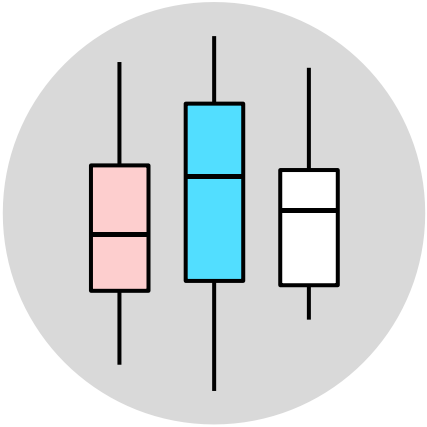
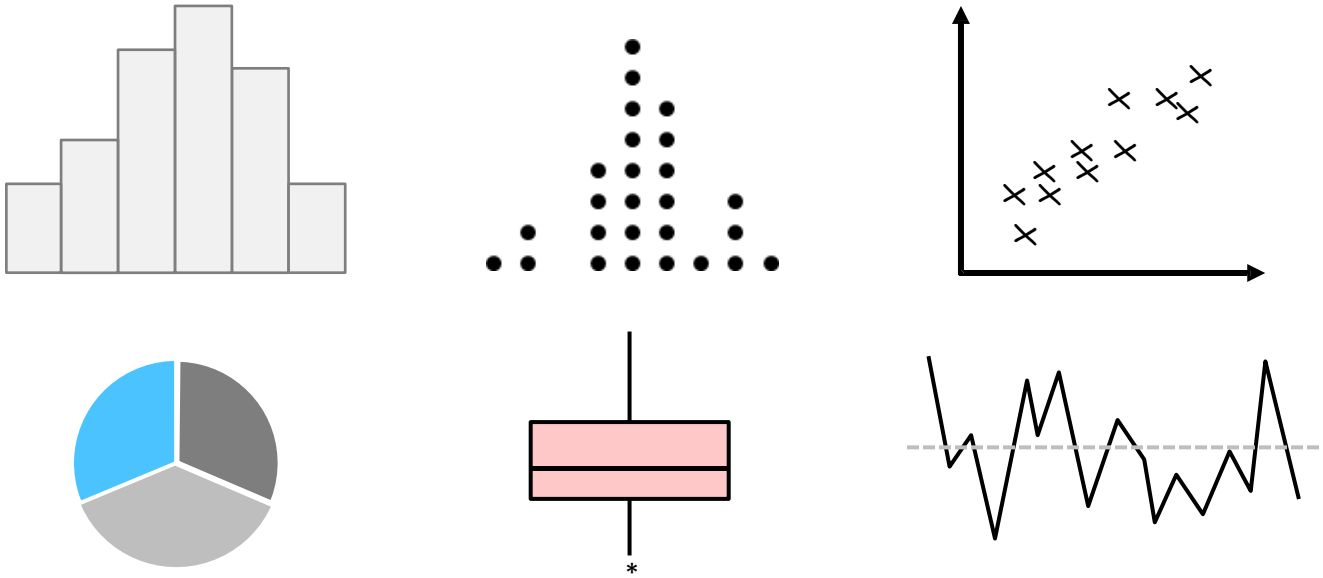
Box Plot



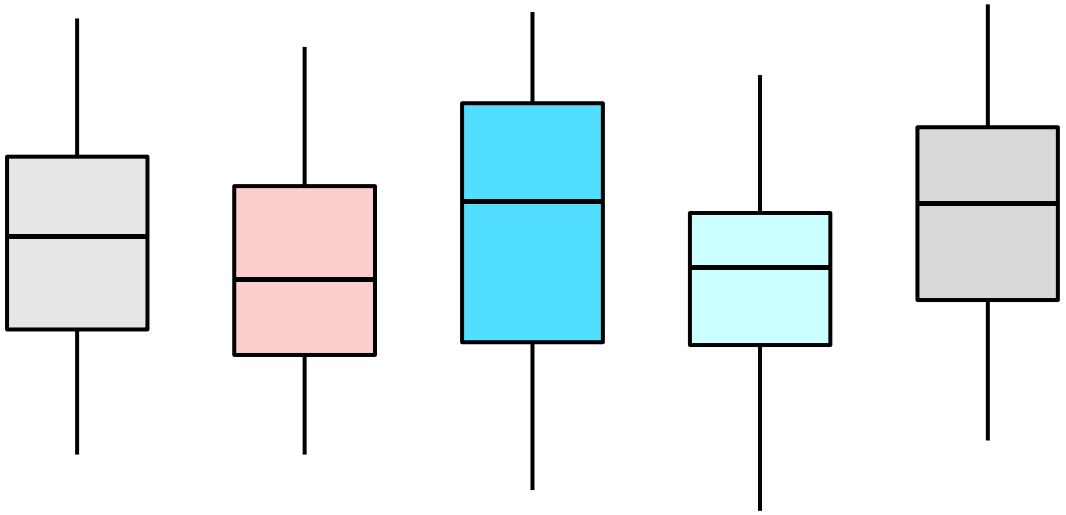
Box Plot

One of the best ways to analyze any process is to **plot the data**



Box Plot

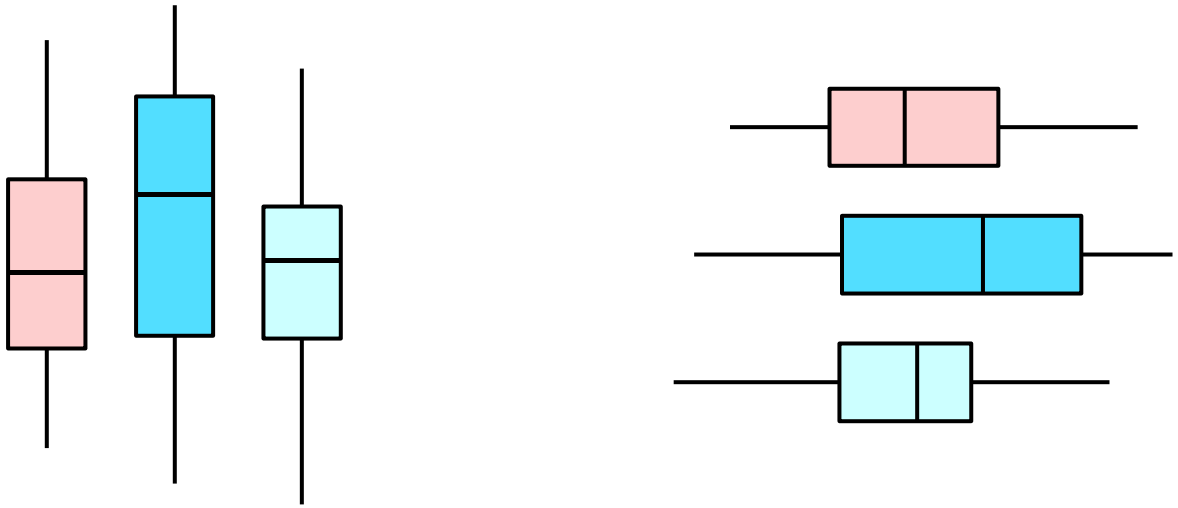
A **box plot** is a graphical way that summarizes the important aspects of the distribution of numeric data



Also referred to as a **Box-and-Whisker Plot** as it displays the data in a box-and-whiskers format

Box Plot

Box plots can be drawn either **vertically** or **horizontally**

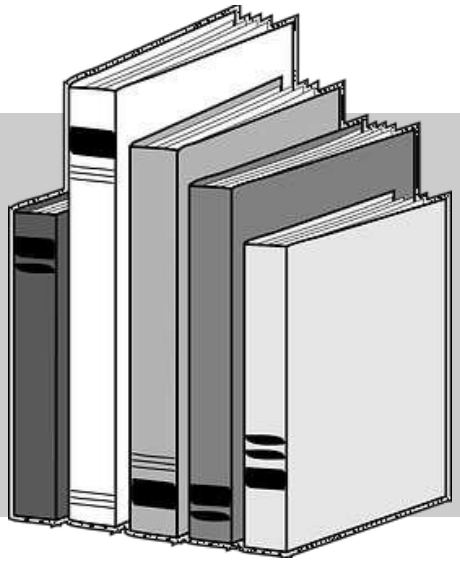


The **length** of the box plot indicates the spread of the data

Box Plot

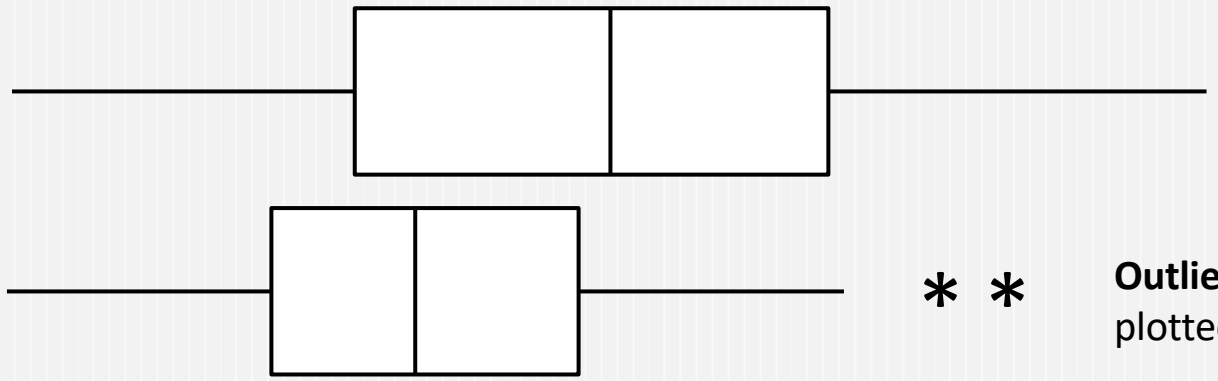
Box plots are widely used in statistics, process improvement, scientific research, economics, and in social and human sciences

Mainly used to **explore** data as well as to **present** the data in an easy and understandable manner.



Box Plot

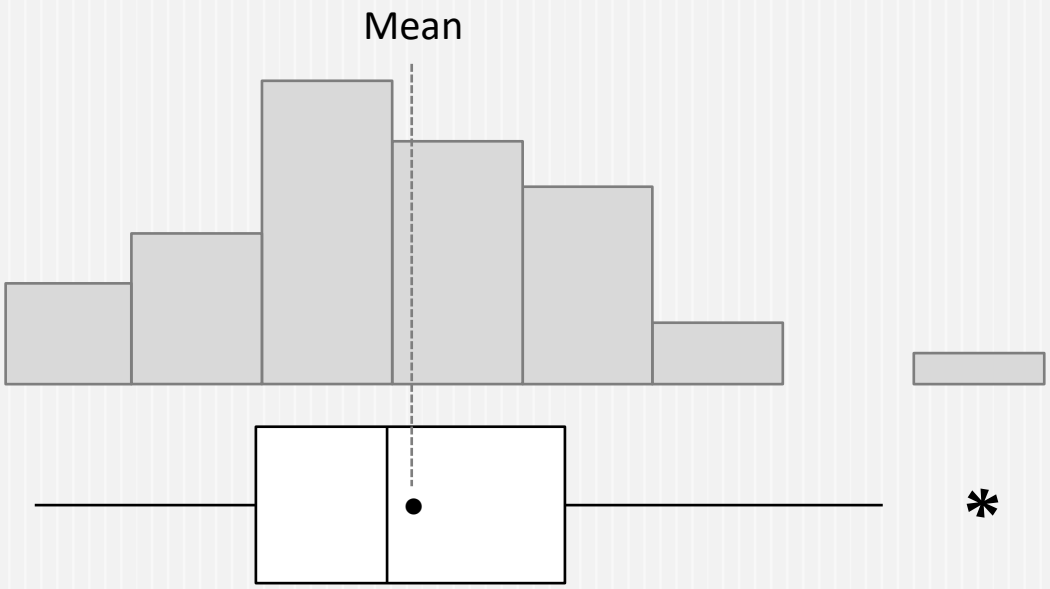
They provide a quick way for examining the **central tendency** and **variation** present in the data



A wider range boxplot indicates **more variability**

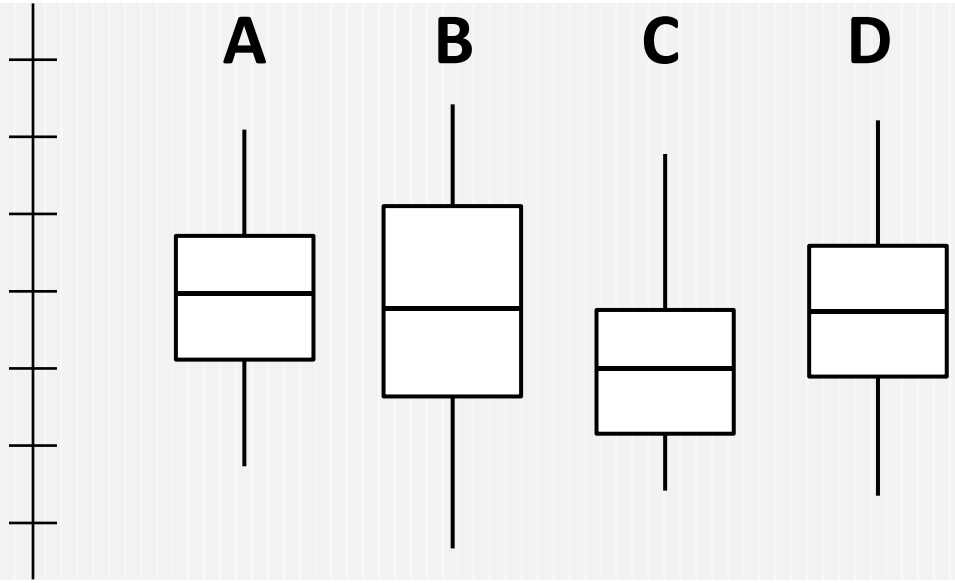
Box Plot

The same continuous data can be presented graphically using histograms and box plots



Box Plot

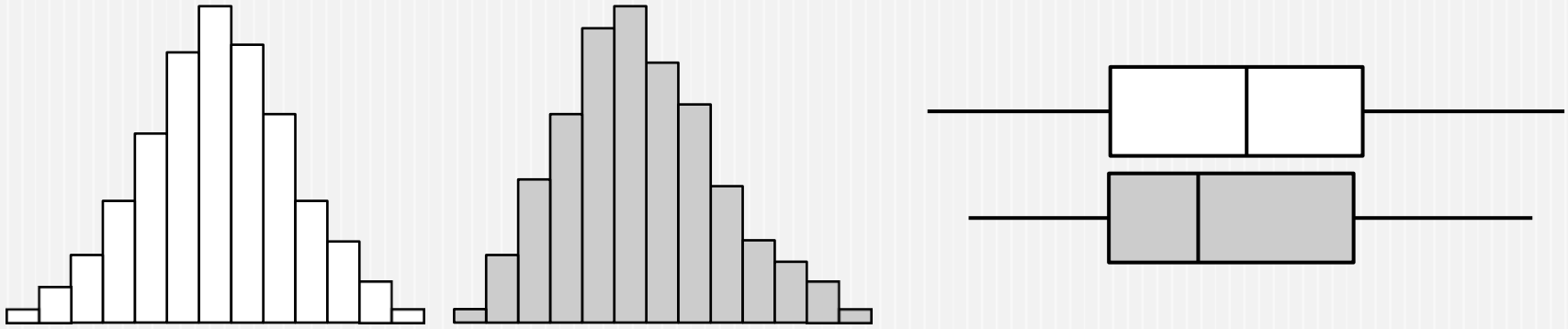
Useful when **comparing** between several data sets



Comparing central tendency and variability

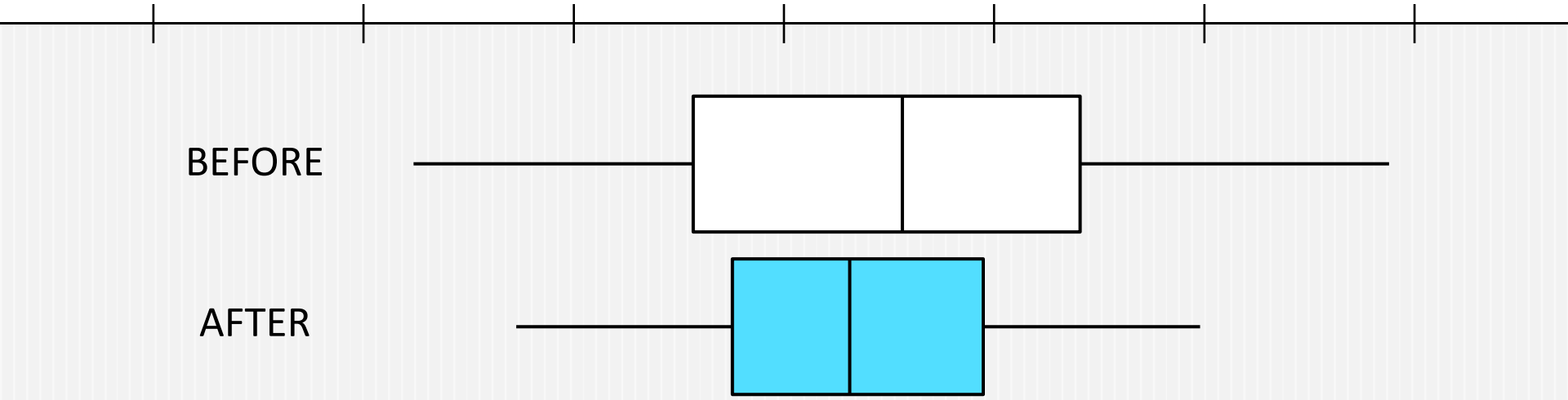
Box Plot

Less detailed than histograms, and take up less space which allows easy comparison of multiple data sets



Box Plot

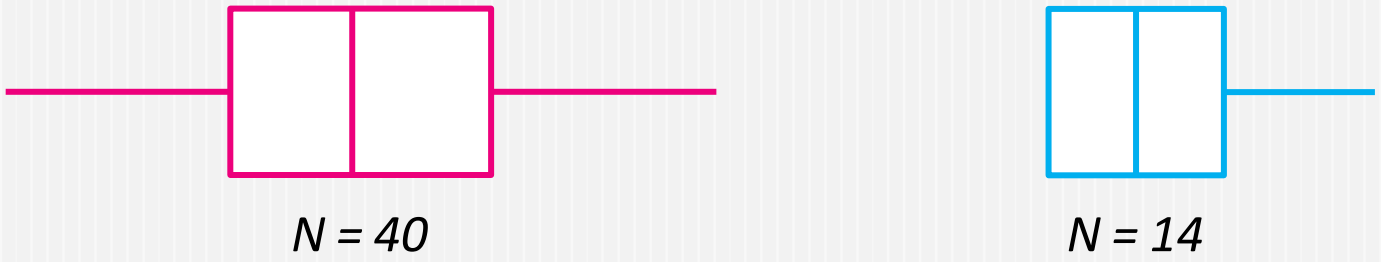
Used to check if there is a significant difference in the process after implementing **process improvement**



In terms of central tendency and variability

Box Plot

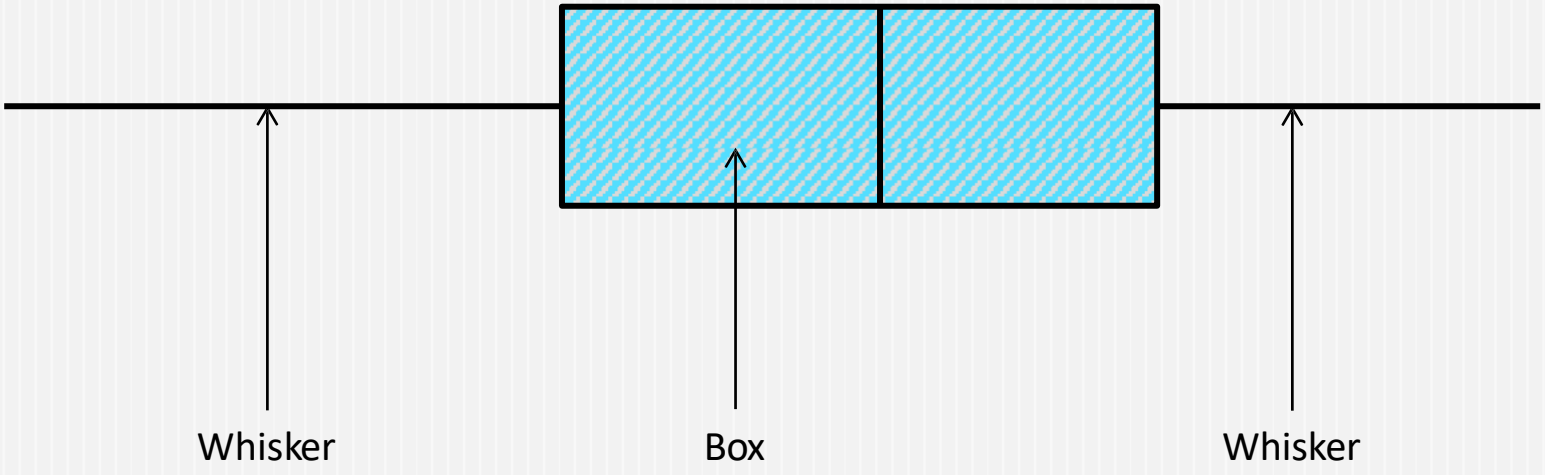
Like histograms, used for moderate to large amount of data



The **size** of the box plot can vary significantly if the data size is too small

Box Plot

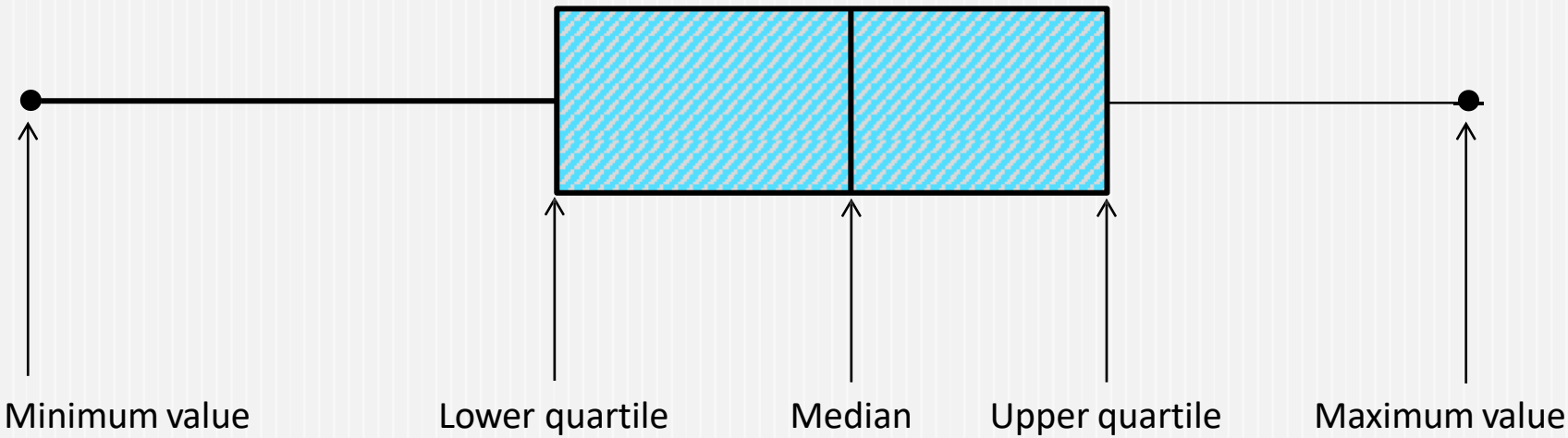
A box plot is made up of a **box** and two '**whiskers**'



The whiskers represent all data values, and the maximum length of a whisker is 1.5

Box Plot

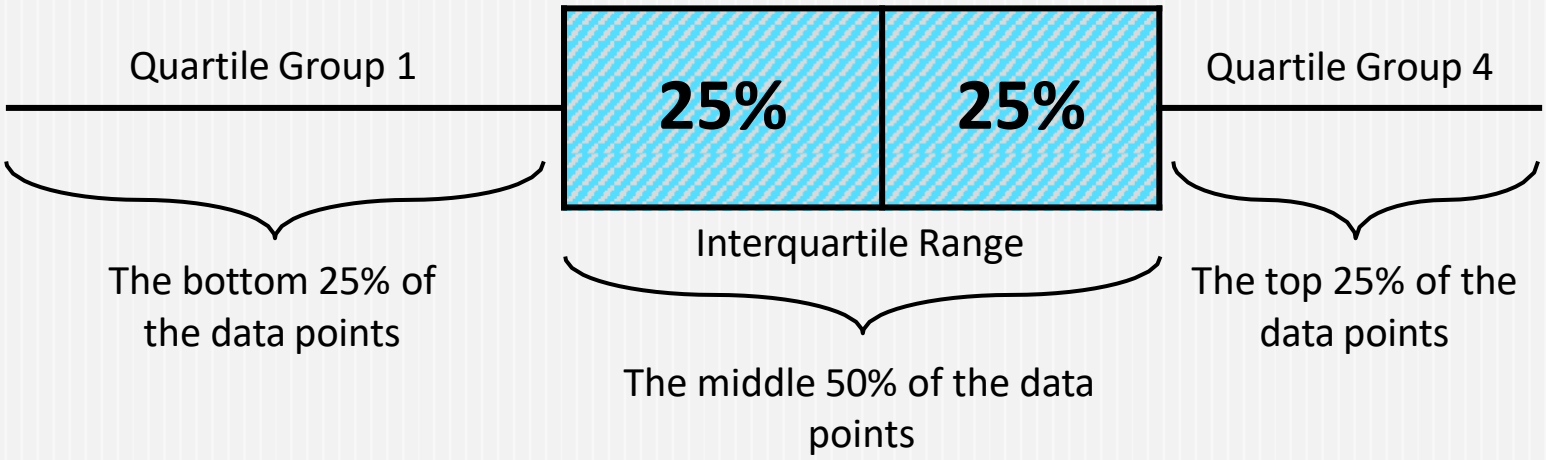
Box plots summarize **key statistics** from the data



Including the median, maximum and minimum values, as well as the lower and upper quartiles (Q1 and Q3)

Box Plot

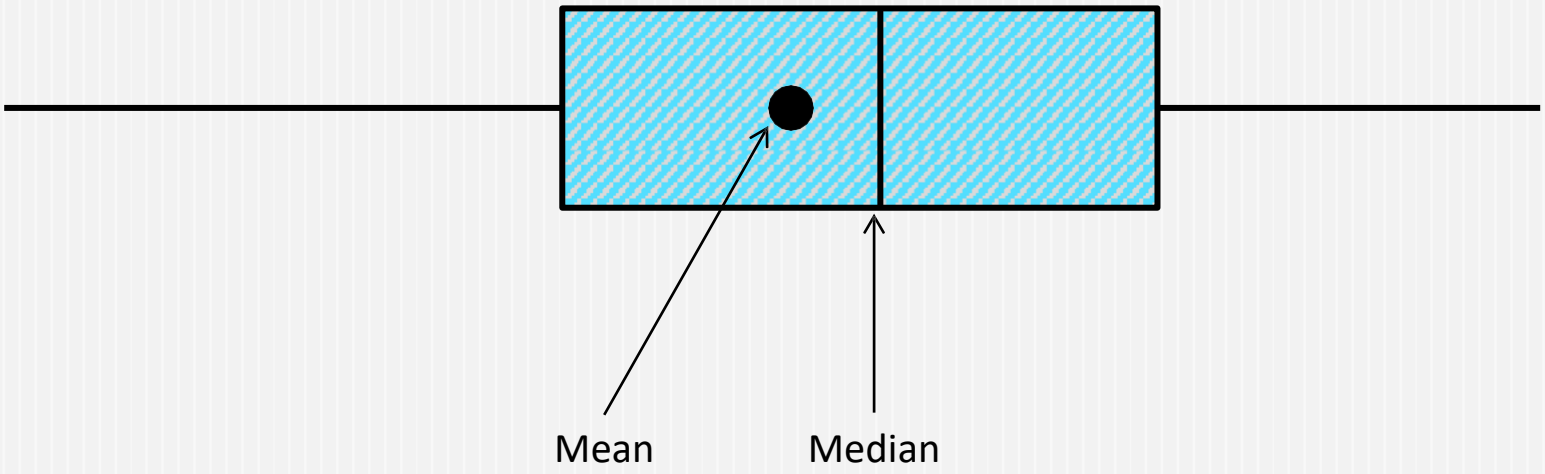
The data is plotted such as . . .



When the median line is not present in the box plot, it suggests that it coincides with one of the quartiles

Box Plot

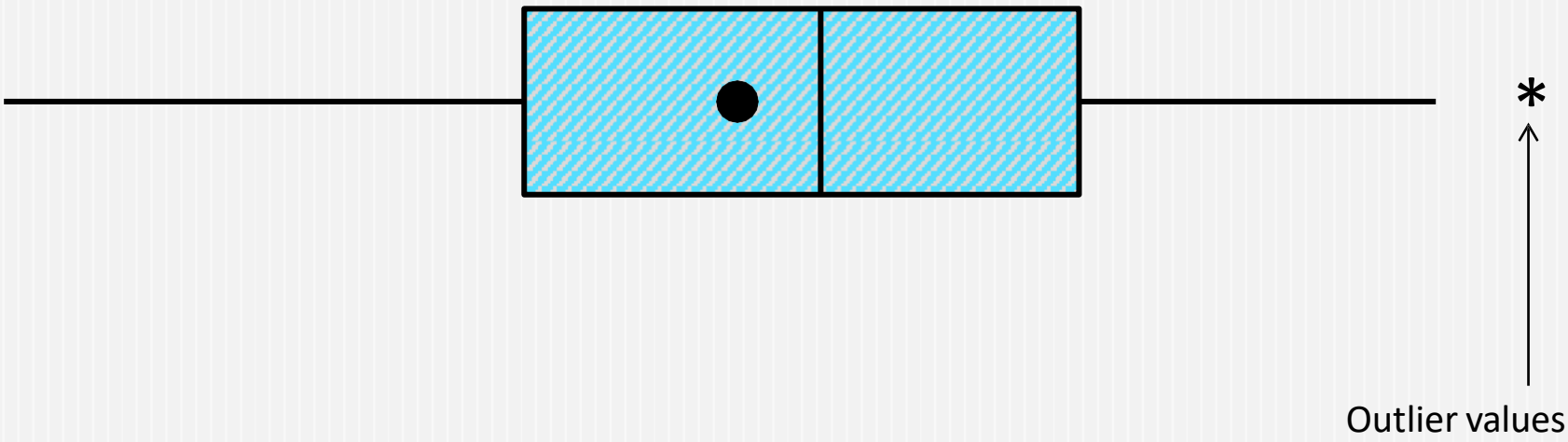
Sometimes they display the **mean** with a special character



Other character can indicate the mean such as a diamond, a plus, etc.

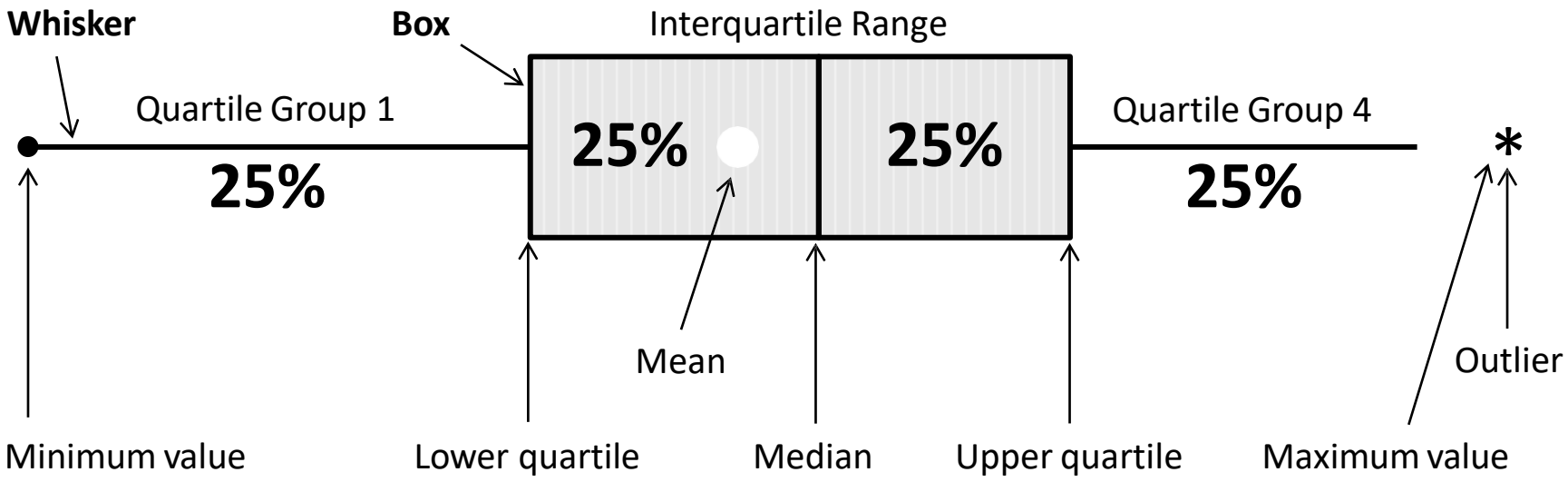
Box Plot

Any data beyond the whiskers are considered **outliers**



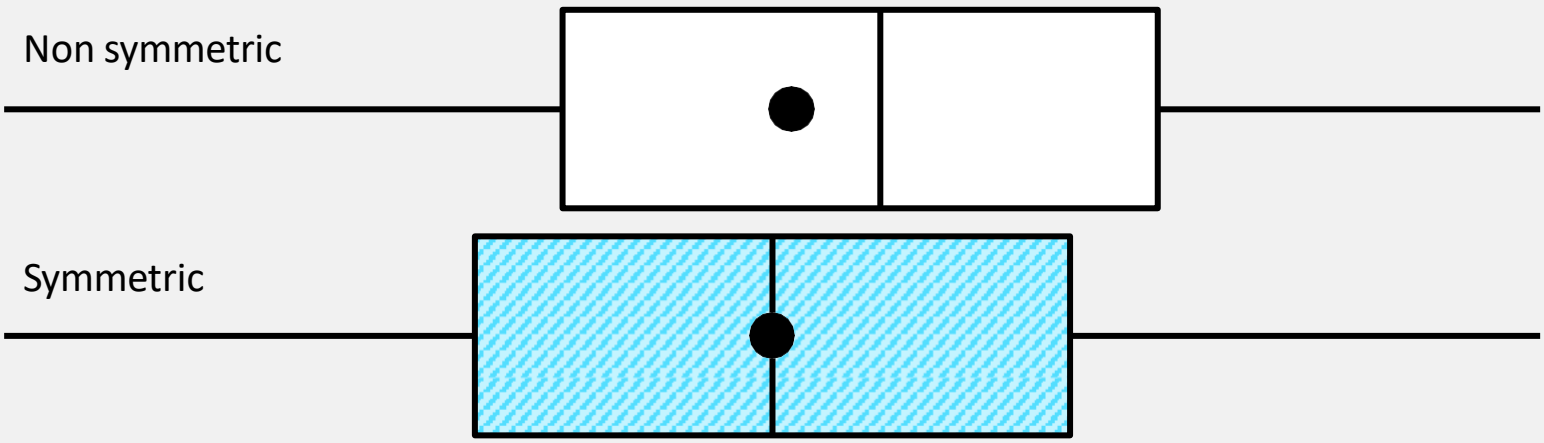
Outliers often reflect errors in data recording or data entry. If the values are real, you should investigate what was going on in the process at that time

Box Plot


Whisker
Box
Interquartile Range
Quartile Group 1
25%
25%
25%
Quartile Group 4
25%
*
Mean
Outlier
Minimum value
Lower quartile
Median
Upper quartile
Maximum value

Box Plot

Can tell whether the **distribution** is symmetrical or skewed

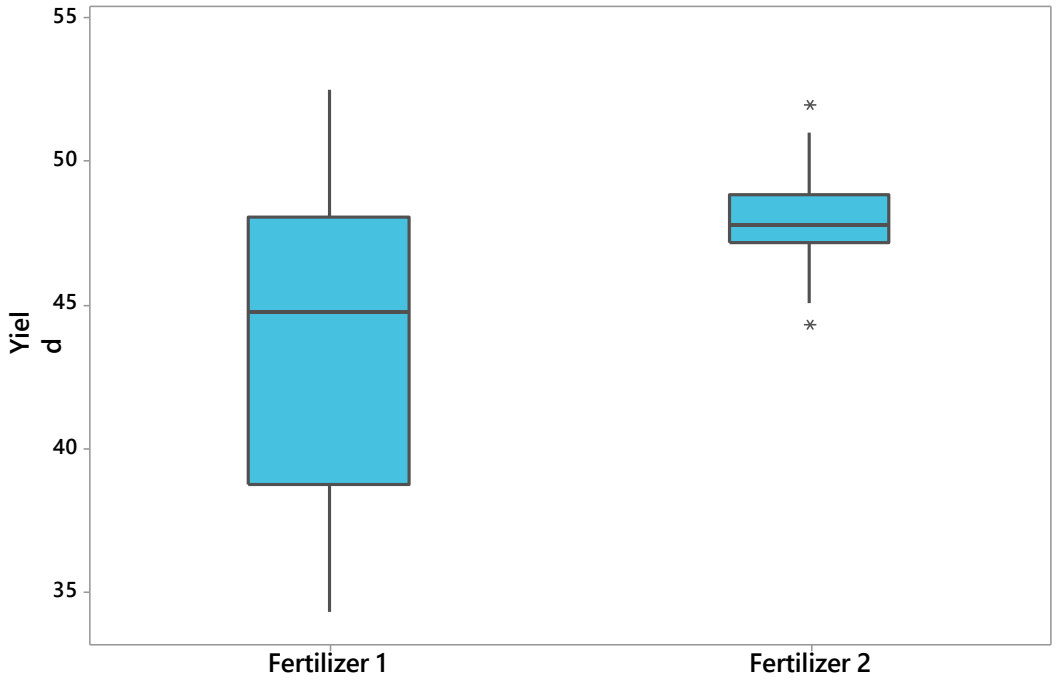


In a **symmetric distribution**, the mean and median are nearly the same, and the two whiskers has almost the same length

Box Plot

Example – Fertilizers

Data source: Minitab

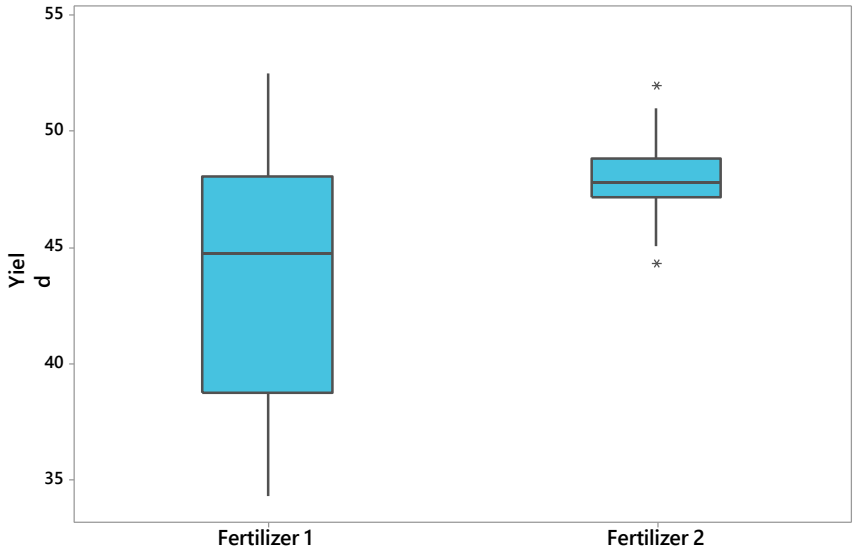


Fertilizer 2 appears to have a higher yield value than Fertilizer 1

Box Plot

Example – Fertilizers

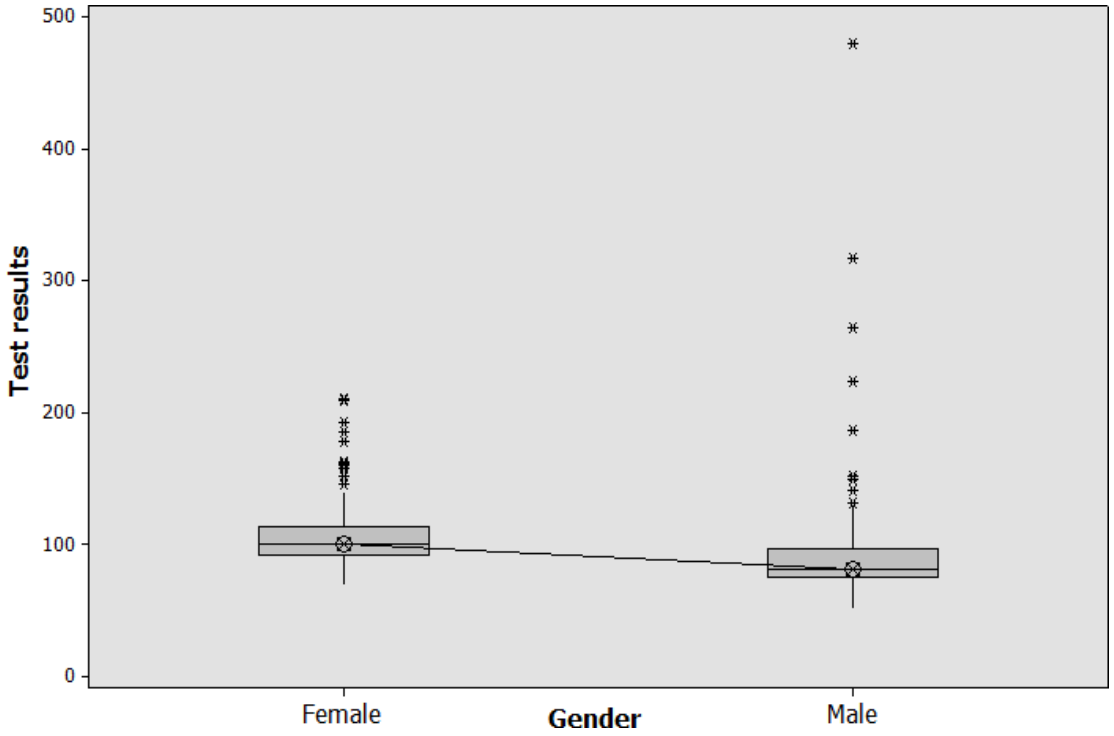
Data source: Minitab



What other comments would you make about the above boxplots?
Think about the variation as well as the presence of any unusual values

Box Plot

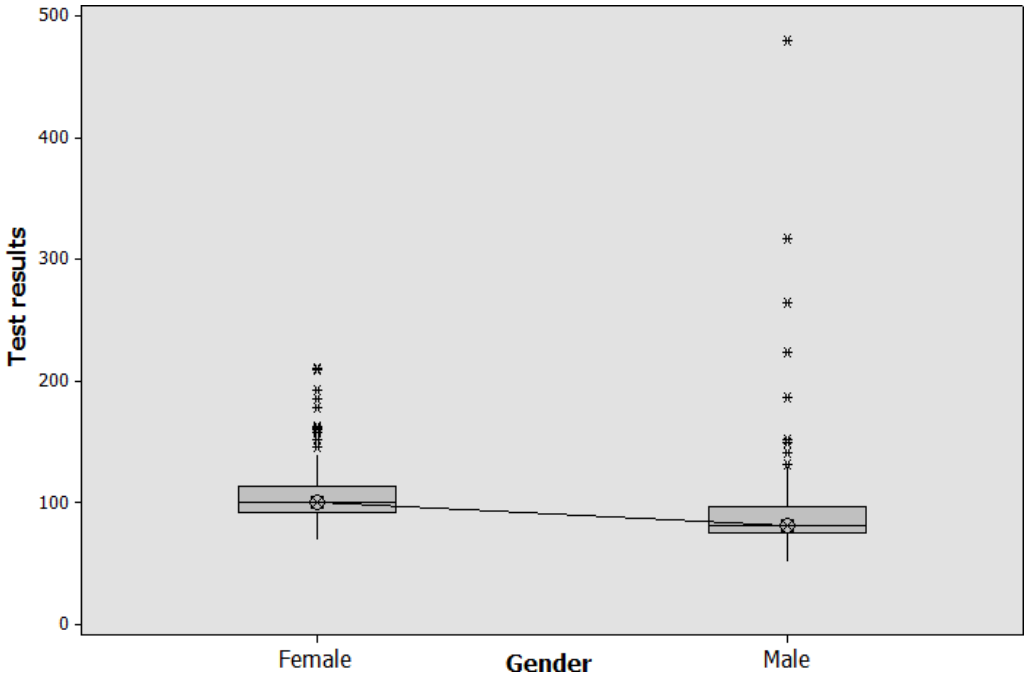
Example – Presence of Diabetes



It is evident that females have in general higher glucose levels than males

Box Plot

Example – Presence of Diabetes



ANOVA can be used here to test the **significance** of the difference between the two means

Given a set of data

5, 7, 12, 13, 18, 1, 14, 7, 15, 11, 6, 9, 13

1. List data in order from least to greatest

1, 5, 6, 7, 7, 9, 11, 12, 13, 13, 14, 15, 18

2. Next we need to find the 5 number summary.

    1. Minimum value
    2. Lower Quartile (Q1) – Median of the lower half of the data
    3. Median
    4. Upper Quartile (Q3) – Median of the upper half of the data
    5. Maximum value

1, 5, 6, 7, 7, 9, 11, 12, 13, 13, 14, 15, 18

Min – 1
Max – 18
Median – 11

Lower half - 1, 5, 6, 7, 7, 9 Q1 = 6.5

Upper half - 12, 13, 13, 14, 15, 18 Q3 = 13.5

If we had an even number of values in our data set, the median would be the mean of the two middle values. In that case we would cut the data set into two equal halves to find the quartiles.

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

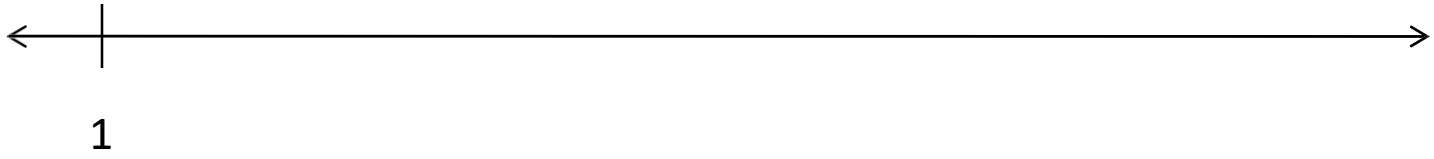
Median – 5. 5

Lower half - 1, 2, 3, 4, 5 Q1 = 3

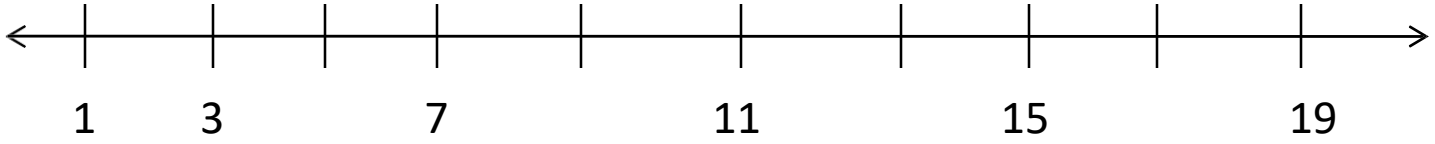
Upper half - 6, 7, 8, 9, 10 Q3 = 8

3. Now we are ready to make our box plot. Start with a number line. The first tick mark I generally make is at the minimum. To determine the scale I am going to use, I will generally find the range (Max – Min) and divide by 10, and then round to the nearest integer.

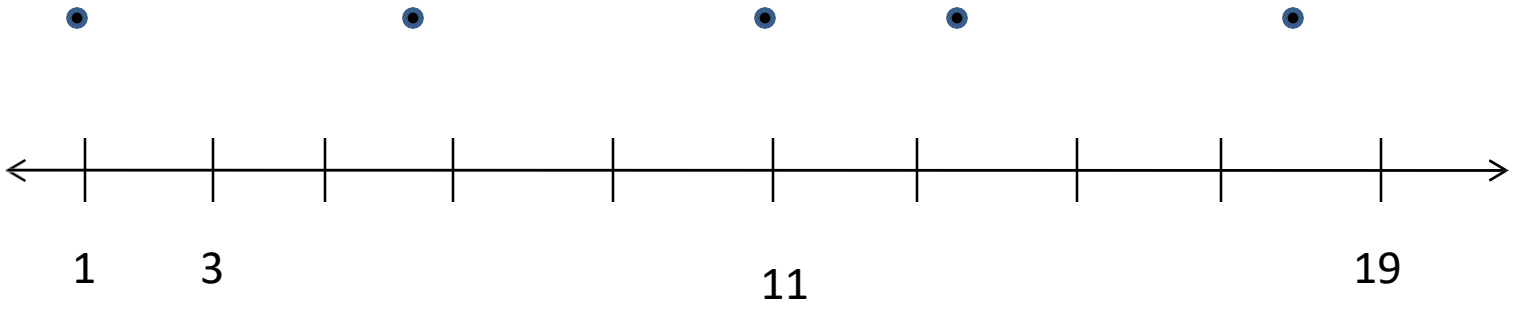
5 number summary - (1, 6.5, 11, 13.5, 18)



Range = 18 – 1 = 17 divide by 10 = 1.7, will use a scale of 2

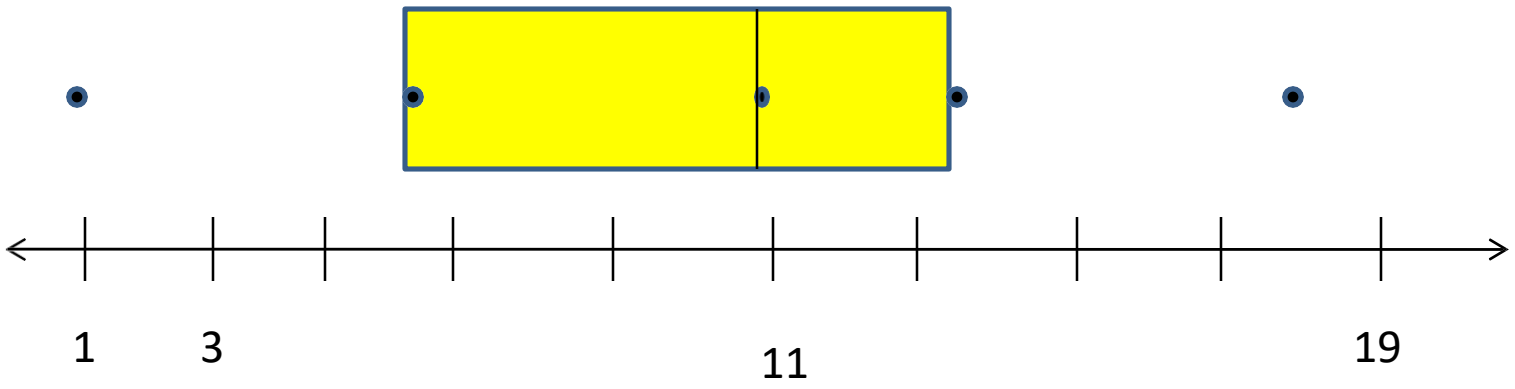


4. Now to construct the boxplot. Put a point above the number line for each value in the 5 number summary (1, 6.5, 11, 13.5, 18)



5. Draw vertical lines through Q1 and Q3 and then finish the box connecting them with two horizontal lines. Also draw a vertical line through the median. This is the box.

(1, 6.5, 11, 13.5, 18)



6. Now draw horizontal lines from the outer edge of the box to the minimum and maximum. These are your whiskers.

(1, 6.5, 11, 13.5, 18)