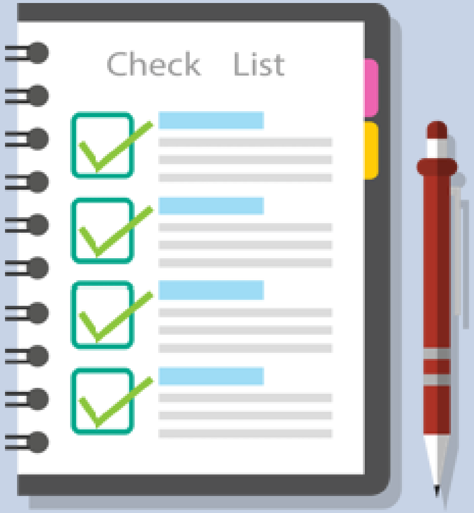
INTRO- DATA SCIENCE

Shah Ayub Quadri

Ayub.quadri89@gmail.com

Index



Basic Maths - Statistics

Statistics

Basic Terminologies

- Descriptive Stats
- Inferential Stats

Central Tendencies

- Mean
- Median
- Mode

Measure of Spread

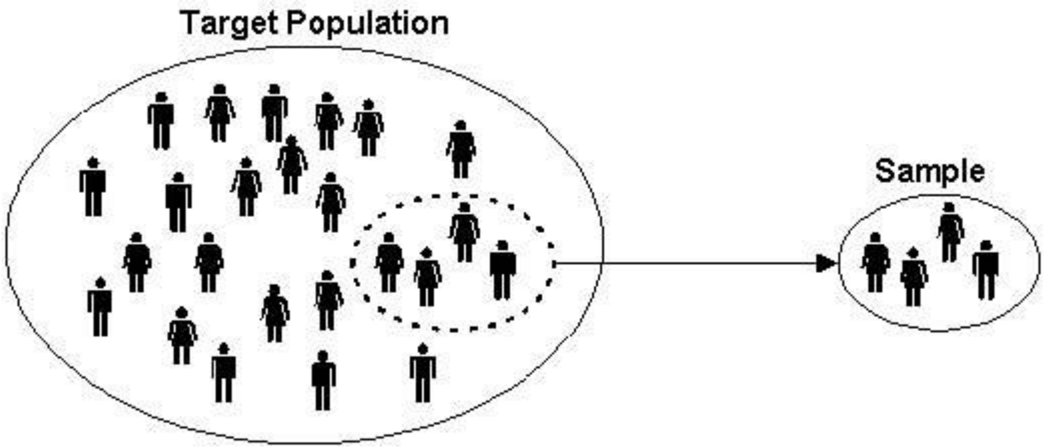
- Variance
- Standard Deviation
- Range
- Quartiles
- Box or Whisker Plot
- Standard Score (z-score)

Statistics

- It is the art of Learning from data.
- Statistics provides a way of organizing the data to extract information based on the data points rather than the intuition or personal experience

Eg: Principal of a School claims that at an average his school will score more than 70% in examination

- Evaluation: To Validate the statement, lets collect sample from a class say 9th standard
- Condition: If the Average score of the class is greater than 70%
- Conclusion: Claim of principal can be accepted
- Conclusion is made based on data points rather than the intuition or assumptions.



Statistics Terminology

Population:

- Information about the total collection of element or data
- Eg: Total strength of students in entire School

Sample:

- The subset of population is referred as sample
- Eg: Say Class IX

Note: population is often too large for us to examine each of its elements. In such cases, we try to learn about the population by choosing & examining a subgroup of its elements.

Parameter:

Descriptive Measure of population

Greek – population parameter

- Mean (μ)
- Variance (σ^2)
- Standard Deviation (σ)

Statistic:

Descriptive measure of Sample

Roman – Sample statistics

- Mean (σ)
- Variance (S^2)
- Standard Deviation (S)

Descriptive Stats:

- Data Gathered about a group to reach conclusion about the same group is called descriptive stats

Eg: If one wants to understand the distribution or spread of the data, Descriptive stats allows us to do this.

- Central Tendency (mean, median, mode)
- Measure of Spread (variance, standard deviation)

Inferential Statistics:

- Data gathered from a sample & the statistics generated to reach conclusion about the population also know as Inferential stats or inductive statistics

Eg: when we don’t have access to the whole population and interested in investigating the data, inferential statistics helps out to do the inference of the population using the sample.

With the help of

- Estimation of parameters (sample mean, variance, std deviation)
- Hypothesis testing (z-score, t-test, chi-square test)

Variables & Data:

- Data represents the complete row of a data set
- Variable are the columns of a given data set

Variables:

- Dependent variable: The outcome variable(Y) which is dependent on the independent variables (X)
- Independent variable: Variables(X) using which outcome(Y) is derived

Qualitative data:

- Nominal data: Categorical data

  Eg: (cat, dog, parrot, fish) – pets
- Ordinal data: Categorical value with some order or rank

  Eg: (Low, medium, High) – Water Level
- Dichotomous: variables with only two levels (True or False) (Yes or No) (Male or Female)

Quantitative Data:

- Interval: Meaningful difference, but no zero point

  Eg: The interval of 20C to 30C. Celsius & Fahrenheit cant b Ratio as at 0C
- Ratio: Meaningful difference, with a natural starting point.

  Eg: Height, Weight, Temp(kelvin)

Central Tendency

- The central position within the set of data is know as central tendency or measure of central tendency.
- Mean, Median Mode are the valid measure of central tendency.

Mean:

- The ratio of sum of all values in the data set to that of no. of element in the data set.
- It can be used with both discrete & continues data

Sample Mean($\bar{x}$)

$$\bar{x} = \sum \frac{x}{n} = (\frac{x_1+x_2+x_3+ \text{......}+x_n}{n})$$

Population Mean(μ)

$$\mu = \sum \frac{x}{N} = (\frac{x_1+x_2+x_3+ \text{......}+x_N}{N})$$

For Examples: https://www.mathsisfun.com/data/central-measures.html

Advantages:

- It includes every value in your data set as part of the calculation
- Only measure of central tendency where the sum of the deviation of each value from the mean is always zero

Disadvantage:

- Effected by outliers or susceptive to the influence of outliers

Median:

- It is the middle value of a data set that has been arranged in order of its magnitude
- Median is less effected by outliers & skewed data
- Based on the data set size
  - If ‘n’ is odd – median = value at the position $(\frac{n+1}{2})$
  - If ‘n’ is even – medina = Avg(value$(\frac{n}{2})$ + value($(\frac{n}{2})$ +1))

Mode:

- Mode is a value that occurs with the greatest frequency in a given data set
- Mode is generally used for categorical values

Advantage:

- Provides the max frequency value

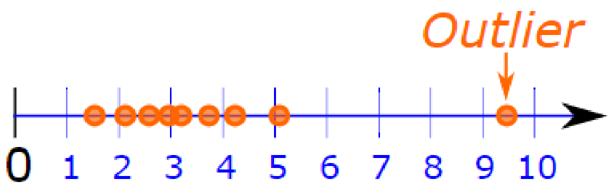
Disadvantage:

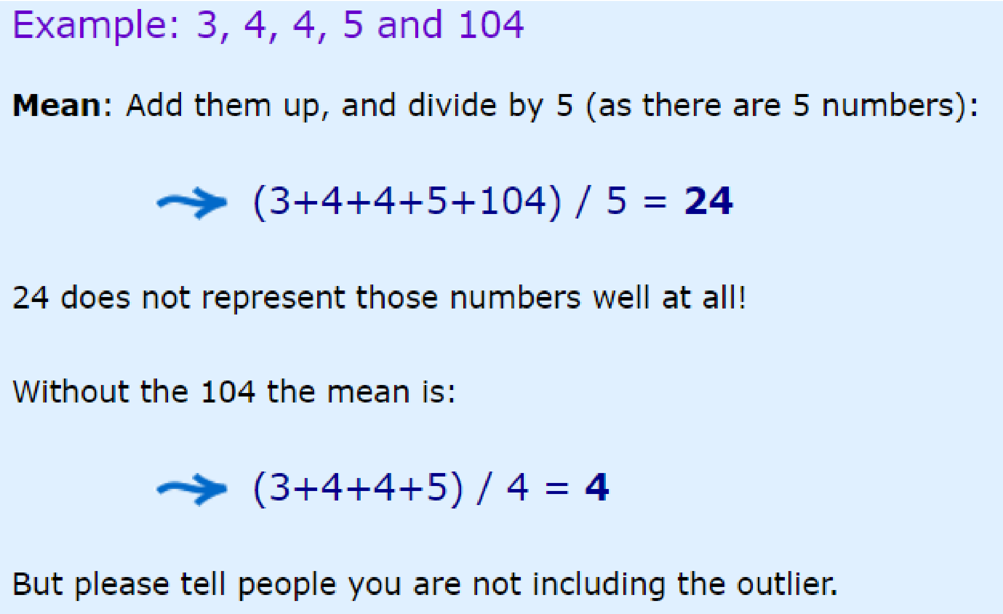
- Mode is not unique when we have more that one highest frequency value ‘bimodal’ or ‘multimodal’
- Problem occurs when most frequent data is far away from the rest of the data.

For Examples: https://www.mathsisfun.com/data/central-measures.html

Outliers:

- The values that lie-outside the range of values in a data set.
- Mean effected by outliers.





Example: 3, 4, 4, 5 and 104

Mean: Add them up, and divide by 5 (as there are 5 numbers):

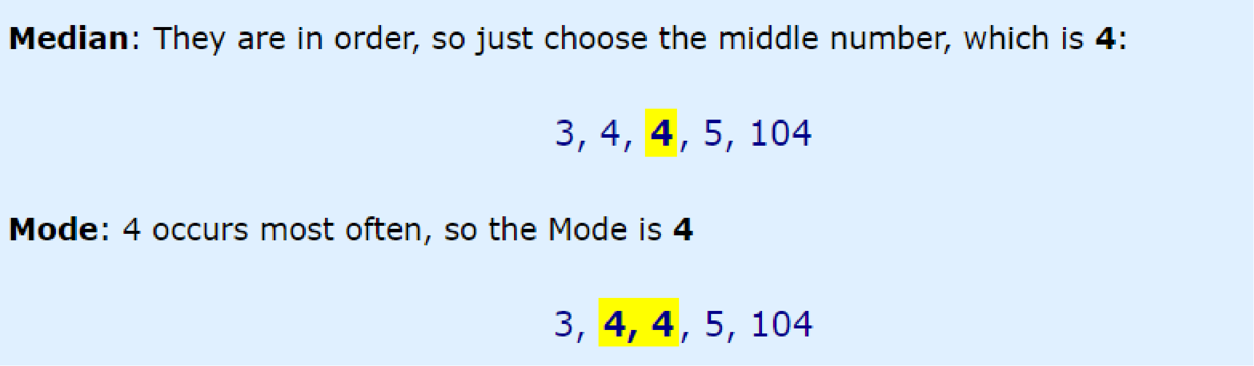
→ (3+4+4+5+104) / 5 = **24**

24 does not represent those numbers well at all!

Without the 104 the mean is:

→ (3+4+4+5) / 4 = **4**

But please tell people you are not including the outlier.



Median: They are in order, so just choose the middle number, which is **4**:

3, 4, **4**, 5, 104

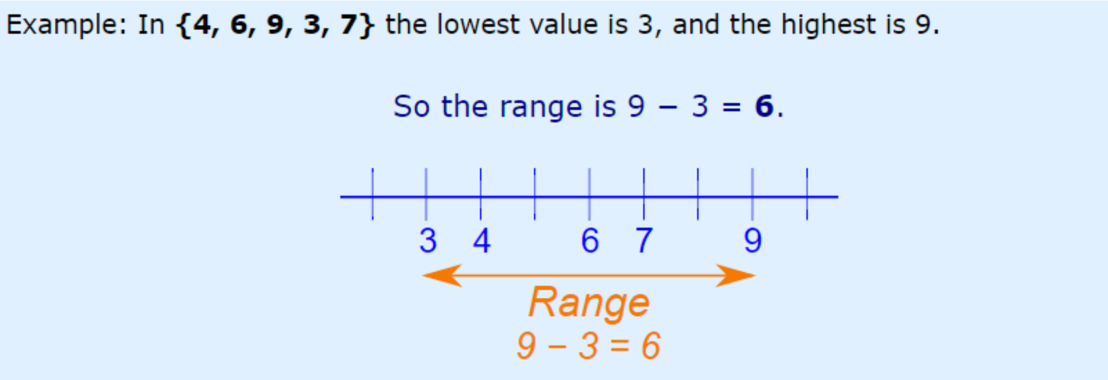
Mode: 4 occurs most often, so the Mode is **4**

3, **4, 4**, 5, 104

Note: Prefer Median over mean or mode when data is skewed (consists of outliers)

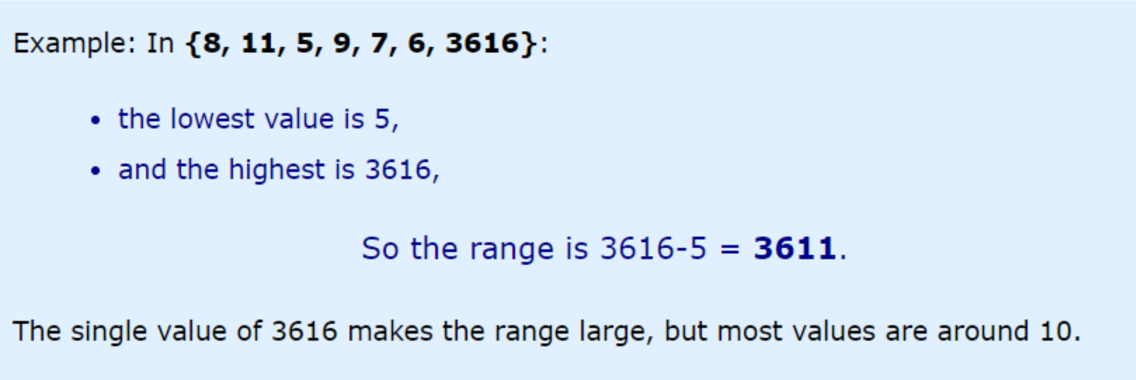
Measure of spread

Range: Difference between highest & lowest value.



Example: In **{4, 6, 9, 3, 7}** the lowest value is 3, and the highest is 9.

So the range is 9 − 3 = **6**.



Example: In **{8, 11, 5, 9, 7, 6, 3616}**:

- the lowest value is 5,
- and the highest is 3616,

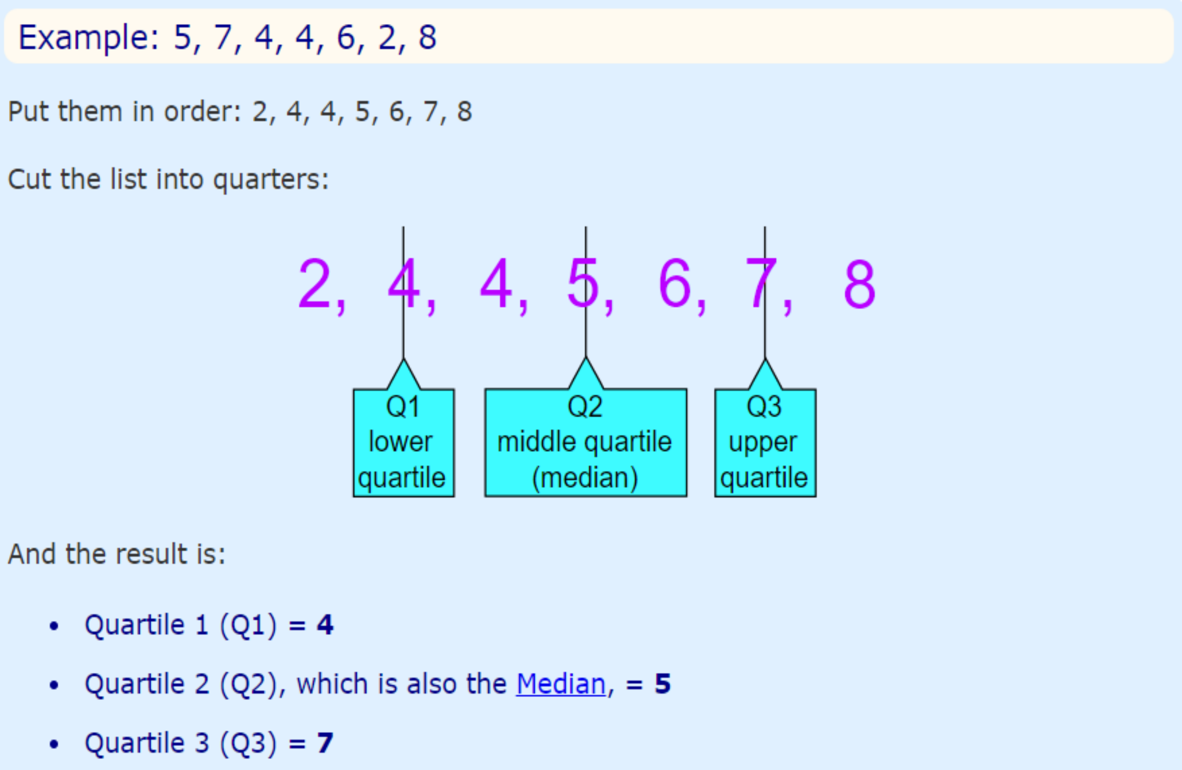
So the range is 3616-5 = **3611**.

The single value of 3616 makes the range large, but most values are around 10.

- Range can be misleading with outliers, thus use *standard deviation* or *Inter quartile range*

Quartiles:

- Quartiles are the values that divide a list of numbers into four quarters
- Scientifically quartile excludes outliers



Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

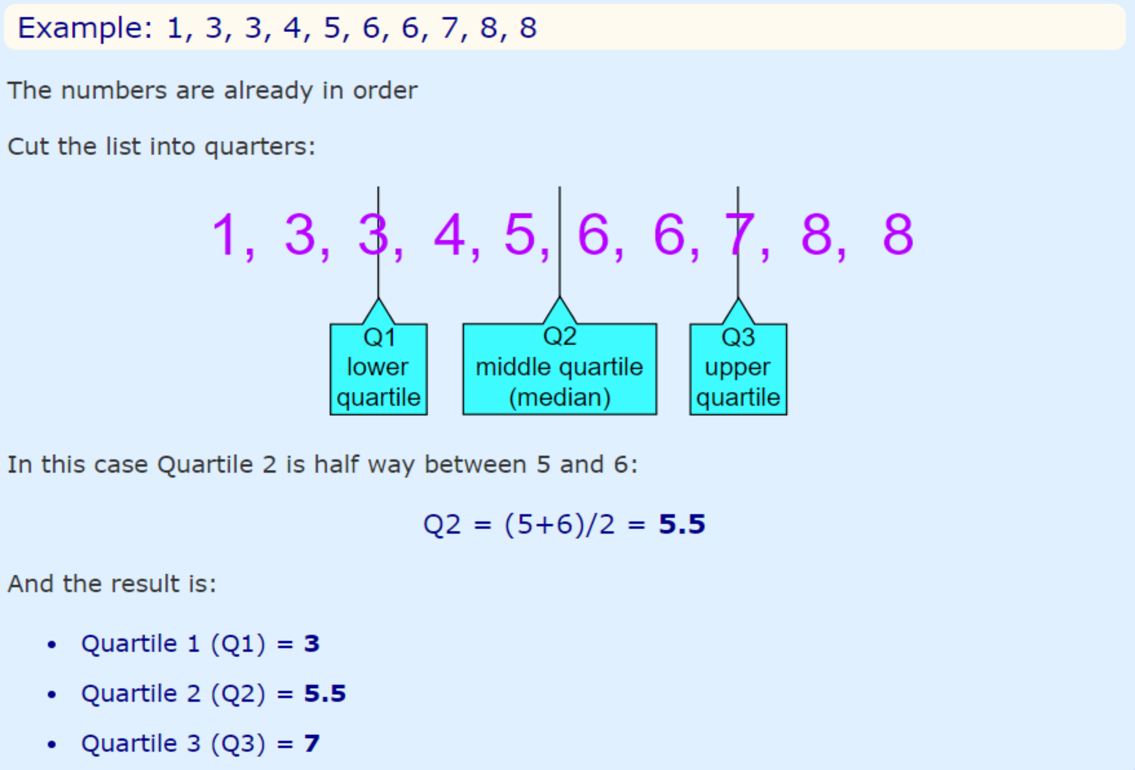
Cut the list into quarters:

2, 4, 4, 5, 6, 7, 8

And the result is:

- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **7**

Fig: odd no easy to get 4 Quartiles



Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:

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

In this case Quartile 2 is half way between 5 and 6:

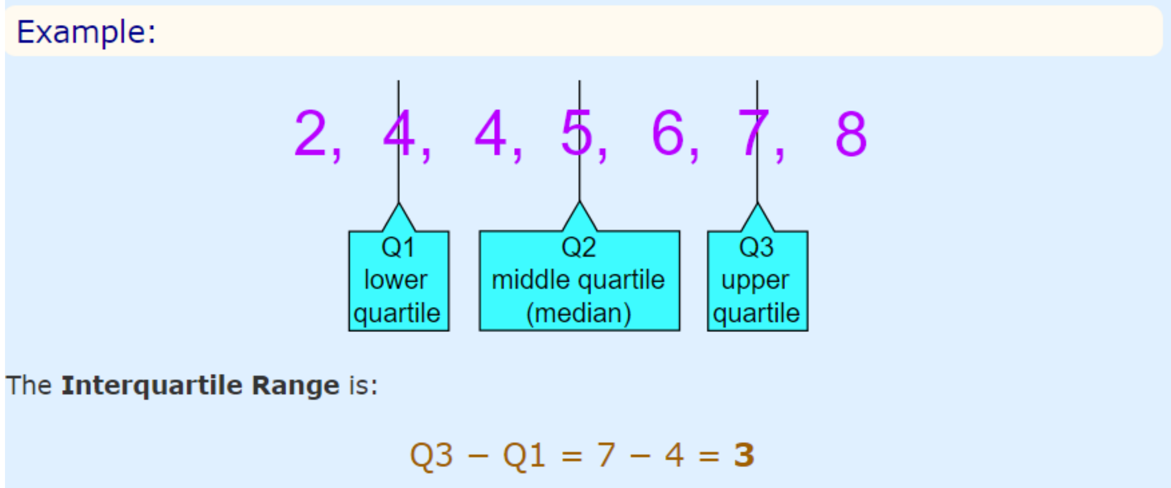
Q2 = (5+6)/2 = **5.5**

And the result is:

- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

Fig: even no get 4 Quartiles

Interquartile Range: the range between Q1 & Q3

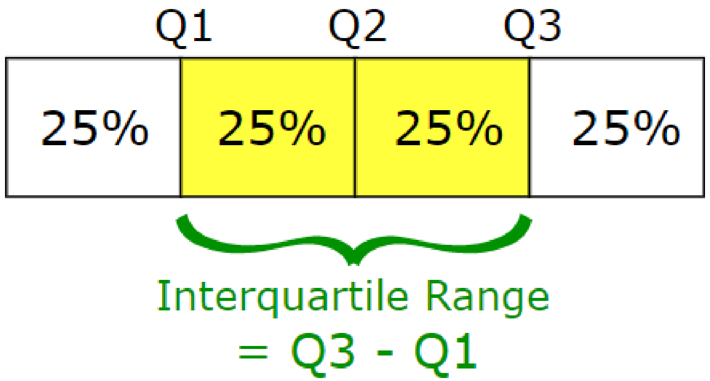


Example:

2, 4, 4, 5, 6, 7, 8

The **Interquartile Range** is:

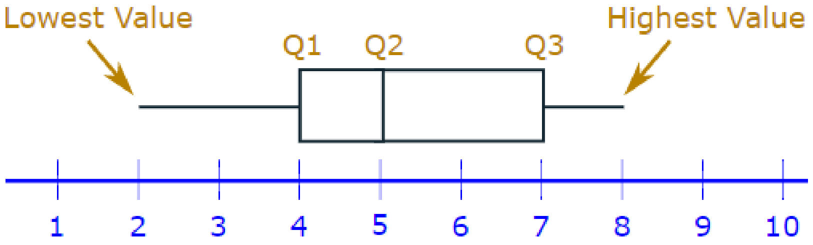
Q3 − Q1 = 7 − 4 = **3**

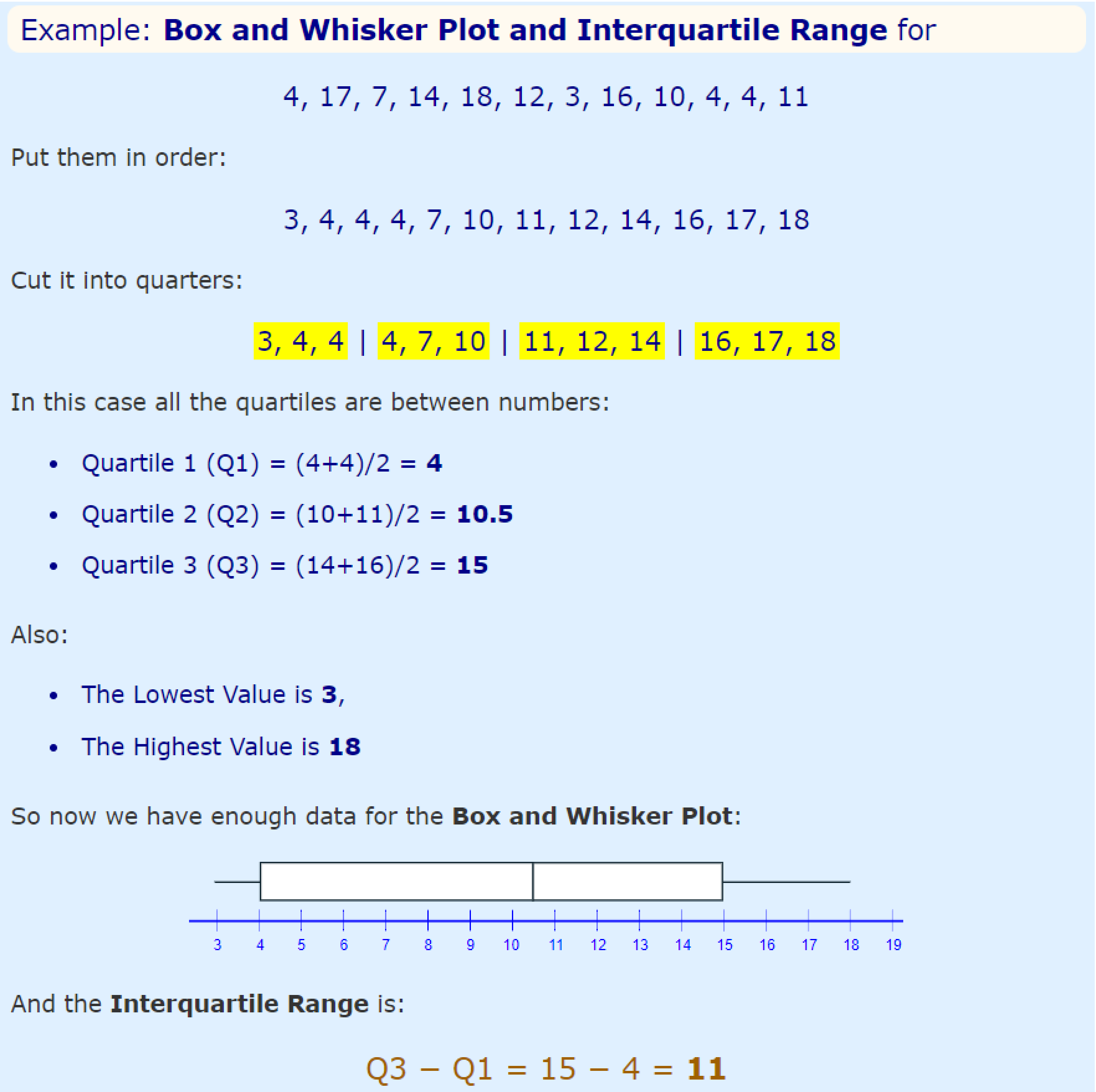


Box or Whisker Plot:

It represents the spread of the data ranging from lowest value to highest values, along with the Interquartile range

Outliers can be easily detected in Box plot.





Example: **Box and Whisker Plot and Interquartile Range** for

4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11

Put them in order:

3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

Cut it into quarters:

3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

In this case all the quartiles are between numbers:

- Quartile 1 (Q1) = (4+4)/2 = **4**
- Quartile 2 (Q2) = (10+11)/2 = **10.5**
- Quartile 3 (Q3) = (14+16)/2 = **15**

Also:

- The Lowest Value is **3**,
- The Highest Value is **18**

So now we have enough data for the **Box and Whisker Plot**:

And the **Interquartile Range** is:

Q3 − Q1 = 15 − 4 = **11**

Variance:

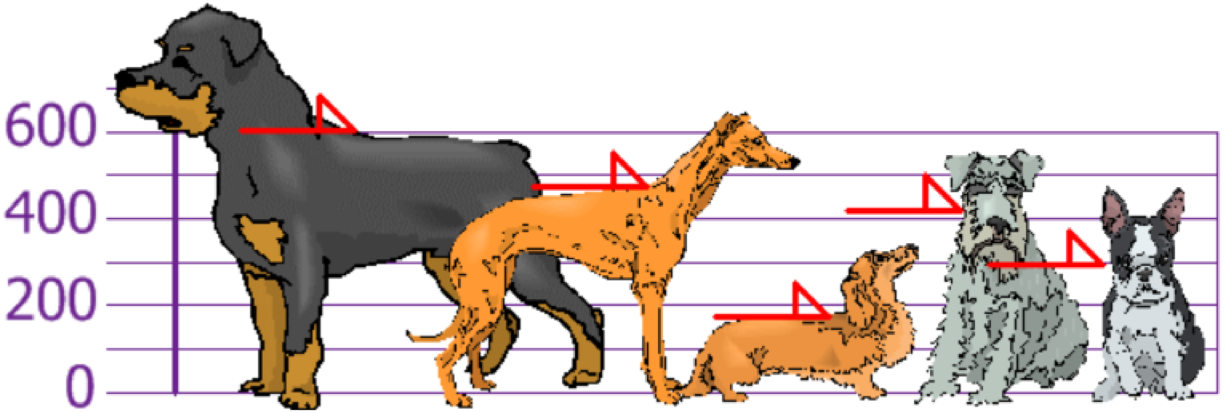
- Defined as the sum of squared distance of each term in the distribution from the mean(), divided by the number of terms in the distribution

$$\sigma^2 = \sum \frac{(X - \mu)^2}{N}$$

- It measures how far away is the given value from the mean.

Example

You and your friends have just measured the heights of your dogs (in millimeters):



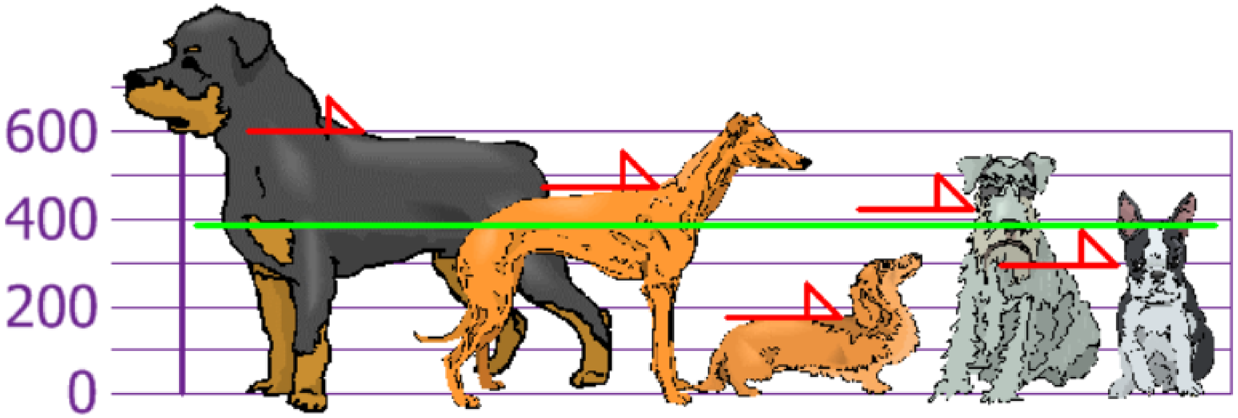
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

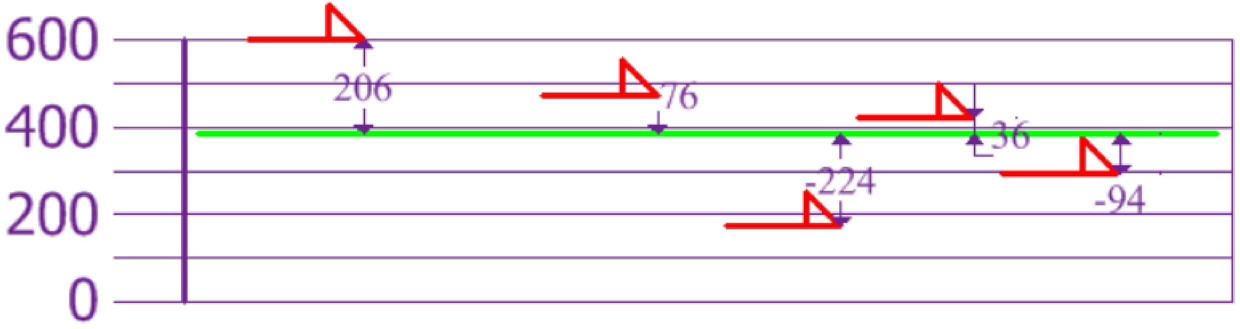
Step1: Find the mean

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



Step2: Find the variance (how far away from mean)



To calculate the Variance, take each difference, square it, and then average the result:

$$\text{Variance: } \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42,436 + 5,776 + 50,176 + 1,296 + 8,836}{5}$$

$$= \frac{108,520}{5} = 21,704$$

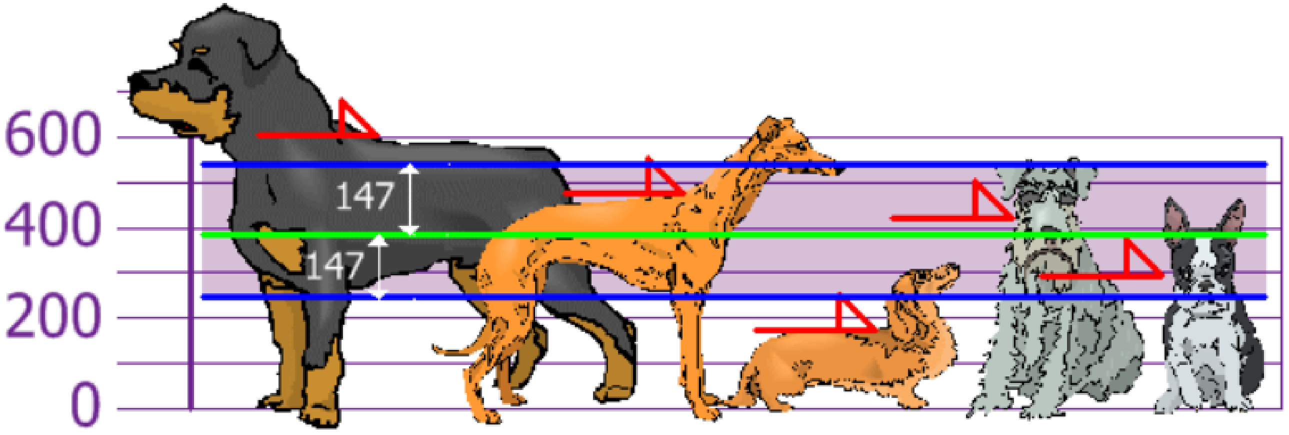
So the Variance is **21,704**

Step3: Find the Standard Deviation

Standard Deviation

$$\sigma = \sqrt{21{,}704}$$
$$= 147.32...$$
$$= 147 \text{ (to the nearest mm)}$$

And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short

Resources

Descriptive stats & inferential stats:

- https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

Types of variables:

- https://statistics.laerd.com/statistical-guides/types-of-variable.php

Variance

- https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/measuring-spread-quantitative/v/variance-of-a-population

Standard Deviation

- https://www.mathsisfun.com/data/standard-deviation.html
- http://www.mathsisfun.com/data/standard-deviation.html#WhySquare