1. Hypothesis Testing

* Validate or reject a hypothesis using statistical methods
* **Null hypothesis** states that there is **no** relationship between the measured phenomenon (the dependent variable) and the independent variable.
* The first step in testing a hypothesis is to assume that the **NULL** hypothesis is **true.** If the number of samples is sufficiently large (say >30). We can assume that:
  + Sampling distribution of the variable are normally distributed around its mean (as per **Central Limit Theorem**)
  + The population standard deviation is almost the same as the sample standard deviation
* Second step -> compute the probability of the observed mean.

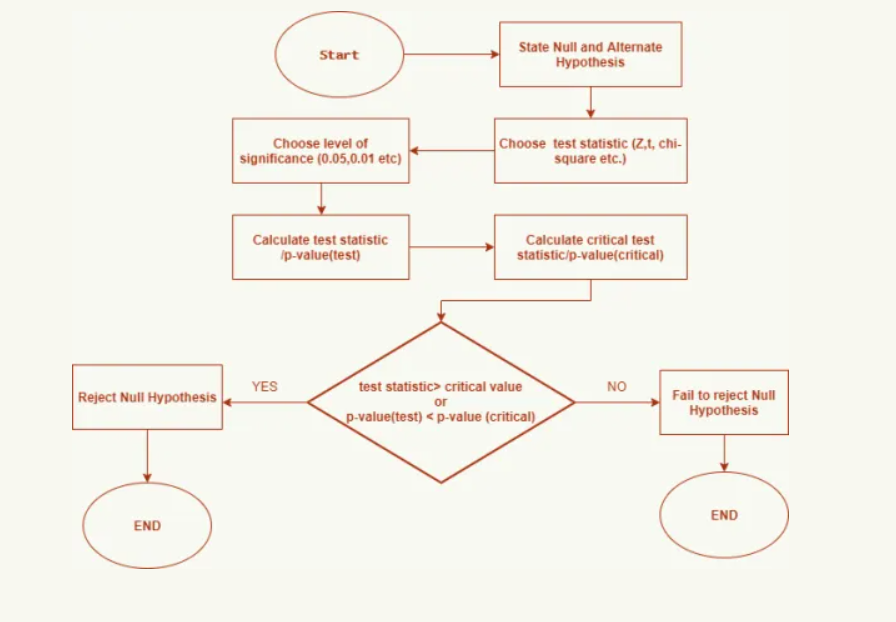
P(data/**Null) = ?**

How extreme is our observation when it is assumed that the **NULL is true?**

* *if our data lies under top 5% extreme values, we can say that given that the* ***NULL*** *is true, our data is highly unlikely, and we* ***REJECT*** *the* ***NULL*** *hypothesis.*
  + If p < alpha, we can reject the NULL hypothesis and assume alternate to be true.
  + If p>=alpha, we would FAIL to reject the NULL hypothesis.
* **P-value:** *The p-value is the probability of getting the observed value given the NULL hypothesis is true. Mathematically we can represent it as,*

*P(sample\_statistic/NULL)*

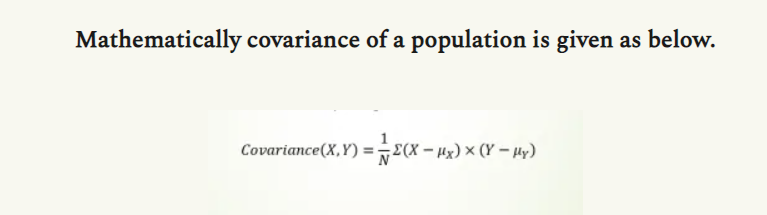
*The smaller the p-value, the greater is the evidence against the NULL hypothesis.*



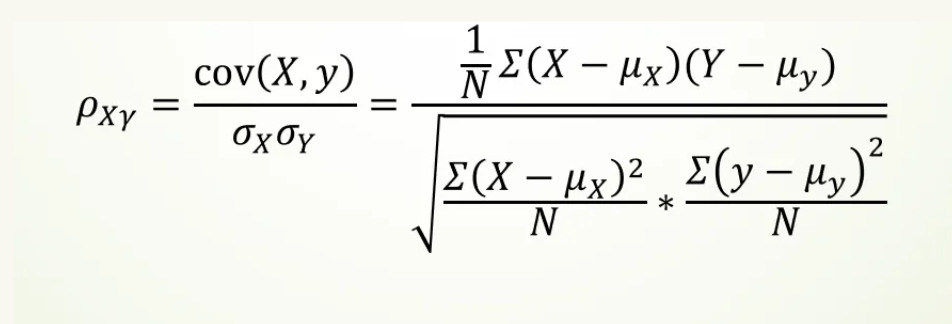
1. Covariance, Correlation and Causation
   * *Studying the relationship between two variable X and Y.*
     1. *When X get bigger, does Y get bigger, or does it get smaller? (****direction) 🡪 covariance***
     2. *Does Y get a lot bigger/smaller, or just a little bit? (****strength)***

*Both answers 🡪* ***Correlation***

* *Covariance measure the directional relation between two random variables. It measures this by comparing the variances of both the variables.*



* + - * *A positive value of covariance means that the variables are positively related*
      * *A negative value of covariance means that the variables are negatively related*
      * *There are two issues with Covariance*
        1. ***Problem of units:*** *The larger the X and Y values, the larger the covariance. A value in smaller units will be small and higher units will be high*
        2. ***Problem of Scale:***
* ***We need to have a metric which is independent of these variations***
* *Correlation: is Covariance where normalization is done with respect to standard deviation of two different distributions. One of the most popular correlation is known as* ***Pearson’s Correlation.***



*Pearson’s correlation coefficient explains about the direction as well as the strength of relationship between two variables. Value range from -1 to 1. Correlation doesn’t mean Causation.*

