[**How Descriptive Stats can be this Structured?**](https://medium.com/@saikirandasari41/how-descriptive-stats-can-be-this-structured-6697b4f22202)

**Statistics Part 2 (Layman Terms 1)**

**Introduction**

Have you ever wondered how a person can cross the river by making the right decision, and without going blindly crossing to reach the other end?

A pharmaceutical engineer develops a new drug that regulates sugar in the blood. Suppose she finds out that the average sugar content after taking the medication is the optimal level. Does this mean that the drug is effective?

So, these are all small Use Cases we need to observe or think logically by using Descriptive statistics and I’ll Explain below!!

Welcome to my next Blog on “Connecting the Dots off Stats” in Layman Terms ‘Statistics Part 2’

**Table of Contents:**

**1] Measures of Central Tendency (aka 1st Moment of Business Decision).…… ->(such as MEAN, MEDIAN, and MODE),**

**2] Measures of Dispersion {2nd Moment of Business Decision (Measure of Variability)}- > (such as RANGE, VARIANCE, and Standard Deviation),…….**

**3] Measures of Position -> (Quartiles, Quantiles ->Deciles, Percentiles). . In Quartiles, we will see “BOXPLOT”**

**4] Measure of Asymmetry (3rd Moment Business decision(Skewness))….. In Skewness prior we take one example and learn “HISTOGRAM” and how it linked with Skewness!!**

**[Right/Positively Skewed Data], [Left/Negatively Skewed Data], [NORMAL Data].**

and what happens with Skewness if Mean, Median, and Mode are Greater or Lower than one another?

**5] 4th Moment Business decision (Kurtosis)**

— — — — — — — — — — — — — — — — — — — — — — — — — —

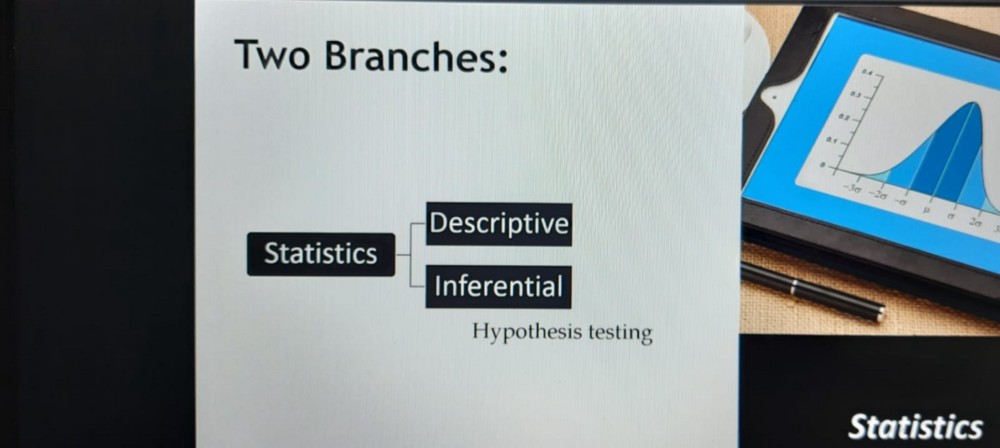
**The ideology of this Blog!**

Whatever process we are working with, we will be dealing with a lot of numbers; numbers related to output, numbers related to the input, all those processes, and how we make sense out of those numbers, that’s where statistics helps us! and I'm going to give my best in terms of QNA and aka Terms of each concept!!!

**-> There are two main branches of Statistics.**

One is Descriptive Statistics and the second one is Inferential Statistics

*These concepts are really important because they help us in conveying something that is happening in the dataset!*



**A] Inferential Statistics? (Remember this “What can be found in our data?)**

Inferential statistics is related to predicting or inferring about the data. So, in inferential statistics, we look at some sample, and based on that sample, we infer or predict the whole population. (Sample and Population were explained in my 1st Blog) [“Statistics Part-1” by Saikiran Dasari for Data Science]

**B] Descriptive Statistics? (Remember this “What can we find in our data?)**

Descriptive statistics helps us in understanding the data we have. They are often used to give an overview of a large data set or to make comparisons between groups of data.

On other hand: They are used to summarize data in an organized manner by describing the relationship between variables in a sample or population.

**When you have data, there are some things you can look they are:**

1. Measures of Central Tendency (aka 1st Moment of Business Decision).…... ->(such as MEAN, MEDIAN, and MODE),

2. Measures of Dispersion {2nd Moment of Business Decision (Measure of Variability)}- >(such as RANGE, VARIANCE, and Standard Deviation),

3. Measures of Position -> (Quartiles, Quantiles ->Deciles, Percentiles)

4. Measure of Asymmetry (3rd Moment Business decision (Skewness))

5. 4th Moment Business decision (Kurtosis).

There are also Measures of Frequency (Count, Percentage) & Measures of Association (such as Correlation and Regression)!!!

— — — — — — — — — — — — — — — — — — — — — — — — — —

**In Detail Explanation of the above 5 topics, we will discuss NOW!!!**

**1] Measures of Central Tendency (aka 1st Moment of Business Decision) ‘MEAN’, ‘MEDIAN’, ‘MODE’**

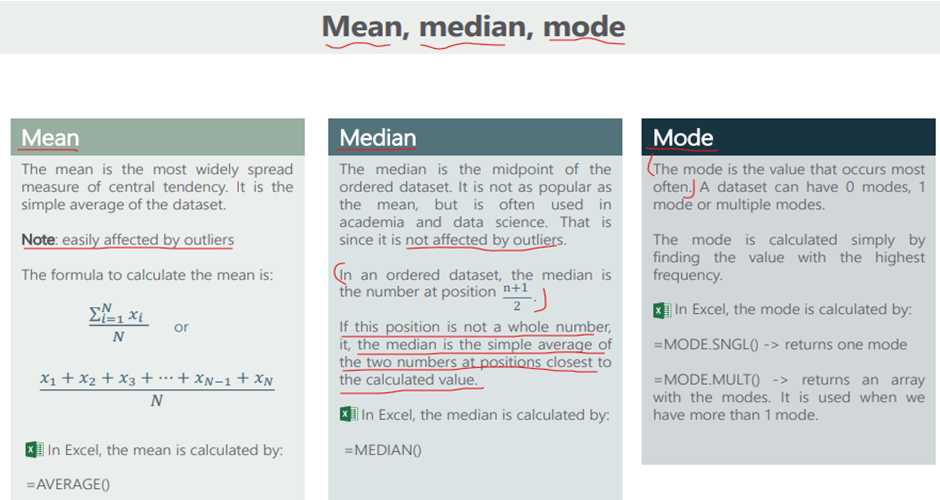
**What is Mean?** The Average of all the Datapoints in a Dataset is called the Mean. It can be calculated by adding together all the numbers in the dataset and then Divided by the Total Number of Values in the Dataset.

**What is the Median?** The Middlemost value when the dataset is sorted In Ascending Order.

**What is Mode?** The Most Repeated Values in the dataset are called Mode.

Nutshell with Formula: (showing with EXCEL formula!! very handy)

Photo representing CTs 3 main Formula



*You know what! In Case of missing values in the dataset, we Impute the values using Mean/Median/Mode*

**Note:** for Numerical continuous values and if there are no Outliers in that columns we try to impute using ‘Mean’ & but if we have missing values with outliers we go with Median imputation, & if the dataset has Categorical data, we use Mode imputation.

**Note:** for “Mode”: When the data items have Total Odd Numbers! prior arrange in Ascending (Take the middle value directly) but if the total numbers are EVEN? then Arrange the data in Ascending take the average of the middle 2 values.

— — — — — — — — — — — — — — — — — — — — — — — — — —

The Mean/Median/Mode does a nice job of telling the Center of the data set, but often we are interested more:

**2] Measures of Dispersion {2nd Moment of Business Decision (Measure of Variability)}** ->(such as RANGE, VARIANCE, and Standard Deviation)

For example, a pharmaceutical engineer develops a new drug that regulates sugar in the blood.

Suppose she finds out that the average sugar content after taking the medication is the optimal level. This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half have dangerously high content.

Instead of the drug being an effective regulator, it is a deadly poison.

What the pharmacist needs is a measure of how far the data is spread apart. This is what the Variance and Standard deviation do!!

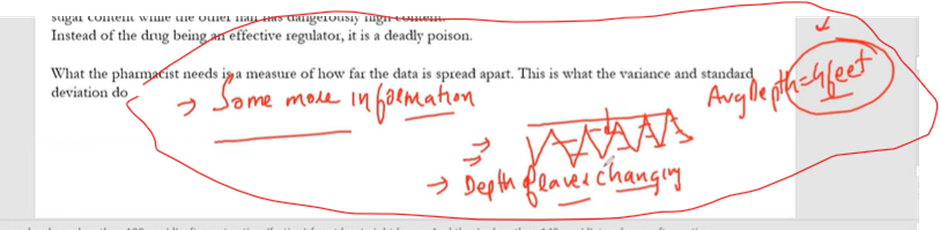
We will first discuss ‘Standard Deviation”

**a. What is Standard Deviation?**

We want to know at each data point How is the Spread changing. Deviation from the Mean values is called Standard Deviation and Standard Dev can be + or –

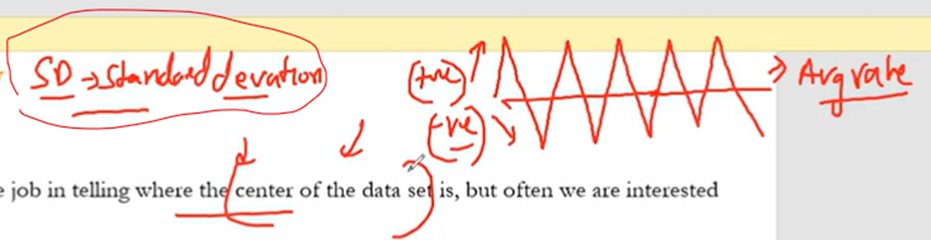
***Eg****: If there is a river and Average height of the river is 4 feet and I want to know if I can be able to cross that river? by finding it from the data we have!*

Image representing river:



Only Mean will not suffice our requirement, we want more information like at each way how much deep is the river if we know that we can answer whether we can cross that river i.e YES/NO!

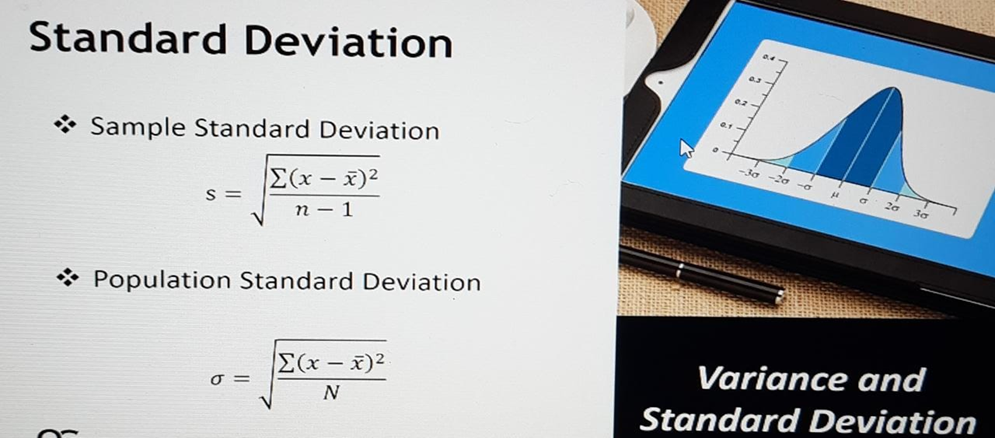
see how the from average depth of the river we can find standard deviation +/-



So, we need to understand if we want to take Better Decisions, we need to have MORE Information, So, we cannot take decisions based on the 1st Moment of Business Decisions(i.e Mean, Median, Mode) it can’t help you :( We should also take the help of the 2nd Moment of Business Decision! (Std-Deviation and Variance)!!!

Always we need to try to achieve the overall Deviation to 0

x is the value of the element at the ith position in the column & x bar is the Sample Mean & n is the number of data points



E.g: If we are preparing a ‘Dal Makhni’ Dish

If we added spices a lot more in the first stage it’s a mistake if we don’t balance it by adding water and if we add more water, we try to add more ingredients as well! likewise, we balance and try to get ‘Final Deviation’ = 0 :-)

The above e.g., reminds me of ‘Trade of between BIAS and VARIANCE’ (by far one of the important concepts in Data Science)

**b. What is Variance going to do?**

**VARIANCE is the variety/useful piece of Information (i.e. More information in the column)! (The more Variance we have, the more beneficial for us in building ML models and making our predictions)**

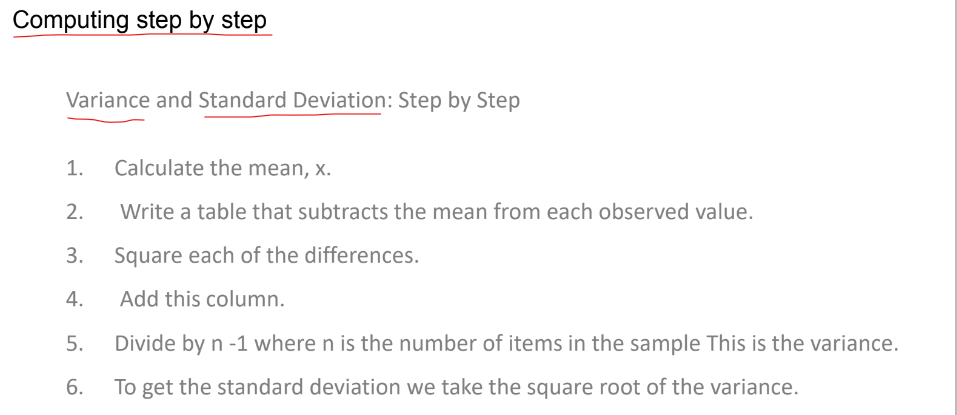
So, just because the standard deviation is 0, we cannot say that the Variation is 0 in the ingredients.

**So, how can you calculate Variations in the Ingredients?**

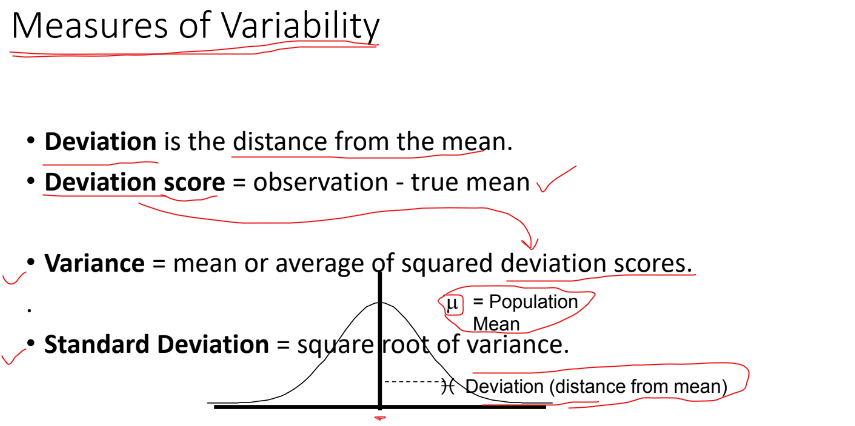
We have to use **VARIANCE**: Variance will Square all the distance from the mean and divided by a total number of values when we do this, we get ABSOLUTE +ve value which is called VARIANCE!

**To find the Standard deviation we first calculate ‘VARIANCE’!!**

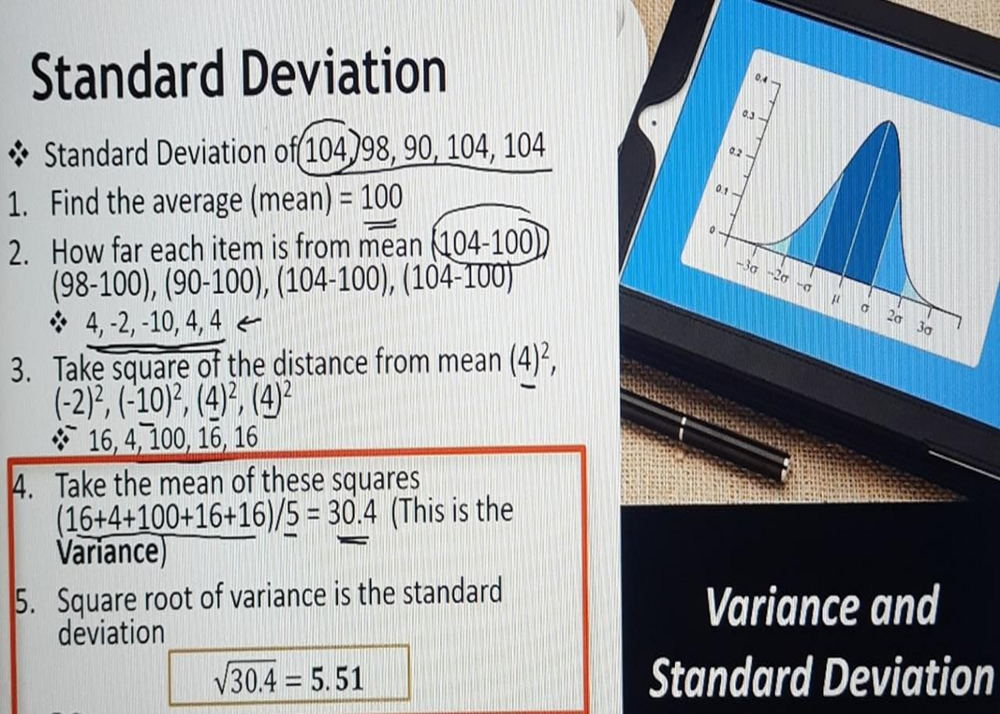
**Step By Step Calculating Variance and Standard Deviation below:**



Terminologies and Understanding the process below fig:



**Executing Now with a small sample!!**

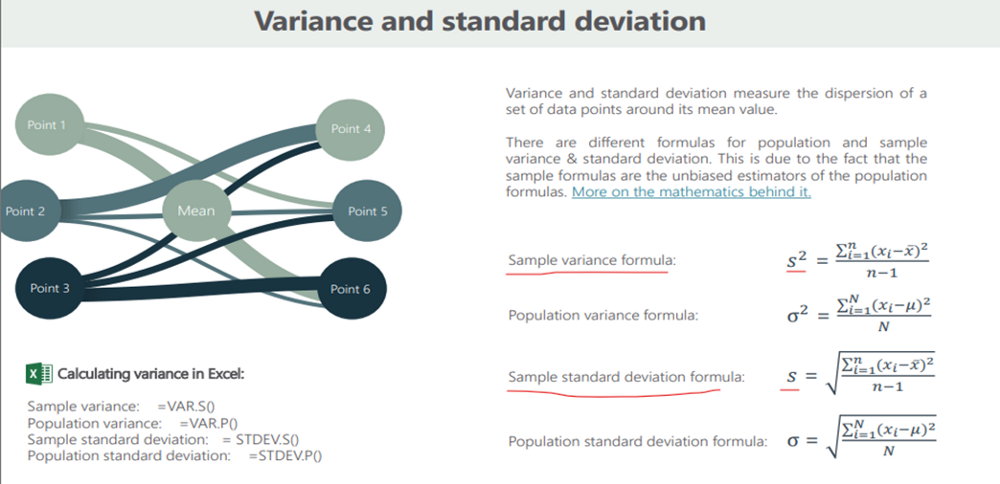


So, the Higher the positive value more is the variation (i.e., the More the Dispersion/Variability in the ingredients THE MORE is the RISK of consuming the dish)

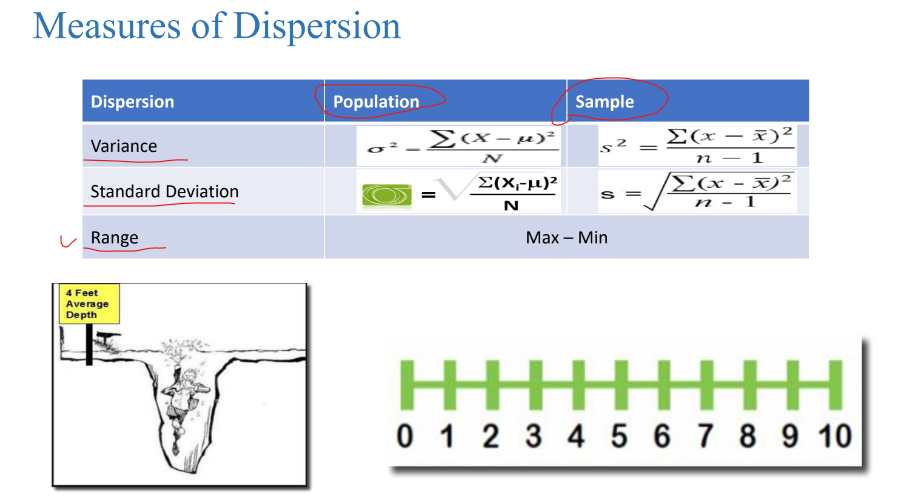
The above calculation is a bit hard way to find out the Standard deviation! isn’t it :(

There are 2 different formulas to find out Variance & Standard Deviation

Image representing Standard deviation and Variance (for Sample and Population each)



The same Image shows the formulas



The difference between these two; the sample and the population standard deviation is in the denominator in the above formula!!

In the sample, if you see it’s n minus one, and in the population, it is divided by N.

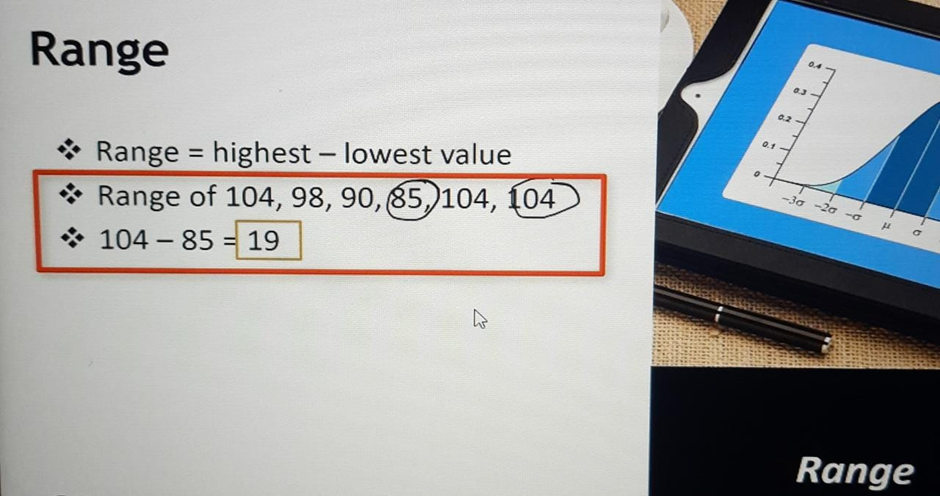
If you take a sample and you find out the standard deviation, you are not interested in the standard deviation of those samples.

What you are doing in actual reality is you are predicting the standard deviation of the whole population based on a sample and that is the reason there is a difference in the formula.

**c. What is Range?**

**Range basically tells us how much the data is Spread!!**

Along with the Mean of the river, if we give Min depth and Max depth feet (additional info) we can take decision to cross the river, that Min and Max are called the RANGE



Now, the problem with the range is just like the mean, the range is also affected by the extreme value. So, instead of 104, if I have 1040, then the range will drastically change, the range will be somewhere around 1000.

— — — — — — — — — — — — — — — — — — — — — — — — — —

**3] Measures of Position -> (Quartiles, Quantiles, Deciles, Percentiles)**

**A. What are QUARTILEs?**

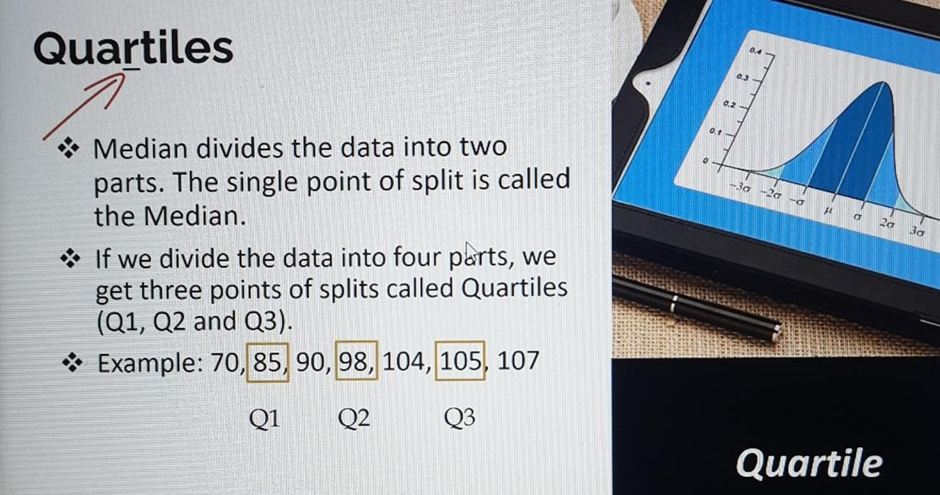
The Quartile Divides the Data into 4 Parts

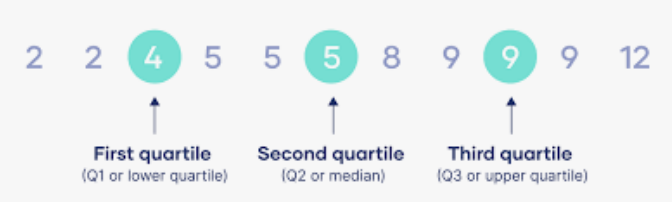
Each of four equal groups into which a population can be divided according to the distribution of values of a particular variable:

When we divide data into four parts, then we will get 3 points of the split. And these points are called Quartile.

So, like in the Median, we are left with one point if we divide the data into two parts. And that point is called the Median.

Quartiles in basic Layman's understanding!!!





And these 3 Quartiles are called Q1, Q2, Q3

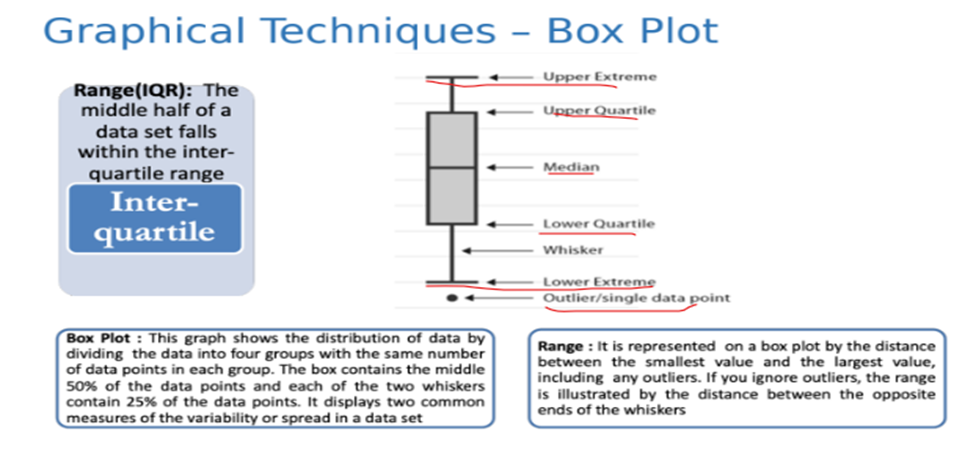
Here, the Median is also called Q2 (2nd Quartile)

And if you want to show this quartile in Visuals, we use something known as a “**Box and Whisker Plot**”

**CONNECTING Quartiles with “BOXPLOT”**

Boxplot is a Tool that is used to determine Outliers (they are extreme observations in the columns)

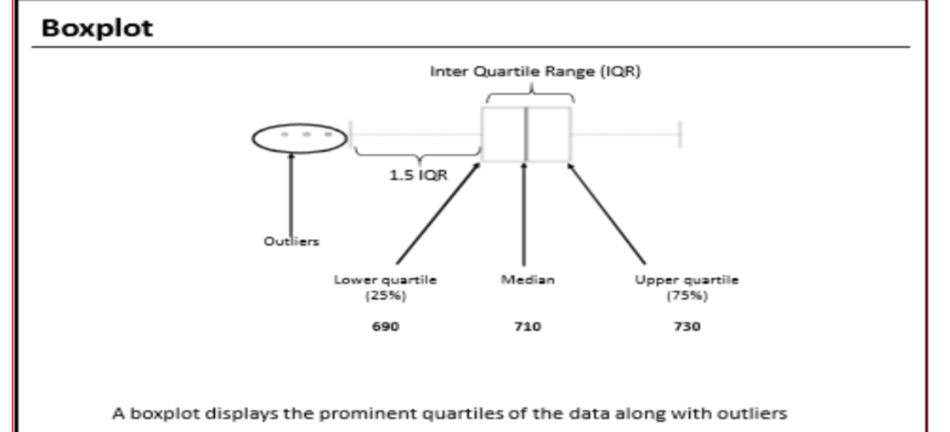
Box and Whisker Plot (Quartiles)



The Width of this plot is called IQR (Inter Quartile Range)

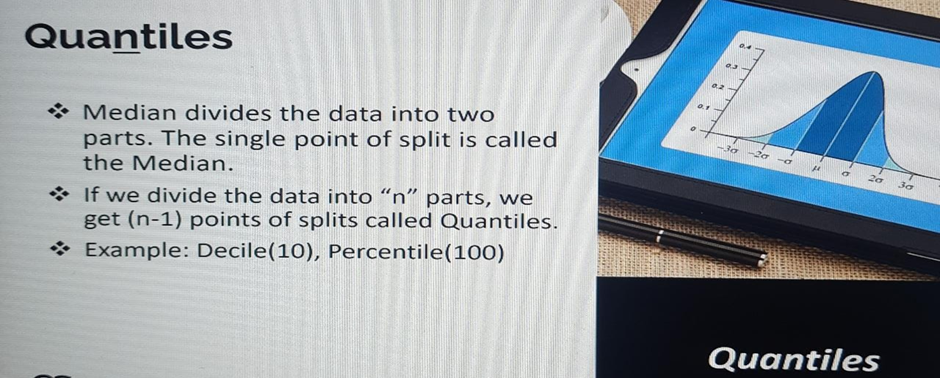
And IQR is: IQR = Q3 — Q1

In Detail Boxplot Figure



**B. What are QUANTILEs (Decile and Percentile?)?**

Quantile (default values is 4), Decile (10) & Percentile (100)



Quantile divides the data into any number of parts.

So, if this divides into any number of parts, then we will get n minus one split point.

And these are split points are called **QUANTILES**.

So, we can divide that up into **10 parts**, then which will be called **DECILE**.

And if we divide this data into **100 parts**, that will be called **PERCENTILE**.

— — — — — — — — — — — — — — — — — — — — — — — — — —

**Before Studying the 3rd and 4th Moment Business Decisions, People ask What is DISTRIBUTION in Statistics?**

-> The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur. A distribution is simply a collection of data or scores on a variable.

-> How each data points are far away from the mean centered on a scale.

**4] Measure Asymmetry (3rd Moment of Business Decision)**

**SKEWNESS?**

The most commonly used tool to measure asymmetry is skewness.

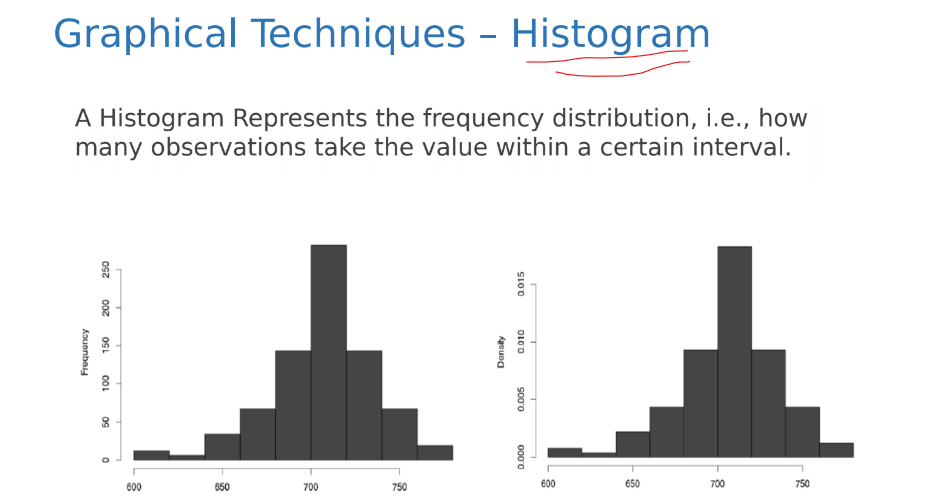
A distribution is asymmetrical when its left and right side do not mirror images

**Understand what is Symmetric and Asymmetric terms**

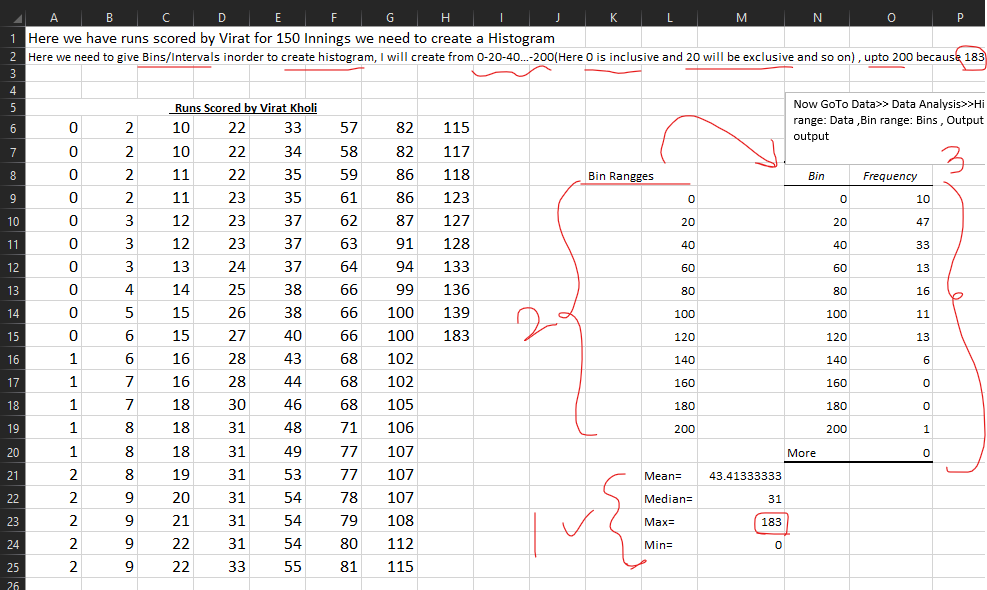
Symmetric data is observed when the values of variables appear at regular frequencies or intervals around the mean. Asymmetric data, on the other hand, may have skewness or noise such that the data appears irregular.

**Prior to Learning Skewness let's see what is HISTOGRAM as this plot will connect with the concept of SKEWNESS!!!**

The ideology of the Histogram Figure:



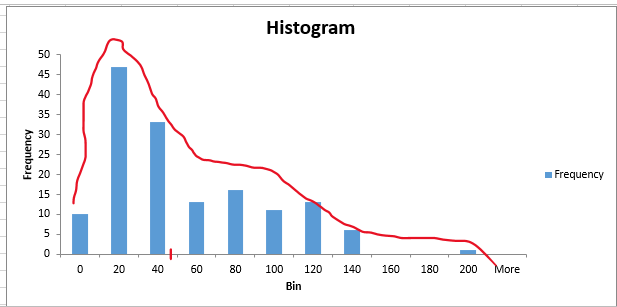
See the Example now which I Performed in EXCEL!!!



We can do this in PYTHON in seconds!!

So, now we will PLOT the Histogram Based on the above Frequency Distribution Column.

Histogram from the above Score Example, and ‘Mean’ Center is 43.4133



Now here we can infer the ‘SKEWNESS’ from the above PLOT!!

Here, by taking the Mean into consideration, we can see there is maximum data density on the LOWER side (Left Side) of the Mean value,

i.e., The bars are TALLER bars on the Lower side and SHORTER bars on the Higher side of the Mean value. (Or vice versa let’s say)

**Such Kind of Data is called “SKEWED Data”.**

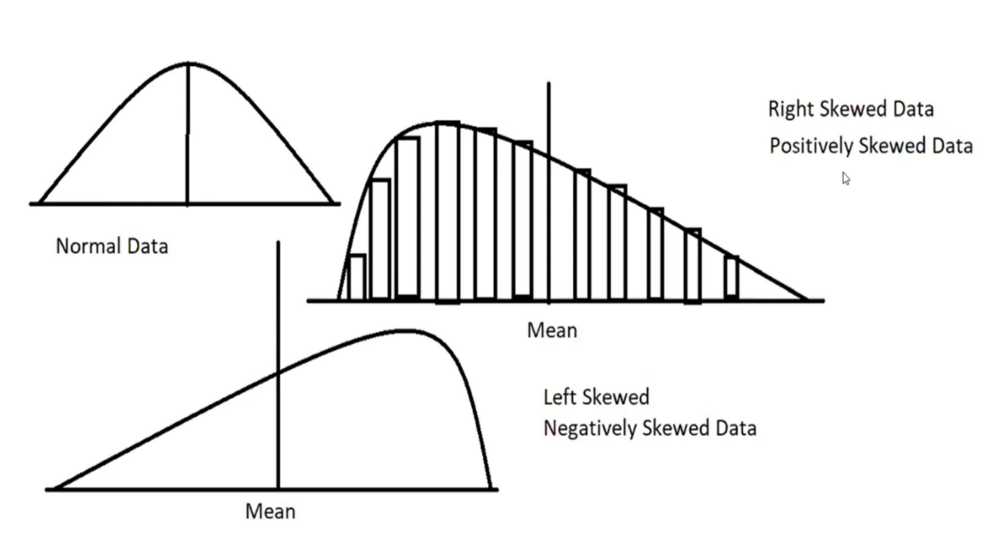
**There are Two Types of Skewness we need to understand!!**

**a. Right/Positively Skewed Data:** Whenever we have TALLER bars (more frequencies) on the left side/Lower side and SHORTER bars on the right side/Higher side of the Mean value. (And tail on the Right Side) then it is called Right/Positively Skewed Data.

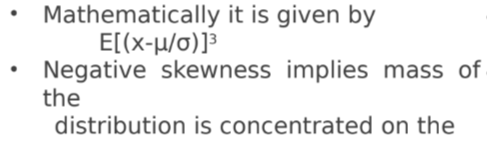
**b. Left/Negatively Skewed Data:** Vice versa of Right/Positively Skewed Data.

**c. NORMAL Data(Zero Skewness):** If the Data is Centrally located and there is a drop in frequencies on both sides equally is called NORMAL Data.

Figure representing the 2 types of Skewness:



(Formula for Skewness):

 RIGHT Side.

We need to understand this!!

The mean, median, and mode can be used to figure out if you have a Positively or Negatively skewed distribution.

**Q) What is the nature of skewness when the mean, median, and mode of data are equal?**

When mean, median, and mode are equal there is no skewness and there will be symmetric distribution or Normal Distribution, and unimodal, the skewness value will be 0.

**Q) What is the nature of skewness when Mean > Median?**

If the Mean is greater than the Median, then the distribution will be Right Skewed (Positively skewed).

**Q) What is the nature of skewness when Median > Mean?**

If the Median > Mean, then the distribution will be Left skewed (Negatively skewed).

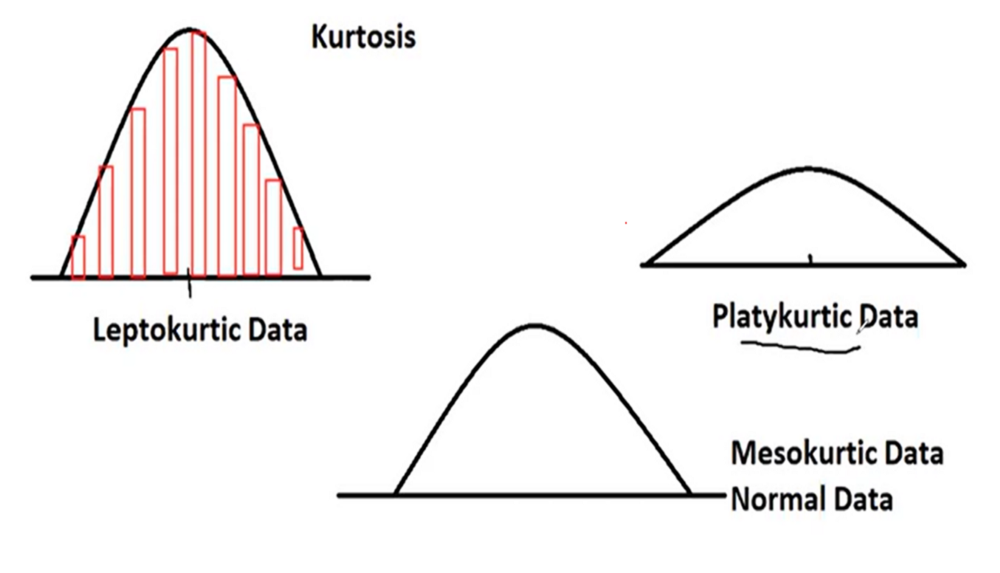
— — — — — — — — — — — — — — — — — — — — — — — — — — — —

**5] 4th Moment of Business Decision**

**KURTOSIS?**

**This basically tells us the ‘Peakedness’ of the Data Distribution**

There are 3 Types of KURTOSIS



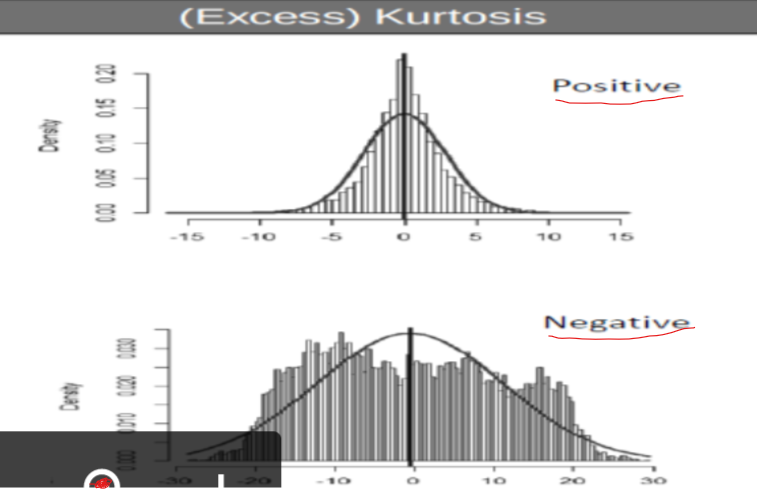
Each KURTOSIS figures

**a. Leptokurtic Data (Positive Kurtosis):** If the Data concentrated is very close to the Mean and very tall bars around the means. Such kind of histogram is Leptokurtic

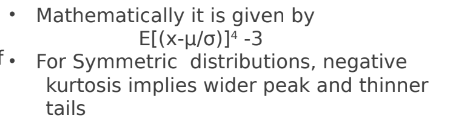
**b. Platykurtic Data (Negative Kurtosis):** If the data is spread wide apart from both extremes of Mean and has shorter bars Such kind of histogram is Platykurtic.

The Data where the std-Dev and variance are extremely high is also called Platykurtic.

**c. Mesokurtic Data/Normal Data:** If the data is in a centrally located region, which is neither spread out wide away from the mean nor is really concentrated closer to the mean, that is called Mesokurtic or Normal data



**Kurtosis Formula**



— — — — — — — — — — — END — — — — — — — — — — — — — — — — —

**Summary**

**What have we learned? We have learned —**

1. We took 2 Use Cases and saw how to not blindly conclude anything!!

2. Two main branches of Statistics (Explained): Inferential and Descriptive

3. Drilled into “DESCRIPTIVE Statistics”

1) Measures of Central Tendency (aka 1st Moment of Business Decision).…… ->(such as MEAN, MEDIAN, and MODE), ……{With Examples and Formulas}

2) Measures of Dispersion {2nd Moment of Business Decision (Measure of Variability)}- > (such as RANGE, VARIANCE, and Standard Deviation),…….{With Examples and Formulas}

3) Measures of Position -> (Quartiles, Quantiles ->Deciles, Percentiles)…………{With Examples and Formulas} …. In Quartiles, we discussed “BOXPLOT”

4) Measure of Asymmetry (3rd Moment Business decision(Skewness))….. {With Examples and Formulas}… In Skewness prior we took one example and learned “HISTOGRAM” and how it linked with Skewness!!

About [Right/Positively Skewed Data], [Left/Negatively Skewed Data], [NORMAL Data].

and what happens with Skewness if Mean, Median, and Mode are Greater or Lower than one another?

5) 4th Moment Business decision (Kurtosis)… {Understanding}

about {[Leptokurtic Data (Positive Kurtosis)], …………………………[Platykurtic Data(Negative Kurtosis)], [Mesokurtic Data/Normal Data]

— — — — — — — — — — — — — — — — — — — — — — — — —

I have tried to write a detailed article and I hope that I am successful in doing so. I’ll try to keep on adding More Content that will link to each other!!