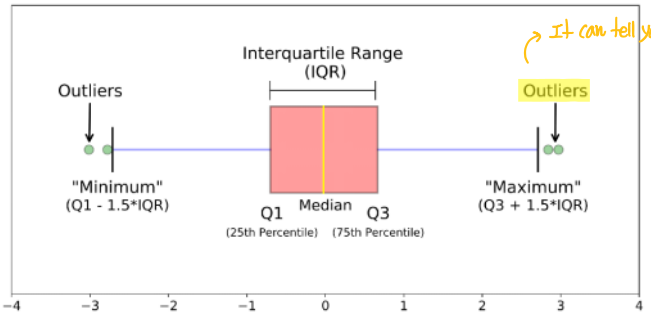
Box plot

Tuesday, August 17, 2021 09:04

For the glory of God

What is a box plot ?

- A box plot is a standardized way of displaying the distribution of data based on a five number ;
  - ↳ minimum, maximum, first quartile (Q1), third quartile (Q3), median



↗ It can tell you about your outliers.

- A box plot gives you a good indication of how the values in the data are spread out.

Elements of box plot

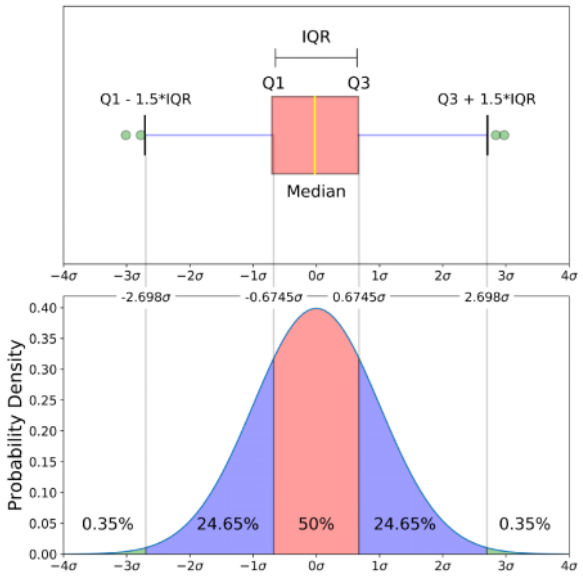
- Median
  - The middle value of the dataset
  - Note that median is more robust than mean (or average) especially if there are many outliers in the dataset.
- First quartile (Q_1 or 25th percentile)
  - The median of the lower half of the dataset
- Third quartile (Q_3 or 75th percentile)
  - The median of the upper half of the dataset
- Minimum
  - The lowest data point excluding outliers
  - It can be computed by : $Q_1 - 1.5 \times IQR$

↓ It stands for Interquartile Range. (25th to the 75th percentile)

- Maximum
  - The largest data point excluding outliers
  - It can be computed by : $Q_3 + 1.5 \times IQR$

Comparison between box plot and normal distribution

- The image below may help us get better understanding of a box plot.



Example : Finding the five-number summary

- Let's say that we have the following dataset :

  25, 28, 29, 29, 30, 34, 35, 35, 37, 38

- Step 1 : Order the data from smallest to largest

  ↳ The dataset is already in order.

- Step 2 : Find the median

  25, 28, 29, 29, 30, 34, 35, 35, 37, 38

  ↓

  $\frac{30 + 34}{2} = 32$

  * Note that the median is the mean of the middle two numbers if n is odd

- Step 3 : Find the quartiles

  a) The first quartile is the median of the data points of the left of the median (32)

  25, 28, 29, 29, 30

  $Q_1 = 29$

  b) The third quartile is the median of the data points of the right of the median (32)

  34, 35, 35, 37, 38

  $Q_3 = 35$

- Step 4 : Find the minimum / maximum values

a) minimum : 25, 28, 29, 29, 30

minimum

b) Maximum : 34, 35, 35, 37, 38

maximum

- Step 5 : Compute IQR and define outliers

↳ outliers can be defined by computing IQR if there is any outlier in the dataset.