**Hypothesis Testing**

Let’s consider an assumption that you need to buy good bike with good average mileage:

* You go to the market initially.
* Collect the quotations.
* Discover the requirements such as good mileage.

To buy a bike you look at the required parameter such as good mileage and you don’t directly buy a bike you first look at the market feedback about the bike weather the mileage of bike is good or bad and then make decision.

**Hypothesis is nothing but an assumption based on the parameter value.**

1. Simple and composite hypothesis:

Simple Hypothesis: When an assumption specifies an exact value is called Simple hypothesis.

Eg: buy a bike with good average let’s say 100KMPL.

Complex hypothesis: When an assumption specifies a range of value is called complex hypothesis.

Eg: Average height of students is greater than 20.

1. Null hypothesis:

The null hypothesis is the **hypothesis which is to be tested for the possible rejection under the possible assumption** i.e. true.

We assume that the null hypothesis is true until we don’t have enough evidence to prove that it is false.

The null hypothesis is denoted by **H0**.

1. Alternative null hypothesis:

It is opposite of Null hypothesis and denoted by **H1**.

*Note: Hypothesis is similar to the person standing in court and waiting for judgment, based on the available facts judge will make the decision whether the person is guilty or not guilty,*

*According to the law, we always assume that the person is innocence until we have enough evidence to prove that the person is guilt. This is called Null hypothesis.*

*Whereas the person is proven guilty based on the evidence, this is called Alternative null hypothesis.*

Whenever we are considering the H0 value consider it as Positive value and H1 as negative value.

Example: A soap company claims that the product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis.

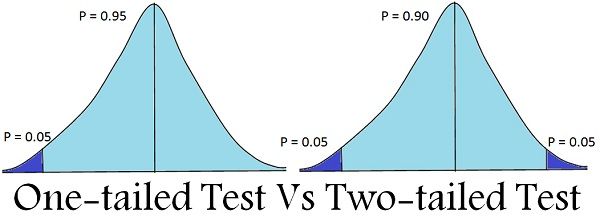
Null Hypothesis (H0): Average =99%

Alternate Hypothesis (H1): Average is not equal to 99%.

**Hypothesis Testing**

It is the process of testing the hypothesis whether based on the available condition we need to accept the null hypothesis or to reject null hypothesis and accept the alternative hypothesis.

1. One tail test and two tail test:



From the above image we can say that “*If the alternative hypothesis gives the alternate in* ***both the direction*** *(Less than or greater than) the value of parameter assign by the null hypothesis is called* ***two tail test.*** *If the alternate hypothesis gives the tail at only* ***one direction*** *(either less or greater than) of the value of parameter assign by null hypothesis is called* ***one tail test****.*”

Let’s consider the example: Given statement is H0: mean = 100, H1: mean is not equal to 100.

H0=100, H1 = can be any value greater or lesser than 100.

According to H1 mean value can be anything greater or lesser than H0 this is example of two tail test.

If the condition is given as H0>=100, then H1 will be <100 this is what called one tail test.

**Critical Region**: The critical region is the region in the **sample space in which if the calculated value lie we are going to reject the null hypothesis (H0).**

Eg: Suppose you are looking to rent an apartment. You listed out all the available apartments from different real state websites. You have budget of Rs. 15000/ month. You cannot spend more than that. The list of apartments you have made have price ranging from 7000/month to 30,000/month.

You have now random selected an apartment. Then Ho and H1will be:

H0: You will pay the rent.

H1: Not pay the rent.

As your budget is 15000/- so you will be reject the apartment whose rent is above 15000/-

Here if the price greater than 15000/- then it lie sample space or it become critical region. At this point you are going to reject the H0 and accept H1 else if it doesn’t lie in the critical region or sample space them we will accept H0 and reject H1.

To determine whether critical region fall under one tail or two tail is based on the probability distribution curve according to the H1.

**Critical region is denoted by alpha (α).**

Image based on Two tail test:

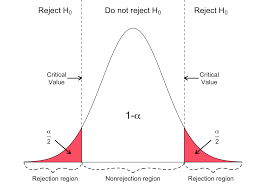
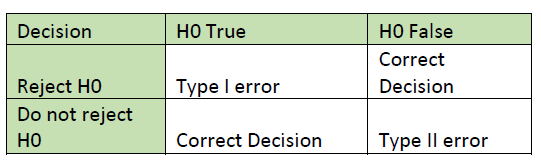


Image based on one tail test:



Type I and Type II error:



**Type I error:** suppose my null hypothesis (H0) is true and I don’t have enough information to prove the null hypothesis (H0) is true, then it is directly going to reject null hypothesis (H0) and accepting the alternative null hypothesis (H1).

α = P[reject H0 when H0 is true]

Type II error: suppose my alternative null hypothesis (H1) is true, but I don’t have enough information to prove it, in this situation we are going to reject alternative null hypothesis (H1) and accept null hypothesis (H0).

β = P[accepting H0 when H1 is true]

Level of significance (α): If we are going to accept H0 at a significance level of 5% then only we can say that our null hypothesis is true with 95% of assurance.

*Note:* ***Steps involve in Hypothesis Testing***

*1) Setup the null hypothesis and the alternate hypothesis.*

*2) Decide a level of significance i.e. alpha = 5% or 1%*

*3) Choose the type of test you want to perform as per the sample data (z test, t test, chi squared etc.) (we will study all the tests in next section)*

*4) Calculate the test statistics (z-score, t-score etc.) using the respective formula of test chosen*

*5) Obtain the critical value for in the sampling distribution to construct the rejection region of size alpha using z-table, t-table, chi table etc.*

*6) Compare the test statistics with the critical value and locate the position of the calculated test statistics i.e. is it in rejection region or non-rejection region.*

*7)* ***I)*** *If the critical value lies in the rejection region, we will reject the hypothesis i.e. sample data provides sufficient evidence against the null hypothesis and there is significant difference between hypothesized value and observed value of the parameter.*

***II)*** *If the critical value lies in the non- rejection region, we will not reject the hypothesis i.e. sample data does not provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of parameter is due to fluctuation of the sample.*

**P-value:** It is the smallest level of significance at which a null hypothesis can be rejected.

For right tailed test:

P-value = P[Test statistics >= observed value of the test statistic]

For left tailed test:

P-value = P[Test statistics <= observed value of the test statistic]

For two tailed test:

P-value = 2 \* P[Test statistics >= |observed value of the test statistic|]

Decision making with P-value:

The P-value is compared to the significance level(alpha) for decision making on null hypothesis.

1. If the P-value is **greater than alpha, we do not reject the null hypothesis.**
2. If the P-value is **smaller than alpha, we are going to reject null hypothesis.**

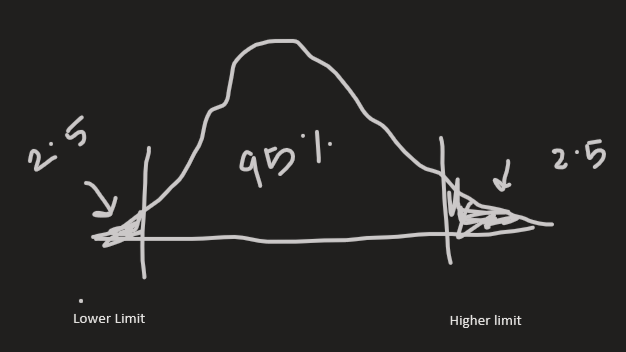
Point Estimator (PI):

This point estimator is important to calculate Confidence interval. Usually most of the expt. You will not know the population mean as the size of the population is very high. Instead of that we can calculate the sample mean which will be represented as x bar. With the help of the x bar we will try to estimate the parameter of population mean. In this case x bar represent the point estimator.



Based on central limit theorem: If we collect multiple sample form the population data and try to calculate the sample mean then it can be approximately equal to the population mean.

Confidence Interval (CI): By considering the point estimator and with some confidence interval 95% I will be able to find out the range population mean probably from what limit to what limit it may fall out and the range may be 95% confidence interval.



There are two way to solve the problem:

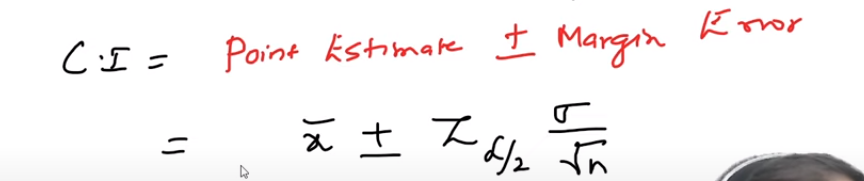
1. If you are given population standard deviation.
2. If you are not given population standard deviation.

Example: Please find out the average size of sharks in the sea.

Suppose that **population standard deviation is given** say =100

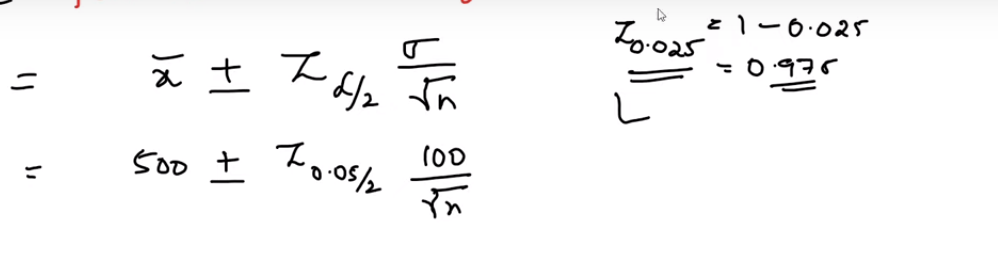
Sample size (n)>=30, sample mean random calculation say x bar = 500,CI = 95%

Formula:

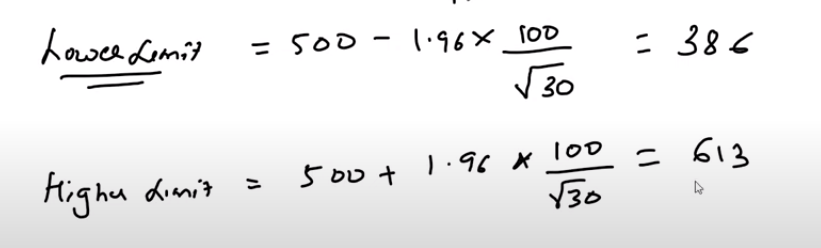
where α is the significance value and the representation is 0.05 where 95% CI



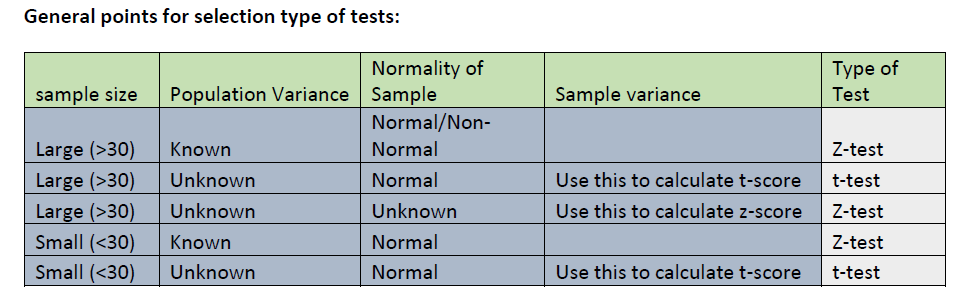
Here we are using z-test as population standard deviation is given.

****

As my entire region is 1 so we need to subtract 1-0.05/2= 0.975

****

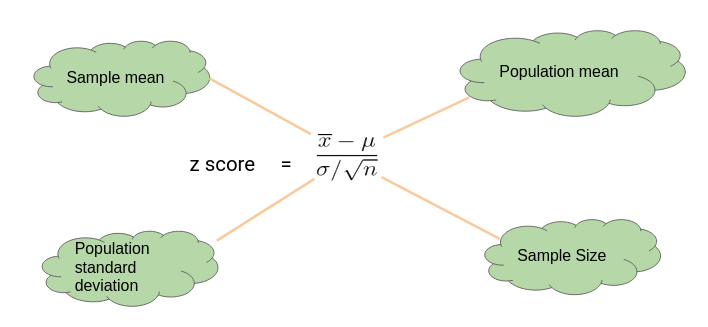
Now we have calculated the range and it says that when I have a confidence interval 95% then my data will lie between 386 to 613.



Z-test: It is the statistical way of testing a hypothesis. To work with Z-test we need to know the Population standard deviation or population variance and the size of the sample should be always >=30.

If my sample size is <30 then I need to use T-test.

One sample Z-test: we need to perform one sample Z-test when we need to compare a sample mean with the population mean

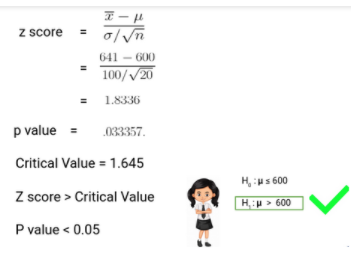


Let’s consider the example: Let’s say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls’ scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our ⍺ value (significance level) to be 0.05.

Random score is collected and result is:

In this example:

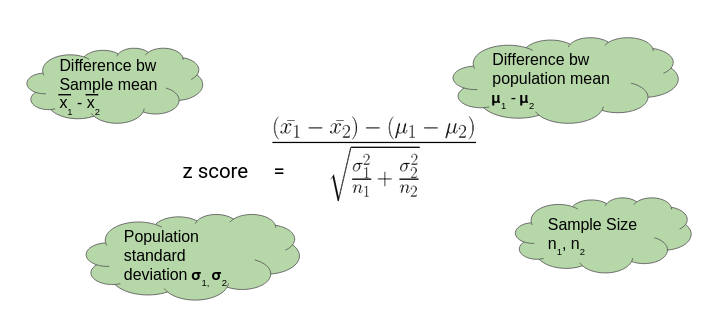
* Mean Score for Girls is 641
* The size of the sample is 20
* The population mean is 600
* Standard Deviation for Population is 100



**Since the P-value is less than 0.05, we can reject the null hypothesis** and conclude based on our result that Girls on average scored higher than 600.

Two Sample Z-test:

We perform a Two Sample Z test when we want to compare the mean of two samples.

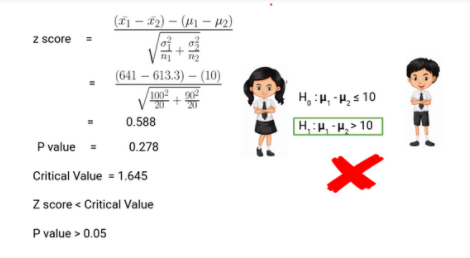


Example to Understand a Two Sample Z Test

Here, let’s say we want to know if Girls on average score 10 marks more than the boys. We have the information that the standard deviation for girls’ Score is 100 and for boys’ score is 90. Then we collect the data of 20 girls and 20 boys by using random samples and record their marks. Finally, we also set our ⍺ value (significance level) to be 0.05

In this example:

* Mean Score for Girls (Sample Mean) is 641
* Mean Score for Boys (Sample Mean) is 613.3
* Standard Deviation for the Population of Girls’ is 100
* Standard deviation for the Population of Boys’ is 90
* Sample Size is 20 for both Girls and Boys
* Difference between Mean of Population is 10



Thus, we can **c**onclude based on the P-value that we fail to reject the Null Hypothesis. We don’t have enough evidence to conclude that girls on average score of 10 marks more than the boys.

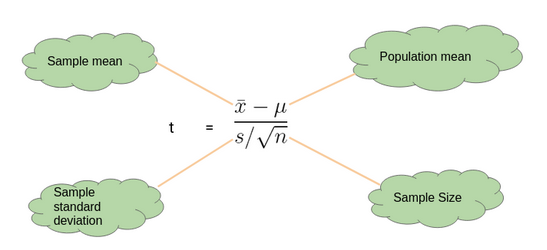
T-test:

T-tests are a statistical way of testing a hypothesis when:

* We do not know the population variance
* Our sample size is small, n < 30

One-Sample t-Test

We perform a One-Sample t-test when we want to **compare a sample mean with the population mean**. The difference from the Z Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.



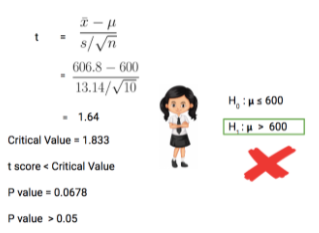
#### Example to Understand a One Sample t-Test

Let’s say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls’ scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our ⍺ value (significance level) to be 0.05 for Hypothesis Testing.

In this example:

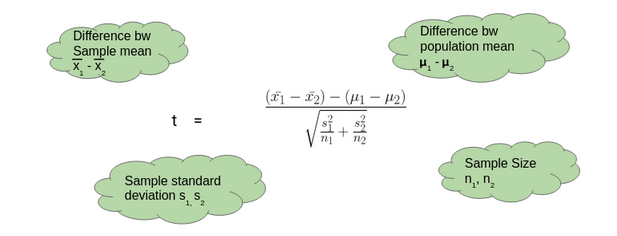
* Mean Score for Girls is 606.8
* The size of the sample is 10
* The population mean is 600
* Standard Deviation for the sample is 13.14

Chi-square Test:



Our**P-value is greater than 0.05 thus we fail to reject the null hypothesis** and don’t have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

Two-Sample t-Test: We perform a Two-Sample t-test when we want to compare the mean of two samples.

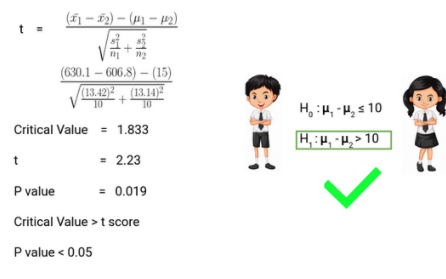


#### Example to Understand a Two-Sample t-Test

Here, let’s say we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls’ scores or boys’ scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks. We choose our ⍺ value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

In this example:

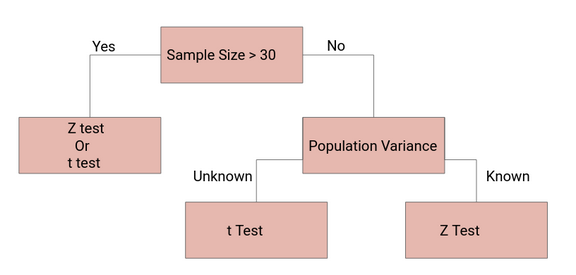
* Mean Score for Boys is 630.1
* Mean Score for Girls is 606.8
* Difference between Population Mean 15
* Standard Deviation for Boys’ score is 13.42
* Standard Deviation for Girls’ score is 13.14



Thus, **P-value is less than 0.05 so we can reject the null hypothesis** and conclude that on average boys score 15 marks more than girls in the exam.

## Deciding between Z Test and T-Test

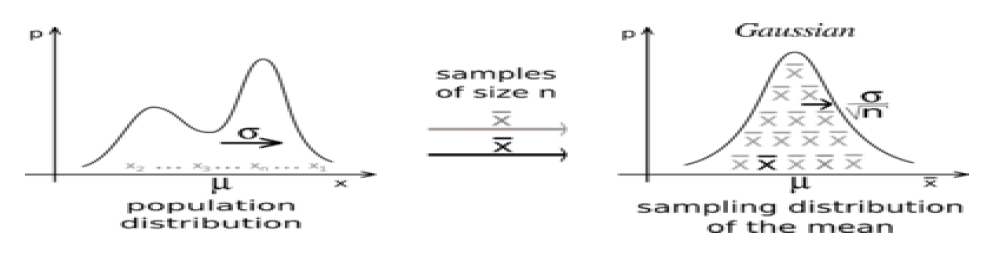
So when we should perform the Z test and when we should perform t-Test? It’s a key question we need to answer if we want to master statistics.



If the sample size is large enough, then the Z test and t-Test will conclude with the same results. For a**large sample size**, **Sample Variance will be a better estimate** of Population variance so even if population variance is unknown, we can**use the Z test using sample variance.**

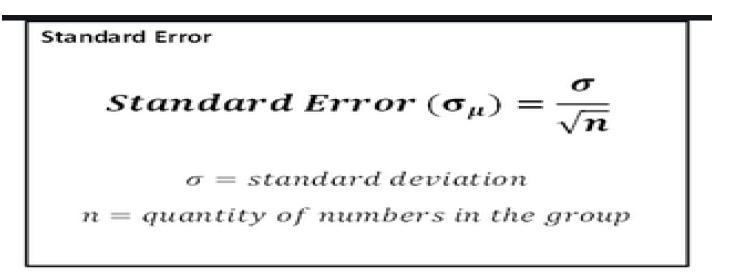
Similarly, for a**Large Sample**, we have a high degree of freedom. And since t**-distribution approaches the normal distribution**,**t**he difference between the z score and t score is negligible.

Central limit theorem: The theorem state that “We are having a large population data with some mean and standard deviation which may or may not belong to Gaussian normal distribution but when we collect some data from that population data known as sample data with some mean and standard deviation where the mean of sample data will be approximately equal to mean of population data and all the sample will follow an approximate normal distribution pattern with same standard deviation for both population and sample where the size of data is > 30.”



Standard error: The standard error is a statistical term that **measures the accuracy for a** sample distribution that represents a population by using standard deviation.

Formula for standard error:

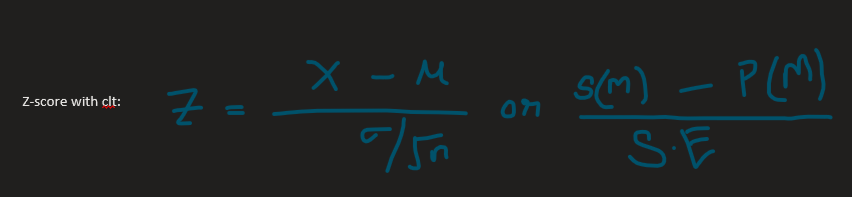


We can see that when my sample size increase the standard error decreases.

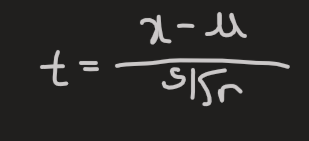
Now let’s relate the statement with central limit theorem:

**The central limit theorem states that the sample mean follows approximately the normal distribution with mean(μ) and standard deviation (σ/√n), where μ and σ are the mean and standard deviation of the population from where the sample was selected. The sample size n has to be large (usually n≥30) if the population from where the sample is taken is non normal.**

So, when we transform our sample data, we will use following formula for the z-score:



While applying central limit theorem by using t-test:



Degree of freedom: It refer to the number of independent observations in a set of observations.

Multi class confusion matrix

PREDICTED VALUE

|  |  |  |  |
| --- | --- | --- | --- |
|  | PASS | FAIL | ABSENT |
| PASS | 16 (*cell 1*) | 0 (*cell 2*) | 0 (*cell3*) |
| FAIL | 0 (*cell 4*) | 17 (*cell 5*) | 1 (*cell 6*) |
| ABSENT | 0 (*cell 7*) | 0 (*cell 8*) | 11 (*cell 9*) |

To calculate TP, TN, FN, and FP

TP: The actual value and the predicted value should be same. So considering pass the value of TP is cell 1.

FN: The sum of values of corresponding rows except the TP value

FN = (cell2 +cell3) = 0+0 = 0

FP: The sum of values of corresponding rows except the TP value

FP = (cell 4 + cell 7) = 0+0 = 0

TN: The sum of values of all columns and rows except the values that class that we are calculating the values for.

TN = (cell 5 + cell 6+ cell 8+cell 9) = 17 +1+0 +11 = 29

Example:

|  |  |
| --- | --- |
| 95 | 28 |
| 24 | 45 |

This is my 2X2 matrixs,

Let’s assign the values:

True\_positive = conf\_mat[0][0]

False\_postive = conf\_mat[0][1]

false\_negative = conf\_mat[1][0]

true\_negative = conf\_mat[1][1]

AUC and ROC curve:

This is basically used for binary classifier. Threshold play’s a very important role to decide which part of probabilities to be selected as 1 and which part to be selected as 0.

Consider the example:

|  |  |
| --- | --- |
| Y | Yhat |
| 1 | 0.8 |
| 0 | 0.9 |
| 1 | 0.4 |
| 1 | 0.3 |
| 0 | 0.2 |
| 1 | 0.7 |

Let’s consider the threshold as [0, 0.2, 0.4, 0.6, 0.8, 1]

Based on the threshold we are construction the table again

**Threshold above 0.5 is 1 and below 0.5 is 0 TPR = TP/(Tp+FN)**

**FPR = FP/(FP+TN)**

**We will compare with yhat**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Y | Yhat | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
| 1 | 0.8 | 1 | 1 |  |  |  |  |
| 0 | 0.9 | 1 | 1 |  |  |  |  |
| 1 | 0.4 | 1 | 1 |  |  |  |  |
| 1 | 0.3 | 1 | 1 |  |  |  |  |
| 0 | 0.2 | 1 | 0 |  |  |  |  |
| 1 | 0.7 | 1 | 1 |  |  |  |  |

When my threshold is 0

True positive: My actual value is 1 and predicted value is 1

True positive = 4/(4+0) = 4/4 = 1

False positive: my actual value is 0 and my predicted value is 1

True positive = 2/(2+0) = 2/2 = 1

Graph construction:

When my threshold is 0.2

True positive: My actual value is 1 and predicted value is 1

True positive = 4/(4+0) = 4/4 = 1

False positive: my actual value is 0 and my predicted value is 1

True positive = 1/(1+1) = 1/2 = 0.5

Similarly we construct it for all others and larger the area under the curve better model you have constructed this is what called ROC and we construct the straight line and the area above the