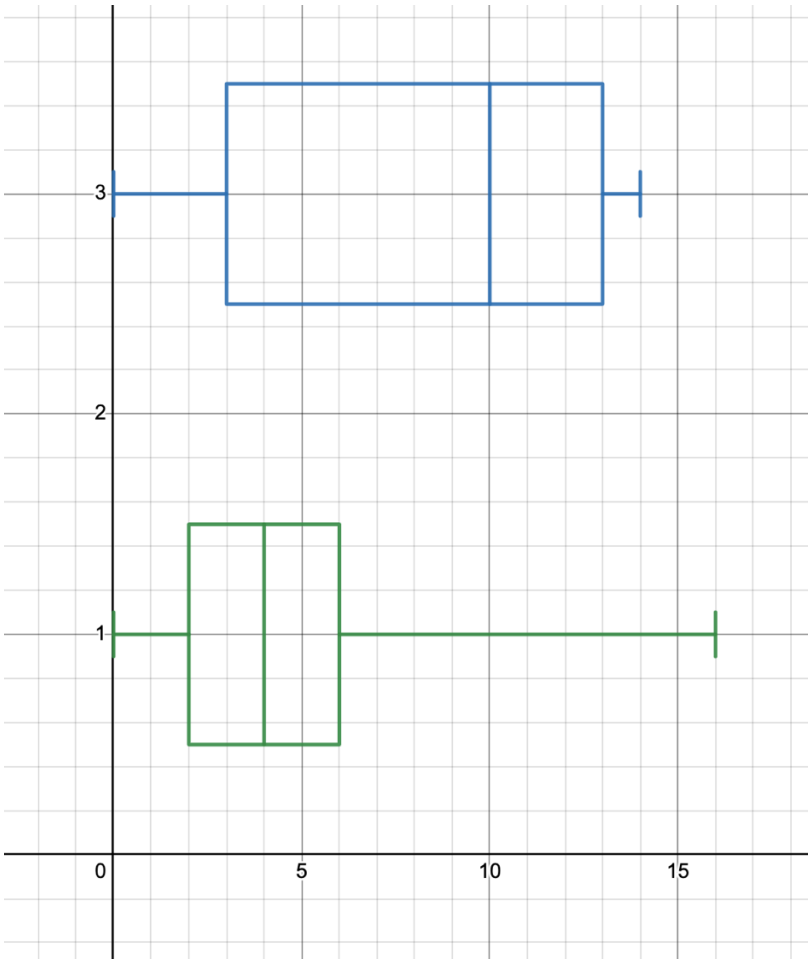
0,4,0,2,3,3,0,4,7,1,9,4,3,2,3,2,16,6,0,13,6,6,5,9,10,5,4,3,3,5,2,3,5,15,5

1. Compute the 5-number summary (**TI-84 Plus CE Calculator**) and construct a Box plot for both sets of data in **DESMOS**.

2. Describe the distribution for both sets of data (left Skew, Right Skew, Normal, inconclusive).



Mother Educated for no more than 6 years.
Mother educated for at least 7 years.



Guidelines for Comparing Box Plots

- Compare the respective medians to compare location.
- Compare the interquartile ranges (box lengths) to compare dispersion.
- Look at the overall spread by the adjacent values.
- Look for signs of skewness if data is not symmetric.
- Look for potential outliers.

Infants with SIRDS

Infants with severe idiopathic raspatory distress syndrome and their birth weights. Is it possible to relate the chances of eventual survival to their birth weight (kg)? The asterisks data denote the weight of the infants who have died.

Birthweight of Infants with SIRDS who Died (kg)

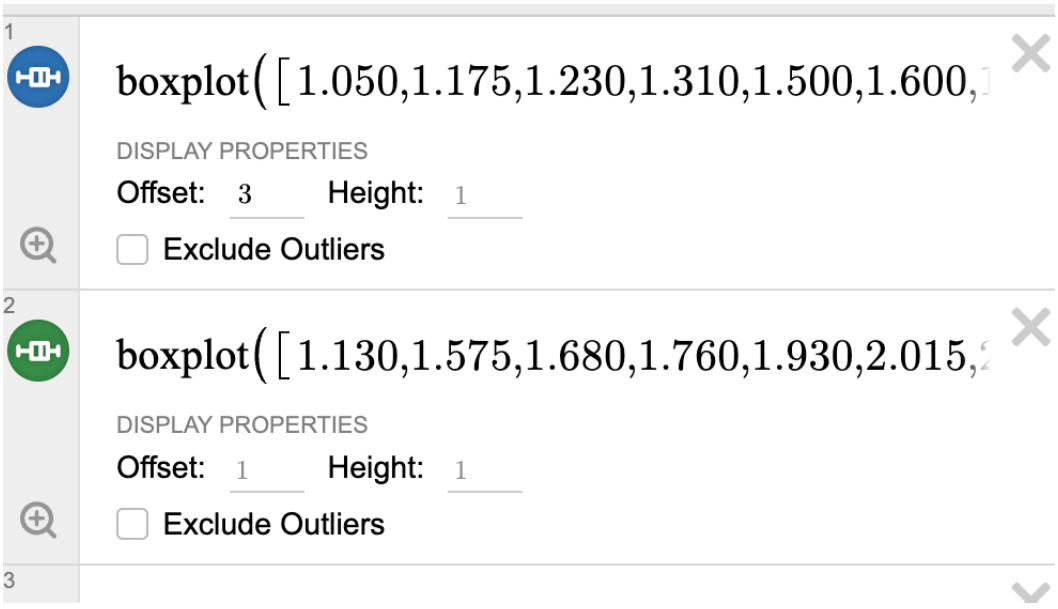
1.050, 1.175, 1.230, 1.310, 1.500, 1.600, 1.720, 2.275, 2.500, 1.030, 1.100, 1.185, 1.225, 1.262, 1.295, 1.300, 1.550, 1.820, 1.890, 1.940, 2.200, 2.270, 2.440, 2.560, 2.730

Birthweight of Infants with SIRDS who Survived (kg)

1.130, 1.575, 1.680, 1.760, 1.930, 2.015, 2.090, 2.600, 2.950, 2,700, 3.160, 3.400, 3.640, 2.830, 1.410, 1.715, 1.720, 2.040, 2.200, 2.400, 2.550, 2.570, 3.005

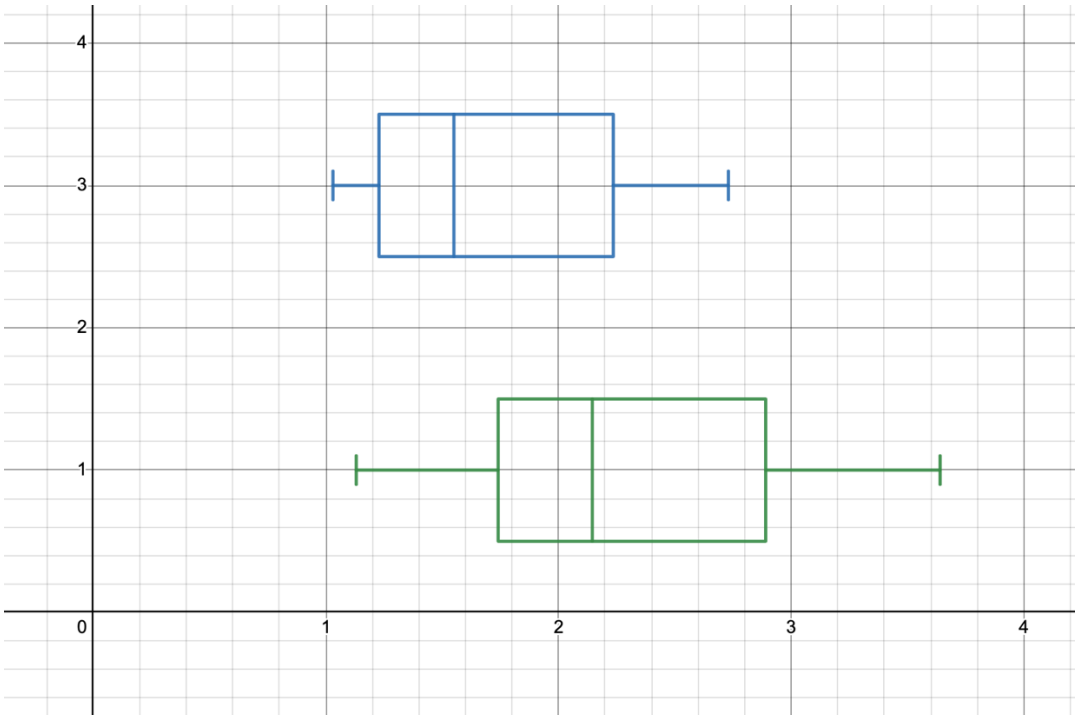
1. Compute the **5-number summary** (TI-84 Plus CE Calculator) and construct a Box Plot for both sets of data in **DESMOS**.

2. Describe the distribution for both sets of data (left Skew, Right Skew, Normal, Inconclusive).



Guidelines for Comparing Box Plots

- Compare the respective medians to compare location.
- Compare the interquartile ranges (box lengths) to compare dispersion.
- Look at the overall spread by the adjacent values.
- Look for signs of skewness if data is not symmetric.
- Look for potential outliers.



When using your **TI-84 Plus CE** and **DESMOS** to evaluate your data, you should make the following conclusions.

Comparison of location shows that the median birth weight of infants who survived is greater than that of those who died.

Comparison of dispersion*:* The interquartile ranges are reasonably similar (as shown by the lengths of the boxes), though the overall range of the data set is greater for the surviving infants (as shown by the distances between the ends of the two whiskers for each boxplot).

Comparison of skewness*:* Though both batches of data appear to be right-skew, and the batch for the infants who died is slightly more skewed than that for those who survived, the skewness is not particularly marked in either case. (In fact, the sample skewness for the birth weights of the infants who survived is 0.25; and for the infants who died, it is 0.53. Both skewness's are positive; the value for the infants who died is rather larger, corresponding to a more marked lack of symmetry, but neither skewness is particularly large.)

***Comparison of potential outliers**:* Neither data set shows any suspiciously far out values which might require a closer look.

***General conclusions**:* Overall, the two batches of data look as if they were generally distributed in a similar way, but with one batch located to the right (larger location) of the other. You can see immediately that the median birth weight of infants who died is less than the lower quartile of the birth weights of infants who survived (that is, over three-quarters of the survivors were heavier than the median birth weight of those who died). So, it looks as if we can safely say that survival is related to birth weight.

You can see how comparative boxplots give a compact, quickly assimilated summary of the data, suggesting that infants who survive and infants who do not may typically have different birth weights.