Business Statistics

Agenda

01 Descriptive Statistics

- Measure of central tendency
- Measure Of Dispersion

02 Probability

- Probability Distributions

03 Inferential Statistics

- Type of sampling Technique
- Central limit theorem
- Confidence interval

04 Hypothesis testing

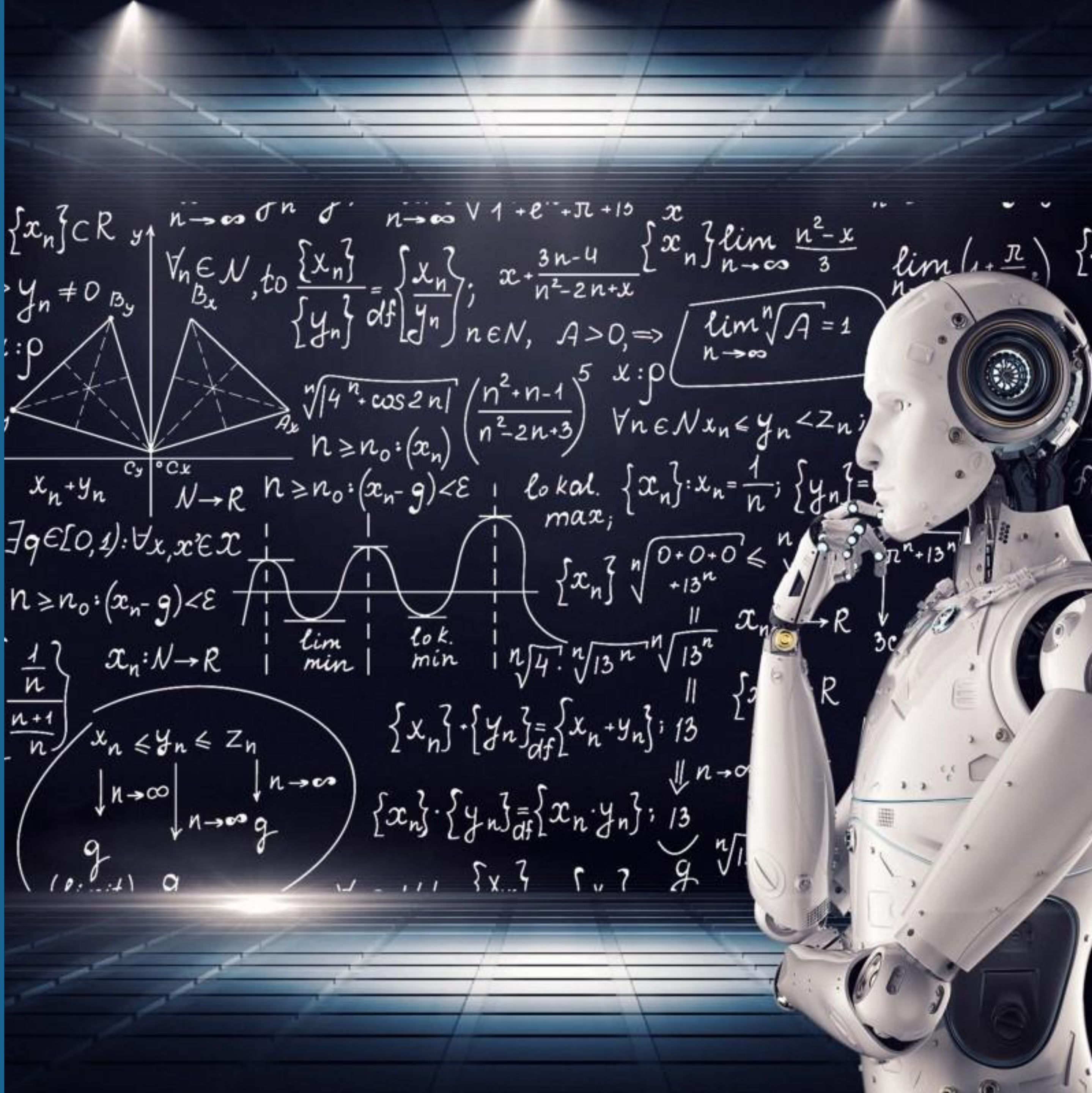
- Significance level
- P-value
- Statistical tests

Statistics

1. **Statistics** is concerned with the describing, interpretation & analyzing of the data
2. Statistics:
    - ○ Descriptive Statistics
    - ○ Inferential Statistics

3. It uses **analytical methods** which provide the math to model & predict variation
4. It uses **graphical methods** to help making numbers visible for communication purposes.

Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data.



$\{x_n\} \subset R$
$\forall n \in N$
$n \in N,\ A > 0 \Rightarrow$
$\lim_{n \to \infty} \sqrt[n]{A} = 1$
$x_n : N \to R$
$N \to R$
$n \geq n_0 : (x_n - g) < \varepsilon$
lim min
lok. min
lokal. max;
$x_n \leq y_n \leq z_n$
$\{x_n\} \cdot \{y_n\} \overset{}{=}_{df} \{x_n \cdot y_n\}$

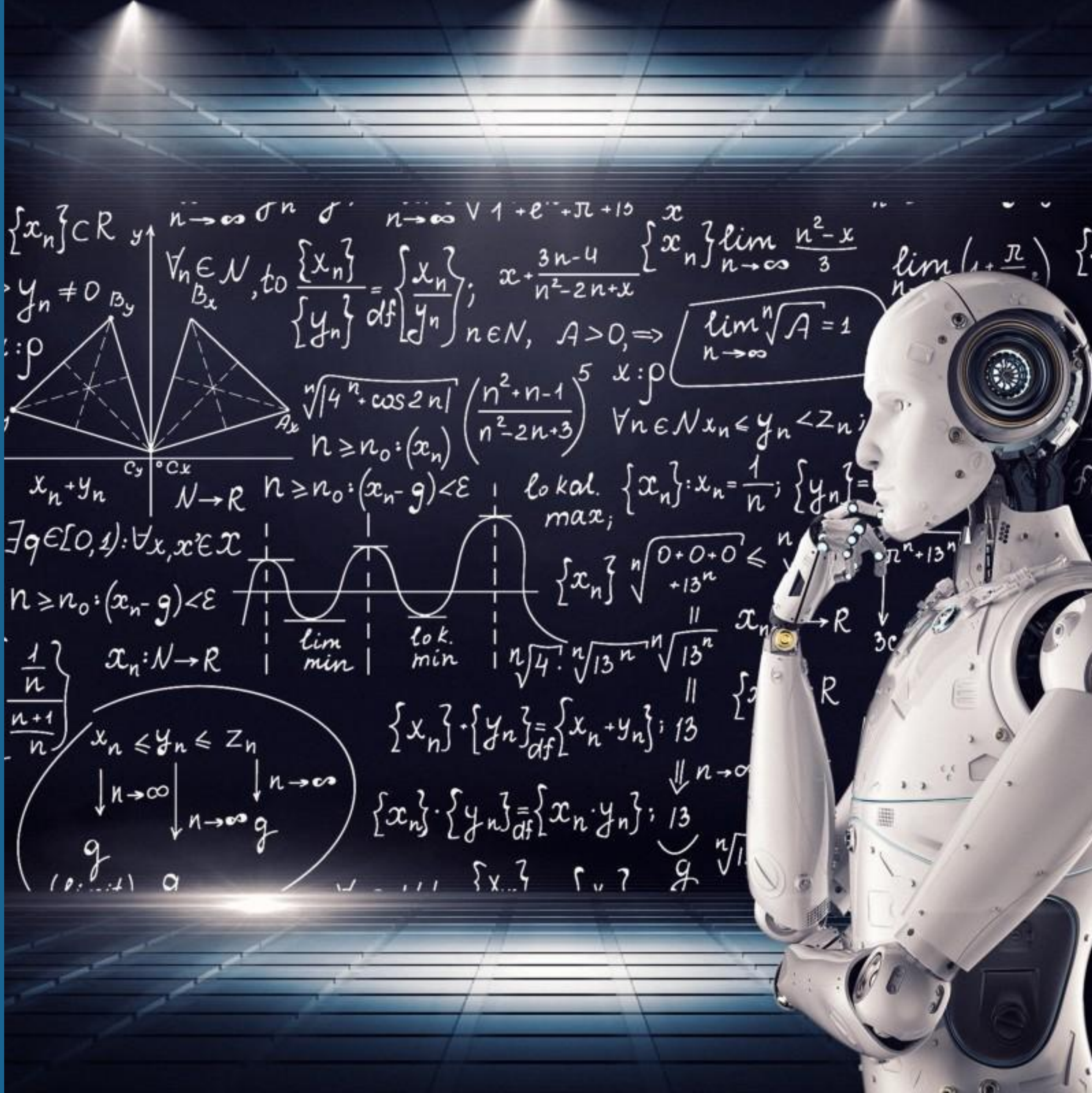
Descriptive Statistics

Descriptive Statistics

Descriptive Statistics describe/summarize the data and provide initial finding for the data. It give us basic understanding for the data. Descriptive statistics doesn’t give us any conclusion about the data but give the brief understanding about the data. It will give us mean/average, median, Mode ,Standard deviation etc about the data.

Types of Descriptive Statistics:-

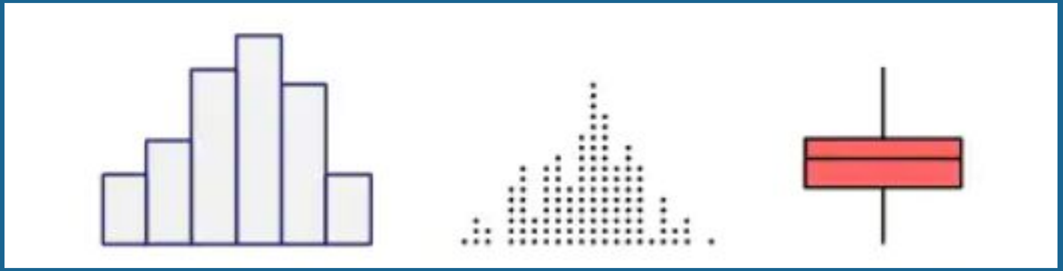
1. Measure of Central Tendency
2. Measure of Dispersion

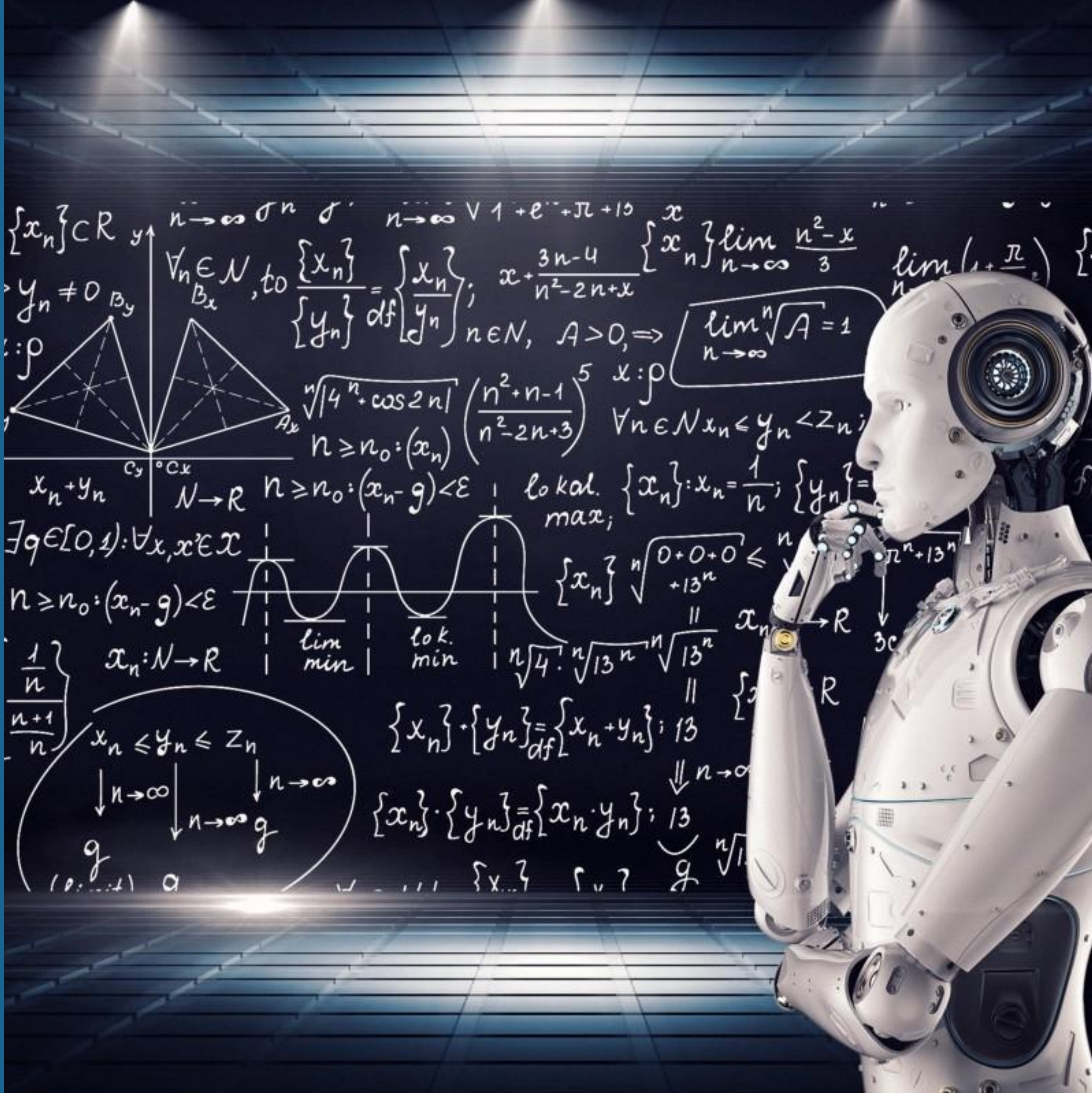


When analyzing a graphical display, you can draw conclusions based on several characteristics of the graph.

You may ask questions such as:

1. Where is the approximate middle, or centre, of the graph
2. How spread out are the data values on the graph
3. What is the overall shape of the graph
4. Does it have any interesting patterns





Outlier:

1. A data point that is significantly greater or smaller than other data points in a dataset
2. It is useful when analyzing data to identify outliers
3. They may affect the calculations of descriptive statistics
4. Outliers can occur in any given dataset and in any distribution

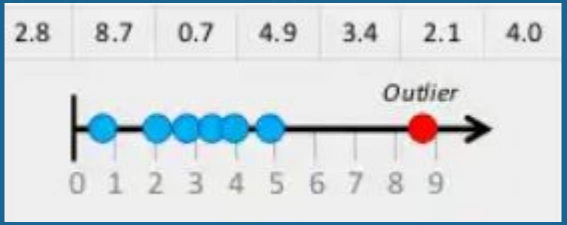
The easiest way to detect them is by **graphing the data** or using graphical methods such as:

1. Histograms
2. Box Plots
3. Normal distribution plots

- Outliers may indicate an experimental error or incorrect recording of data
- They may also occur **by chance**
  - It may be normal to have high or low data points
- You need to decide whether to exclude them before carrying out your analysis
  - An outlier should be excluded if it is due to human error

Outlier:

This example is about the time taken to process a sample of applications

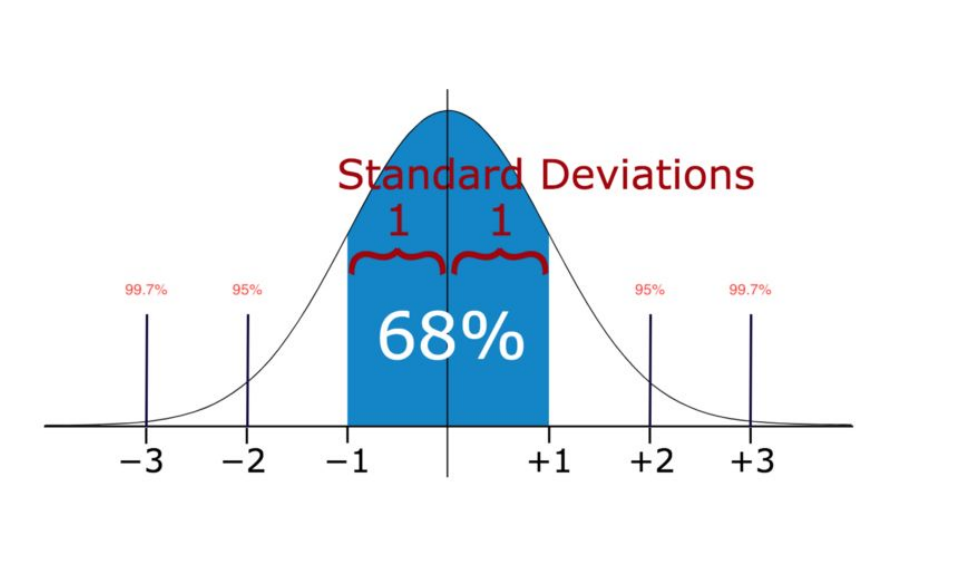


2.8	8.7	0.7	4.9	3.4	2.1	4.0

It is clear that one data point is far from the rest of the values, and hence is called as an **Outlier**

Detecting Outliers:

1. Standard Deviation: In statistics, If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations



Lets, understand what is standard deviation:

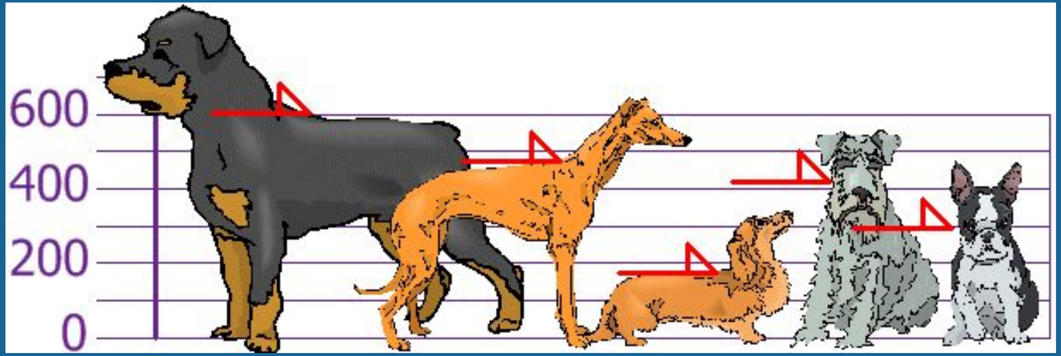
The Standard Deviation is a measure of how spread out numbers are. Its symbol is **σ** (the greek letter sigma)

The formula is easy: It is the square root of the Variance.

So now you ask, "What is the Variance?"

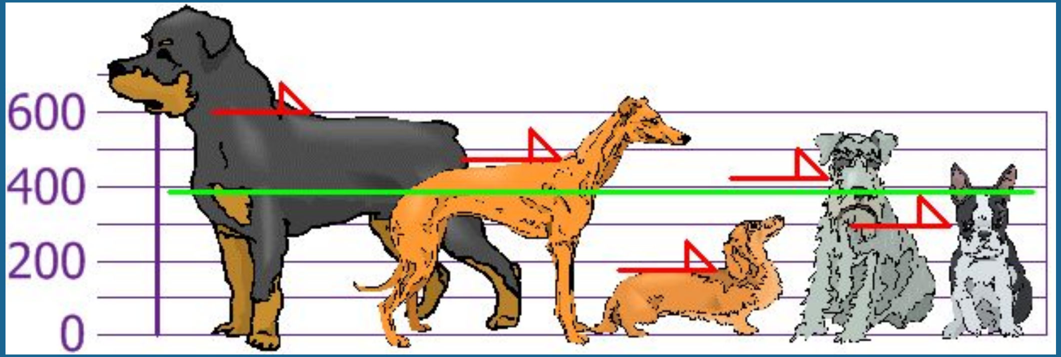
Variance: The average of the squared differences from the Mean.

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

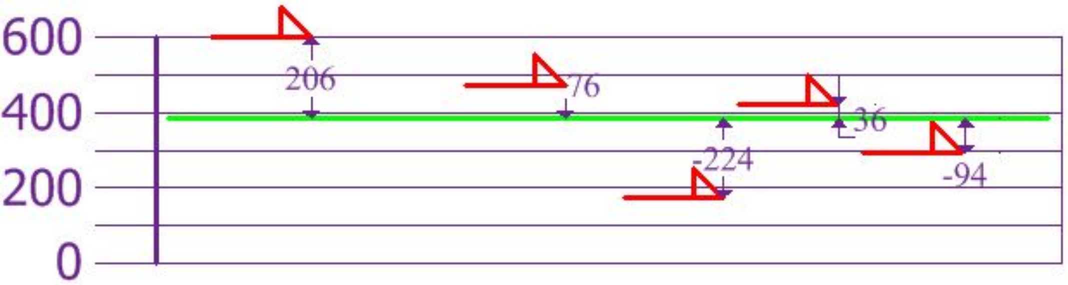


Find out mean, variance & standard deviation

Mean = 600+470+170+430+300/5 = 394



Now we calculate each dog's difference from the Mean:



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

Variance = **square(σ)**

$= 206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2 / 5$

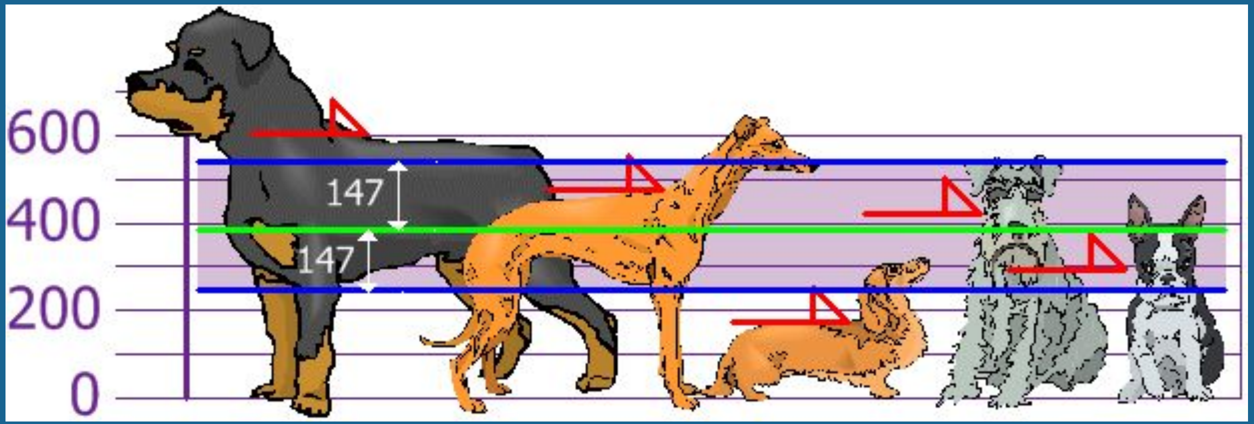
$= 108520/5$

$= 21704$

σ $= \sqrt{21704} = 147$

Example Courtesy: mathisfun.com

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, right?

We can expect about 68% of values to be within plus-or-minus 1 standard deviation.

Example Courtesy: mathisfun.com

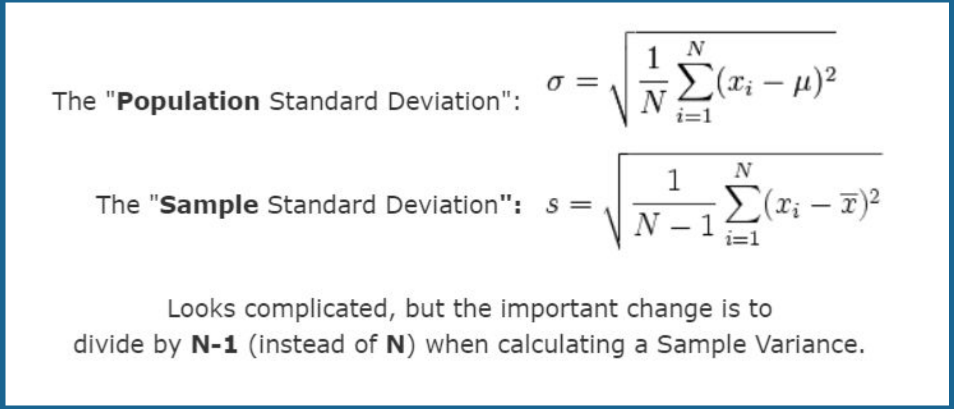
But ... there is a small change with **Sample Data**

Our example has been for a **Population** (the 5 dogs are the only dogs we are interested in).

But if the data is a **Sample** (a selection taken from a bigger Population), then the calculation changes!

When you have "N" data values that are:

- The **Population**: divide by **N** when calculating Variance (like we did)
- A **Sample**: divide by **N-1** when calculating Variance



The "**Population** Standard Deviation": $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

The "**Sample** Standard Deviation": $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$

Looks complicated, but the important change is to
divide by **N-1** (instead of **N**) when calculating a Sample Variance.

Example: if our 5 dogs are just a sample of a bigger population of dogs, we divide by 4 instead of 5, hence:

Sample variance = 108520/4 = 27130

Sample Standard Deviation = Sq root of 27,130 = 165 (nearest value)

Example Courtesy: mathisfun.com

Why **n-1** & why not n??

This is actually called Bessel's correction. The idea behind this is that this is a more unbiased measure of variance than the usual definition.

Imagine you have a huge bookshelf. You measure the total thickness of the first 6 books and it turns out to be 158mm. This means that the mean thickness of a book based on first 6 samples is 26.3mm.

Now you take out and measure the first book's thickness (one degree of freedom) and find that it is 22mm. This means that the remaining 5 books must have a total thickness of 136mm

Now you measure the second book (second degree of freedom) and find it to be 28mm. So you know that the remaining 4 books should have a total thickness of 108mm .

.

.

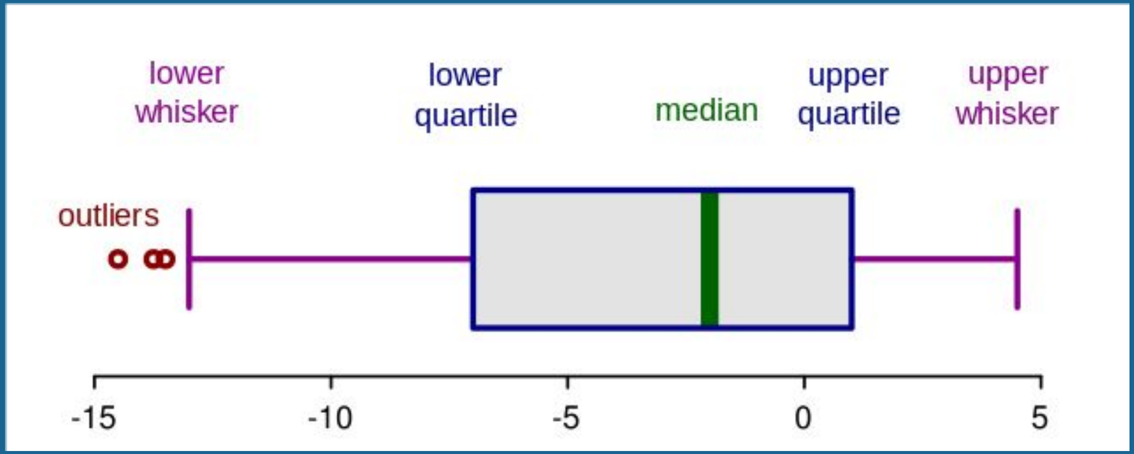
In this way, by the time you measure the thickness of the 5th book individually (5th degree of freedom) , you automatically know the thickness of the remaining 1 book.

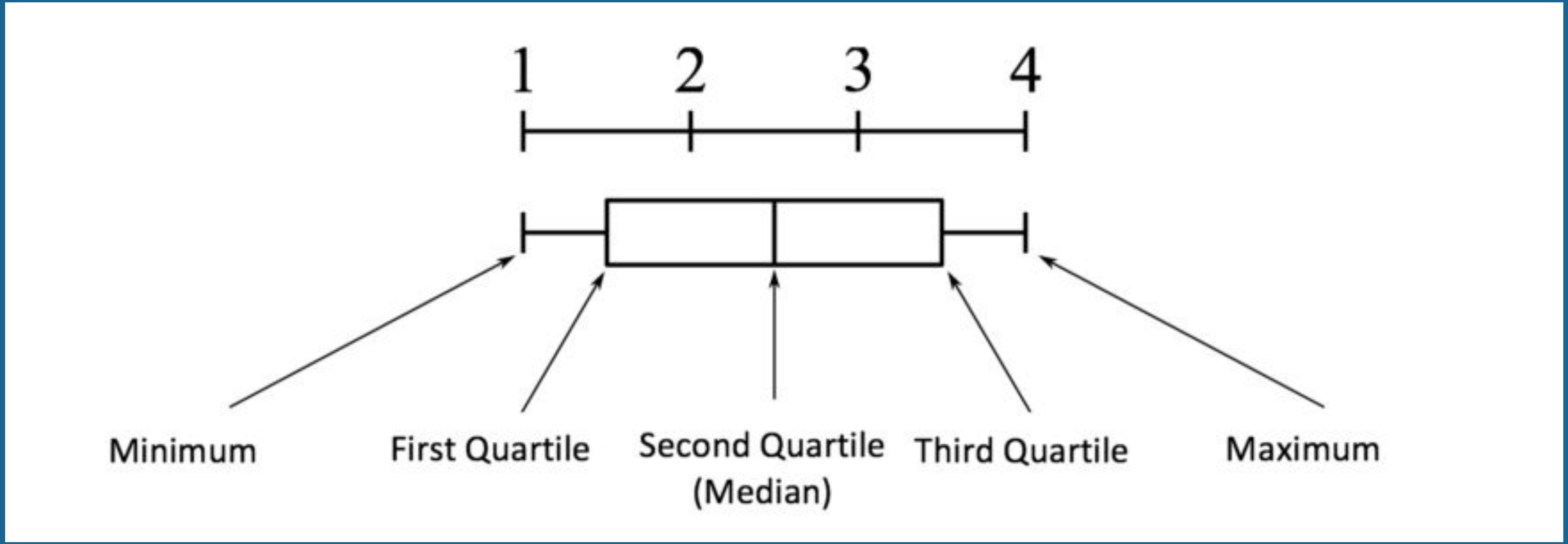
This means that you automatically know the thickness of 6th book even though you have measured only 5. Extrapolating this concept, In a sample of size n, you know the value of the n'th observation even though you have only taken (n-1) measurements. i.e, the opportunity to vary has been taken away for the n'th observation.

This means that if you have measured (n-1) objects then the nth object has no freedom to vary. Therefore, degree of freedom is only (n-1) and not n.

Detecting Outliers:

2. Boxplots: Box plots are a graphical depiction of numerical data through their quantiles.





Detecting Outliers:

3. Clustering Techniques: There are various clustering techniques such as KMeans, Hierarchical & DBScan, and all of them can be used to identify outliers

4. ML Algorithms

Measure of Central Tendency

Central tendency measures the center value of the dataset .It give us idea about the concentration of the value in the central part of the distribution.

Mean/Average.

Median.

Mode

Mean/Average

It’s an average set of observation of the data. It compute the sum of all observation present in the datasets divided by total number of observation.

Steps to find the Mean/Average:

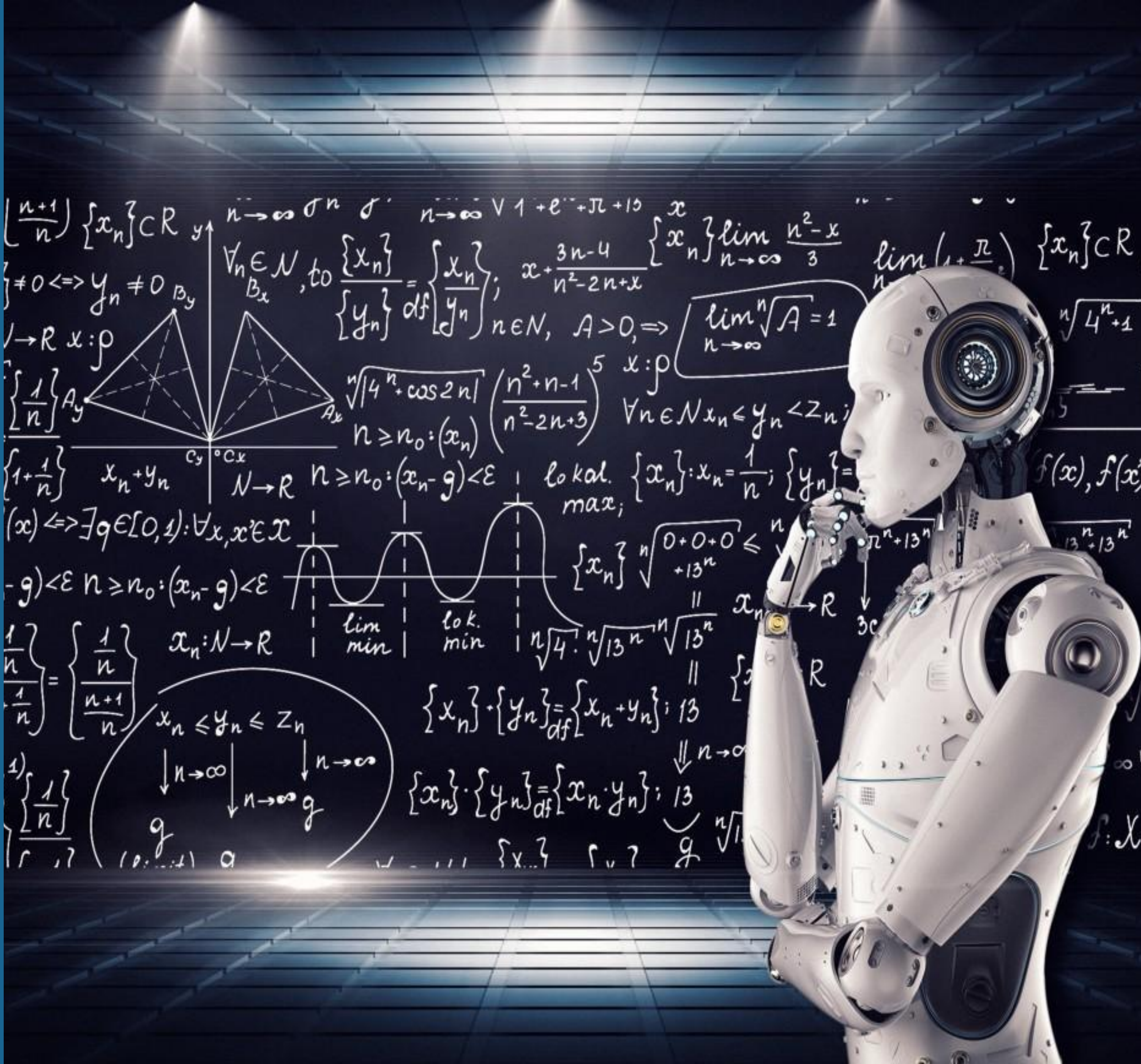
1. Add/Sum each number/observation present in the dataset.

2. Calculate the total number present in the dataset.

3. Divide sum of observation to the total number of observation.

Formula:-

$$\overline{X} = \frac{\sum X}{N}$$



Median

Median is the middle value of the entire dataset. Its split whole dataset in two parts and take the middle value of the datasets. It’s also called the 50th percentile.

Steps to find Median of the datasets:-

(For odd)

1. First Arrange the observation in Increasing order.
2. Divide the observation in two equal parts 50/50.
3. Take the middle value of the data.

(For even)

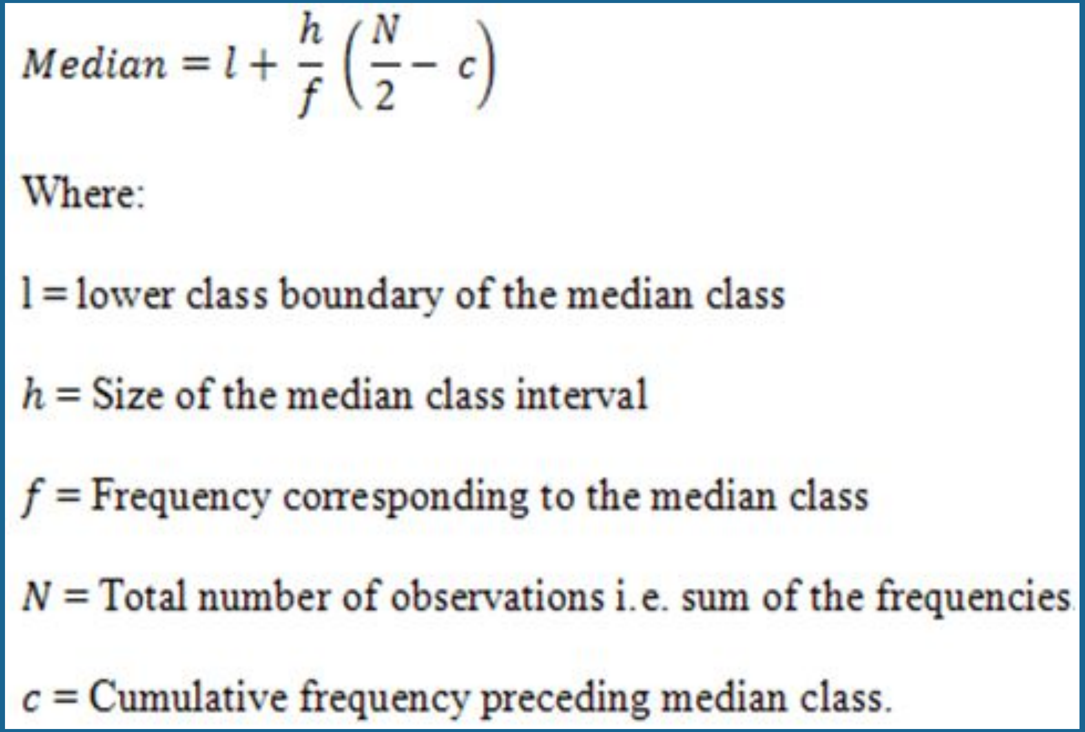
1. First arrange the values in increasing order.
2. Divide the observation in two equal parts 50/50.
3. Now take the middle two value of the dataset which is remain after dividing.
4. Now take the average of the middle two value, that is median of the dataset.

In case of discrete distribution the formula is:

Its obtained by considering the Cumulative frequency

N/2

In case of continuous distribution:



$$Median = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

Where:

l = lower class boundary of the median class

h = Size of the median class interval

f = Frequency corresponding to the median class

N = Total number of observations i.e. sum of the frequencies

c = Cumulative frequency preceding median class.

Mode

Mode is the value which occur most frequently in the set of the observation. Data can have more than one mode as Uni-model, Bi-model, Multi model. It’s Usage depends on the on situation as Max( ),Min( ),Mean( ).

Steps for finding the Mode:-
It’s very easy to find mode of any observation

1. Take the Most frequent value present in the dataset.

Special Cases:

1. If the maximum number of frequency repeated
2. If the maximum frequency is occur at the beginning and end of the observation.
3. If there is irregularities in the distribution.

In all the above cases we find the mode of the observation by using method of grouping.

In case of continuous distribution the formula is:

END OF PART 1