Stat Support Activity: Box Plots and Outliers

Notes

- If a data is not normally distributed (not bell shaped), it is usually skewed right (long right tail) or skewed left (long left tail).
- When data is skewed or not normal, we use the Median (Q_2) for the average, Quartiles Q_1 and Q_3 for typical values, and the Interquartile Range (IQR) for typical spread.
- **Box Plot:** A Box Plot (or Box and Whisker Plot) is a graph of the Median, Quartiles, IQR and Outliers. Box Plots are primarily used for skewed or non-normal data. Box Plots can be drawn horizontally with smaller numbers on the left and larger numbers on the right. Box Plots can also be drawn vertically with smaller number below and higher numbers above.
- **Creating a Box Plot:**
  - **Number Line:** Draw an evenly spaced, labeled number line going from the min and max in your data set.
  - **Box:** Draw a box between Q_1 and Q_3. (The length of your box is the typical spread corresponding to the Interquartile Range IQR. The box also shows you where typical values fall in your data set. Typical values are between Q_1 and Q_3.)
  - **Average:** Put a line in the box corresponding to the Median (Q_2). (This is the average.)
  - **Low Outlier Cutoff:** Calculate the low outlier cutoff $Q_1 - (1.5 \times IQR)$. Remember to calculate parenthesis first, then subtract in the correct order.
  - **Low Outliers:** Any numbers below the low outlier cutoff in your data set are considered unusually small (low outliers). Put stars on the Box Plot corresponding to all of the low outliers. Some data sets have a lot of outliers and some do not have any. If a Box Plot has no stars on the left (lower numbers), then the data set has no low outliers.
  - **High Outlier Cutoff:** Calculate the high outlier cutoff $Q_3 + (1.5 \times IQR)$. Remember to calculate parenthesis first, then add.
  - **High Outliers:** Any numbers above the high outlier cutoff in your data set are considered unusually large (high outliers). Put stars on the Box Plot corresponding to all of the high outliers. Some data sets have a lot of outliers and some do not have any. If a Box Plot has no stars on the right (higher numbers), then the data set has no high outliers.
  - **Lower (left side) Whisker:** Identify the lowest number in the data set that is not a low outlier. If there are no low outliers in the data this will be the smallest number in the data set (min). The lower (or left side) whisker is a line drawn from the box to the smallest number in the data set that is not an outlier. Note: The whisker does not go all the way to the low outlier cutoff. It must end at a number in the data set and that number cannot be an outlier.
  - **Upper (right side) Whisker:** Identify the largest number in the data set that is not a high outlier. If there are no high outliers in the data this will be the largest number in the data set (max). The upper (or right side) whisker is a line drawn from the box to the largest number in the data set that is not an outlier. Note: The whisker does not go all the way to the high outlier cutoff. It must end at a number in the data set and that number cannot be an outlier.

Stat Support Activity: Box Plots and Outliers

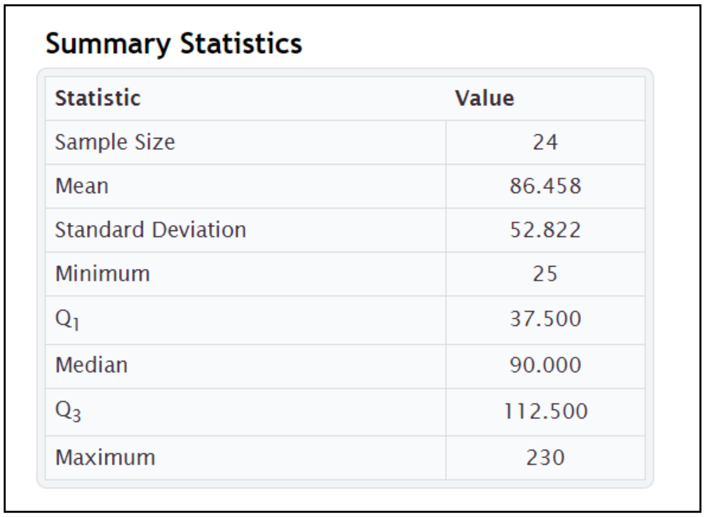
- **Whiskers can be missing:** It is possible for a Box Plot to be missing a whisker. This is rare and only happens when there are certain numbers in the data repeated over and over. If the largest number in the data set (max) and Q_3 happen to be the same number, then there would not be an upper (right side) whisker. If the smallest number in the data set (min) and Q_1 happen to be the same number, then there would not be a lower (left side) whisker.

Problems

1.

Here is a quantitative data set describing the milligrams of potassium per serving in cereals. The data has already been put in order from smallest to largest. Here are the summary statistics calculated from StatKey.

Potassium (milligrams per serving)
25
25
30
35
35
35
40
40
45
55
90
90
90
95
95
105
110
110
115
120
130
160
170
230



Summary Statistics

Statistic	Value
Sample Size	24
Mean	86.458
Standard Deviation	52.822
Minimum	25
Q_1	37.500
Median	90.000
Q_3	112.500
Maximum	230

Stat Support Activity: Box Plots and Outliers

a) Draw a labeled evenly spread out horizontal number line from the minimum value on left to the maximum value on the right.

b) Draw a box above your number line between Q_1 and Q_3. Remember, typical values in a non-normal or skewed data set are between Q_1 and Q_3. So your box shows you the typical values in your data.

c) Fill in the blanks for the following sentence: "Typical cereals in the data set have between _______ mg of potassium and _________ mg of potassium."

d) StatKey did not calculate the Interquartile Range (IQR). Use the formula $IQR = Q_3 - Q_1$ to calculate the IQR. Notice this is the length of your box!

e) Fill in the blanks for the following sentence: "Typical amounts of potassium in the cereal data are within _______ milligrams from each other."

f) Put a line in the box corresponding to the Median (Q_2).

g) Fill in the blank for the following sentence: "The average amount of potassium for the cereals in the data set is _______ milligrams."

h) Calculate the low outlier cutoff $Q_1 - (1.5 \times IQR)$. Remember to calculate parenthesis first, then subtract in the correct order.

i) Look at the Potassium data set in order. Are there any numbers smaller than the low outlier cutoff? These are low outliers (unusually low values). If so make a star at each of the low outliers.

j) Calculate the high outlier cutoff $Q_3 + (1.5 \times IQR)$. Remember to calculate parenthesis first, then add.

k) Look at the Potassium data set in order. Are there any numbers larger than the high outlier cutoff? These are high outliers (unusually large values). If so make a star at each of the high outliers.

l) What is the smallest number in the data set that is <u>not</u> an outlier? (This could be the min if there are no low outliers. If there are low outliers, look for the smallest number in the data that is <u>not</u> an outlier.)

m) Draw your lower (left side) whisker from the box to the lowest number in the data that is not an outlier.

n) What is the largest number in the data set that is <u>not</u> an outlier? (This could be the max if there are no high outliers. If there are high outliers, look for the largest number in the data that is <u>not</u> an outlier.)

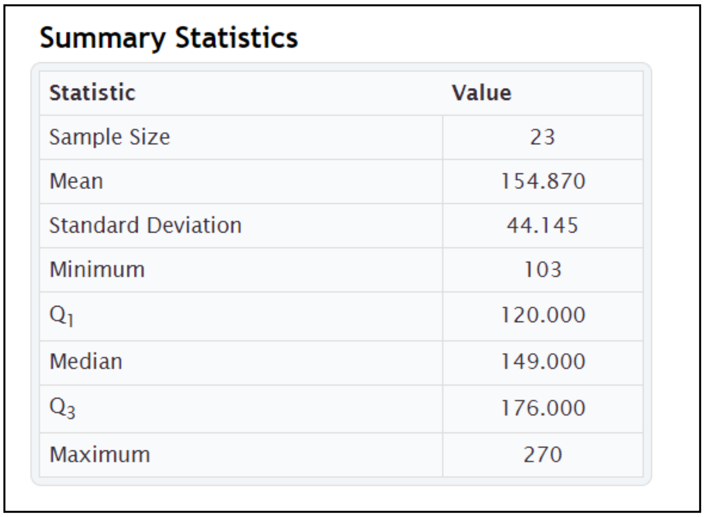
o) Draw your upper (right side) whisker from the box to the largest number in the data that is not an outlier. You are now done with drawing your Box Plot!

Stat Support Activity: Box Plots and Outliers

2.

Here is a quantitative data set listing the weights of some COC students.   The data has already been put in order from smallest to largest. Here are the summary statistics calculated from StatKey.

Weight (in pounds)
103
106
109
110
114
120
120
125
135
135
144
149
155
155
165
170
172
180
185
190
200
250
270



Summary Statistics

Statistic	Value
Sample Size	23
Mean	154.870
Standard Deviation	44.145
Minimum	103
Q_1	120.000
Median	149.000
Q_3	176.000
Maximum	270

a) Draw a labeled evenly spread out horizontal number line from the minimum value on left to the maximum value on the right.

b) Draw a box above your number line between Q_1 and Q_3.   Remember, typical values in a non-normal or skewed data set are between Q_1 and Q_3. So your box shows you the typical values in your data.

c) Fill in the blanks for the following sentence: “Typical weights for the COC students in the data set are between _______ pounds and _________ pounds.”

d) StatKey did not calculate the Interquartile Range (IQR). Use the formula $IQR = Q_3 - Q_1$ to calculate the IQR. Notice this is the length of your box!

e) Fill in the blanks for the following sentence: “Typical weights of these COC students are within _______ pounds from each other.”

f) Put a line in the box corresponding to the Median (Q_2).

Stat Support Activity: Box Plots and Outliers

g) Fill in the blank for the following sentence: "The average weight of the COC students in the data set is _______ pounds."

h) Calculate the low outlier cutoff $Q_1 - (1.5 \times IQR)$. Remember to calculate parenthesis first, then subtract in the correct order.

i) Look at the COC student weight data set in order. Are there any numbers smaller than the low outlier cutoff? These are low outliers (unusually low values). If so make a star at each of the low outliers.

j) Calculate the high outlier cutoff $Q_3 + (1.5 \times IQR)$. Remember to calculate parenthesis first, then add.

k) Look at the COC student weight data set in order. Are there any numbers larger than the high outlier cutoff? These are high outliers (unusually large values). If so make a star at each of the high outliers.

l) What is the smallest number in the data set that is not an outlier? (This could be the min if there are no low outliers. If there are low outliers, look for the smallest number in the data that is not an outlier.)

m) Draw your lower (left side) whisker from the box to the lowest number in the data that is not an outlier.

n) What is the largest number in the data set that is not an outlier? (This could be the max if there are no high outliers. If there are high outliers, look for the largest number in the data that is not an outlier.)

o) Draw your upper (right side) whisker from the box to the largest number in the data that is not an outlier. You are now done with drawing your Box Plot!