Box and Whisker Plots: Outliers

Five-number summary:

A summary of information to approximately describe a distribution. It identifies the following:

- **Maximum and Minimum**—give bounds for the data set—all values fall between them
- **Median**—a measure of center, "typical" value
- **Q1 and Q3**—measure of variability: $IQR = Q_3 - Q_1$

Can use median and quartiles together to get an indication of shape:

- May be skewed if median is much closer to one of the quartiles or if whiskers are not the same length (approximately)

Should not be used alone as indicator of shape of distribution—not enough detail— hides some features such as bimodality. Graphs to find shape of distribution: histogram, dotplot, stemplot.

To draw "full" or "modified" boxplot (generally used):

1. Identify 5-number summary
2. Calculate the Interquartile Range (IQR)
3. Calculate the "Thresholds." The upper and lower limits where data is considered outliers:
   a. $Q_3 + 1.5 \times IQR$ and $Q_1 - 1.5 \times IQR$
   b. Don't draw thresholds on boxplot—they are not data values. Only use them to identify outliers.
4. Identify which numbers are outliers
5. Construct a Box and whisker plot
   a. Draw a scale, or number line, that accommodates all data values. Include numbers and units. The plot may be vertical or horizontal.
   b. Draw rectangular box with ends at quartiles and a line through box at median.
   c. Draw two "whiskers" from corresponding ends of box to most extreme data value that is not an outlier—inside thresholds. Put dots or other marks for each outlier value.

Using Quartiles to Find Outliers: The "1.5 times IQR" rule

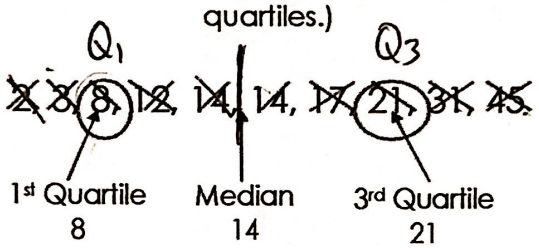
{3, 12, 17, 2, 21, 14, 14, 8, 45, 31}

Step 1: Identify 5-number summary

2, 3, 8, 12, 14, 14, 17, 21, 45, 31

Since there are two numbers in the middle we find the average. So the Median equals 14.

(Note: Since the median was the average of two numbers, we use those numbers when finding the quartiles. If there were only one number as the median we would NOT use it when finding the quartiles.)



Min = 2
$Q_1 = 8$
Med = 14
$Q_3 = 21$
$Q_5 = 45$
↑ 31 is new max

Step 2: Calculate the interquartile range

$Q_3 - Q_1$

3rd Quartile – 1st Quartile
= 21 – 8
IQR = 13

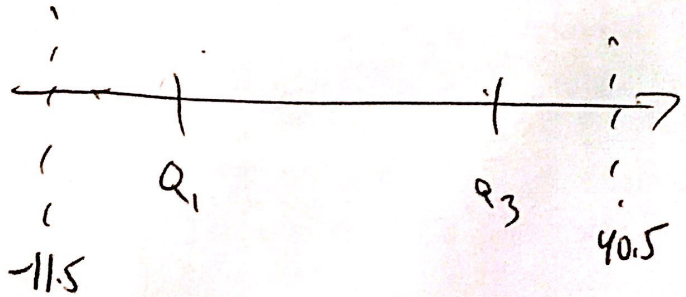
Step 3: Calculate the "Thresholds" – The upper and lower limits

Multiply your interquartile range by 1.5
13 • 1.5 = 19.5

Now find the upper and lower limits by adding this number to the 3rd Quartile value and also subtracting from the 1st Quartile value.

$Q_3 + 19.5 = 21 + 19.5 = 40.5$
$Q_1 - 19.5 = 8 - 19.5 = -11.5$

Upper Limit: 21 + 19.5 = 40.5
Lower Limit: 8 – 19.5 = –11.5



45 is bigger, so 45 is an outlier

Step 4: Identify outliers. Determine if any of your numbers are more/less than your limits.

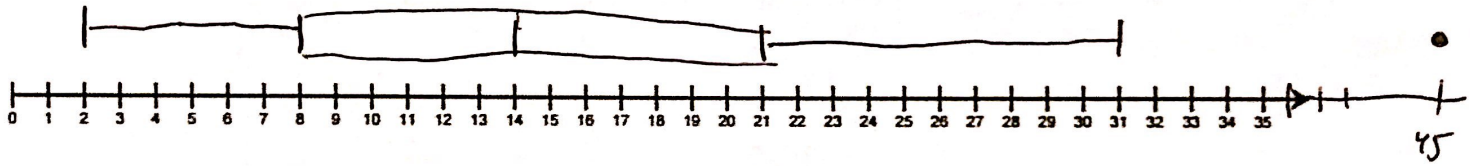
NO Yes

In this scenario, are there any numbers less than –11.5 or greater than 40.5? Looking at our original data in Step 1, we see that there are no numbers less than –11.5 but there is one number greater than 40.5

Therefore, our outlier is 45.

(Note: You can have no outliers, one outlier, or multiple outliers in any given set of numbers.)

Step 5: Construct a Box and whisker plot

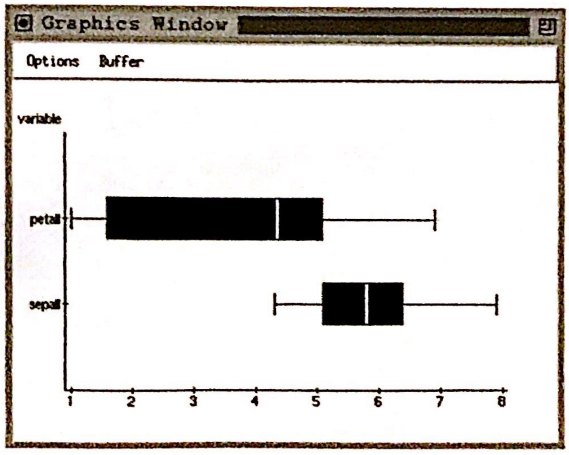


Why 1.5 x IQR?

John Tukey, inventor of boxplots, answered that one was not enough and two was too much.

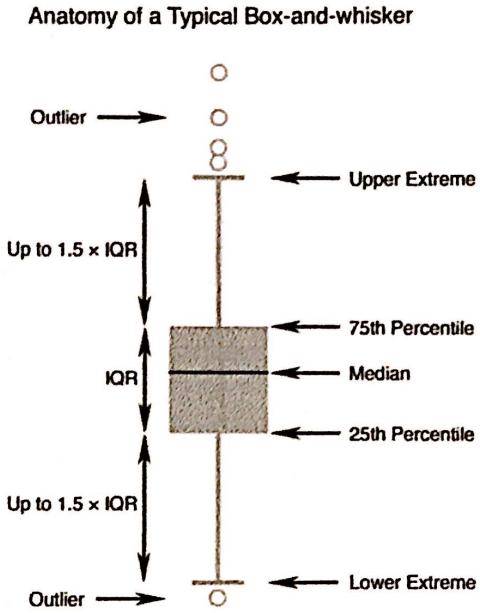
Multiple boxplots

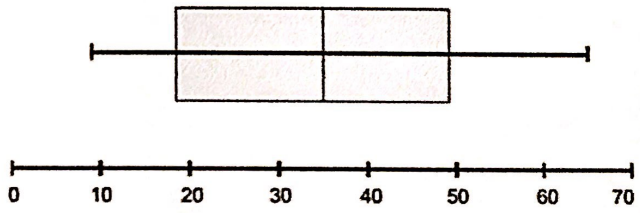
- Useful to compare several boxplots of related data on the same scale.
- Small differences may be meaningless, especially with small data sets.

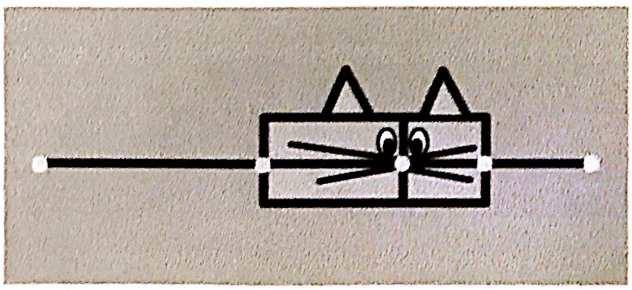


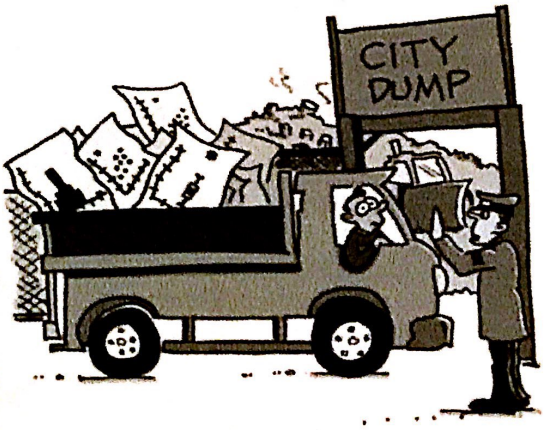
Percentiles

- Q_1 and Q_3 are the 25th and 75th percentiles respectively.
- Q_1 — 25th percentile means 25th percentile means 25% of data at or below this value.
- Q_2— median —50th percentile
- Q_3—75th percentile means 75% of data at or below this value.









"Sorry, we don't let people discard outliers without a good reason."