**Source of the material :** <https://www.youtube.com/watch?v=6fUYt1alA1U&list=PLZoTAELRMXVMhVyr3Ri9IQ-t5QPBtxzJO&index=12> (Statistics Playlist, Krish Naik)

**Random Variable**

What is a random variable?

Any variable (placeholder to store something)

Why is it important?

Every feature in a dataset is essentially a random variable

Types of Random Variable:

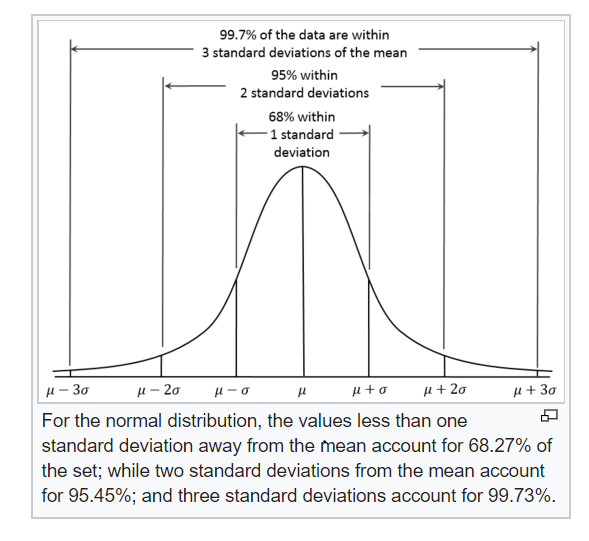
1. Numerical variable -> e.g. Age (numerical values)
   1. Discrete Random variable -> takes the values of a whole number e.g. Number of bank accounts a person holds, population of a state
   2. Continuous Random variable -> takes any value within a range e.g. Height of different people, salary
2. Categorical variable -> e.g. Gender (distinct number of categories)

**Types of Distribution**

**Gaussian distribution (Normal Distribution)**

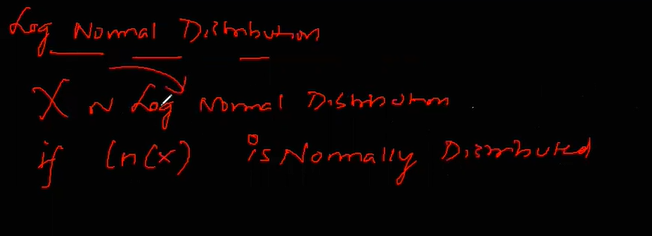
Plotted for a continuous random variable

Empirical formula for Gaussian distribution (forms bell curve) -> specifies the distribution of the data as shown below:



**Standard Normal Distribution: Mean =0,Standard deviation=1**

**Log Normal Distribution**



It takes the shape of a right skewed curve

Example: Income of the people (very less number of rich people. As income increases, the number of people earning that decreases)

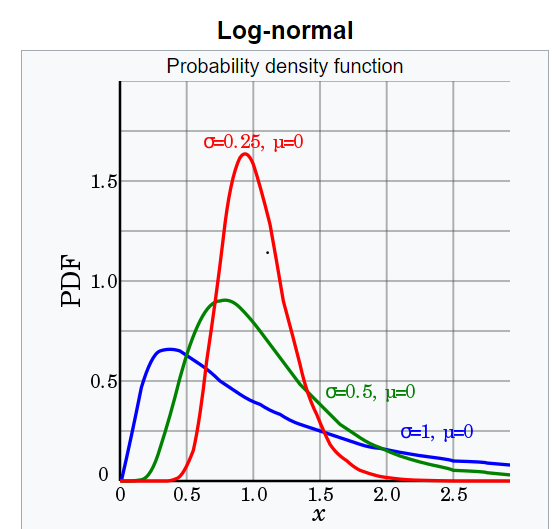
Why do we learn about different types of distributions: Let us take an example of the budget of a company in R&D and Marketing and we have to determine the profit based on this?

R&D and Marketing maybe having completely different

Let’s say R&D follows a Gaussian Normal Distribution(GND), we can standardize it using StandardScaler()

Let’s say Marketing follows a log normal distribution-> I will take log of all the values in this column and then it will follow a GND, which I can standardize using StandardScaler()

Both R&D and Marketing can now be brought on the same scale and various algorithms can be applied



**Measures of central tendency**

Video source: <https://www.youtube.com/watch?v=GvftKv9uctk&list=PLZoTAELRMXVMhVyr3Ri9IQ-t5QPBtxzJO&index=5>

1. Mean(µ)-> Mean is greatly affected by outliers
2. Median( Median helps us to lessen the impact of outliers when denoting the central value)
3. Mode-> Denotes the number with the highest frequency

We handle the missing values by taking a mean of the other numbers for attributes like Age(which are independent) in feature selection

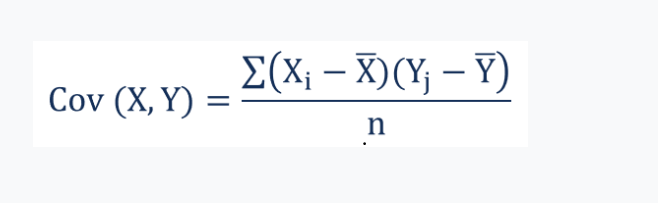
**Covariance**

 A measure of the relationship between two random variables. However, the metric does not assess the dependency between variables. The metric evaluates how much – to what extent – the variables change together.

The values are interpreted as follows:

Positive covariance: Indicates that two variables tend to move in the same direction.(y increases as x increases)

Negative covariance: Reveals that two variables tend to move in inverse directions.(y decreases as x increases)



Where:

Xi – the values of the X-variable

Yj – the values of the Y-variable

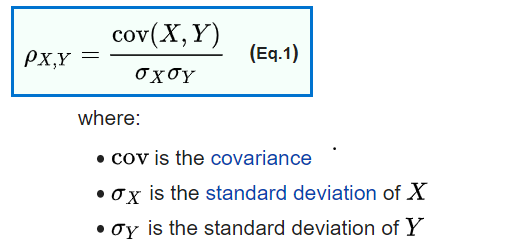
X̄ – the mean (average) of the X-variable

Ȳ – the mean (average) of the Y-variable

n – the number of data points

Cov(x,x)= Var(x)

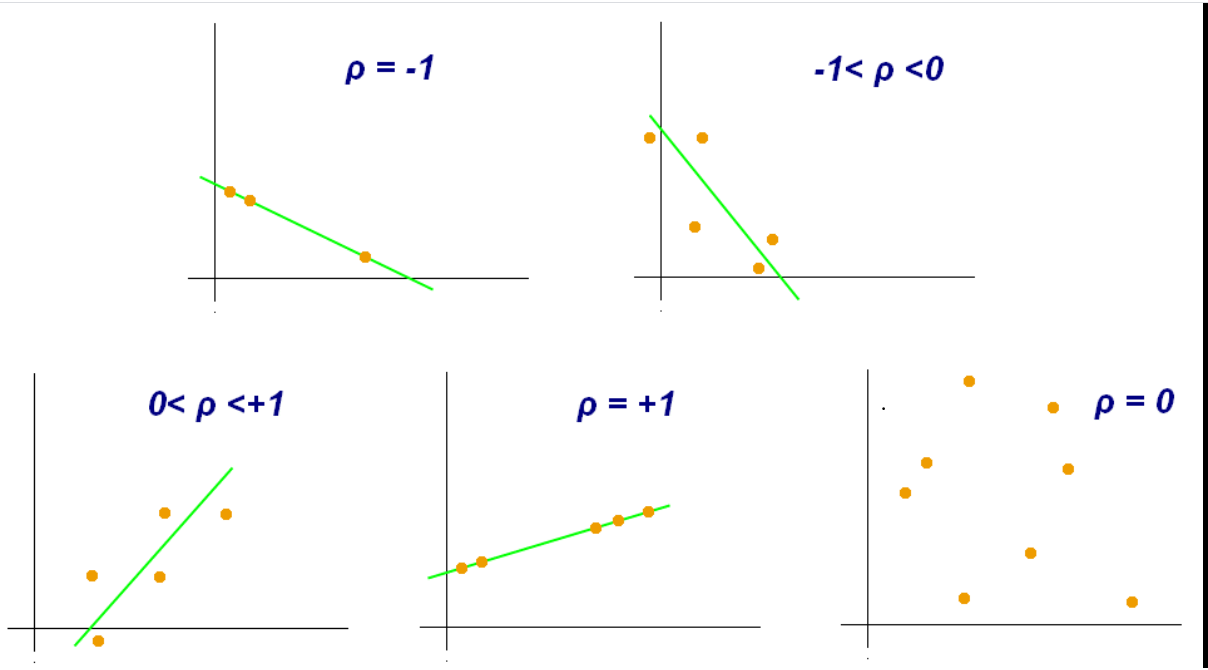
**Pearson Correlation Coefficient**



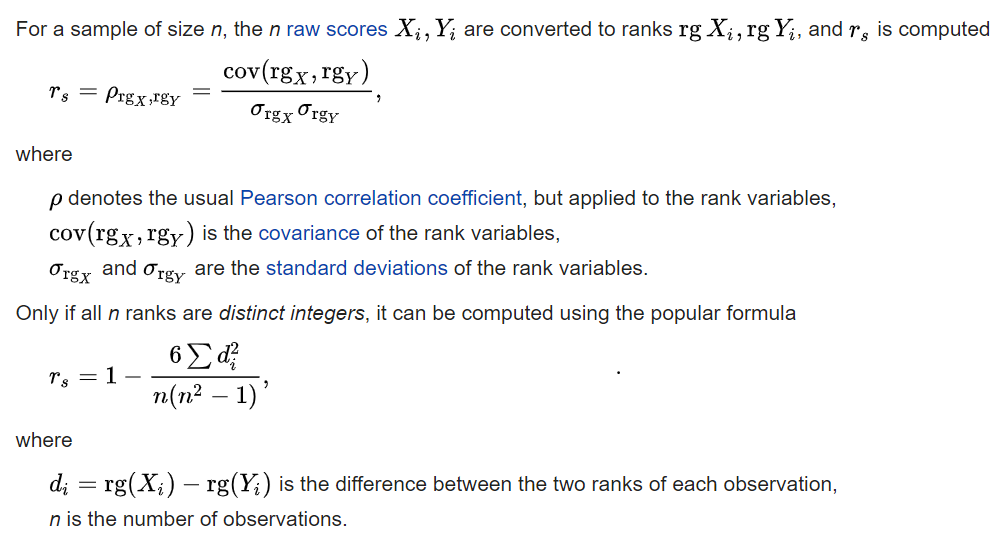
**Based on the covariance, Pearson Correlation Coefficient can give you:**

1. **Strength of the relationship**
2. **Direction of the relationship**

**-1<= Pearson Correlation Coefficient <= 1**



If Pearson correlation coefficient is 1, X and Y are essentially same. Only one can be used in the data analysis

**Spearman’s Rank Relation Coefficient**

**Example :** <https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>

Advantage:

Even though the data is non-linear (unlike Pearson), you can give you a better understanding of the relationship between two variables( can account for the outliers too)

**Most of the correlation in the EDA uses the Spearman Rank Correlation Coefficient**