**Hypothesis Testing Exercise**

**-------------------------------------**

A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you carried out to check validity of the assumptions.

--------------------------------------------------------------------------------------------

**Sol: dataset -> Cutlets.csv**

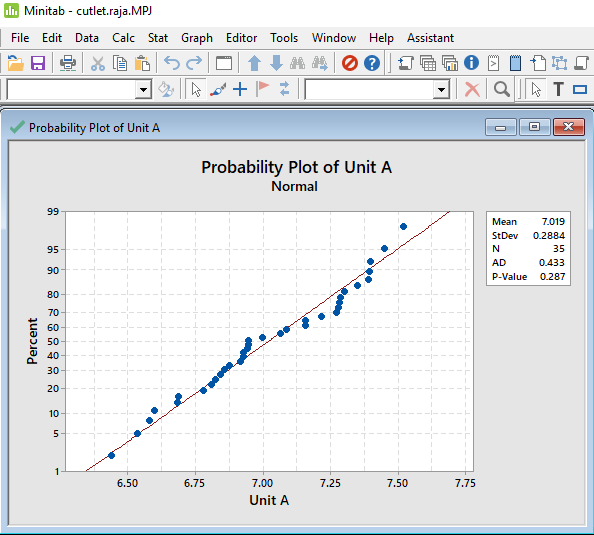
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Unit A | 6.809 | 6.4376 | 6.9157 | 7.3012 | 7.4488 | 7.3871 | 6.8755 | .. |
| Unit B | 6.7703 | 7.5093 | 6.73 | 6.7878 | 7.1522 | 6.811 | 7.2212 | .. |

We know[ Outpur(Y)-Continuous & Input(X)-Discrete in two categories] then, we should to go with : **2-Sample t test**

**Create hypothesis for Unit A:**

Ho= Data is Normally distributed

Ha=Data is not Normally distributed



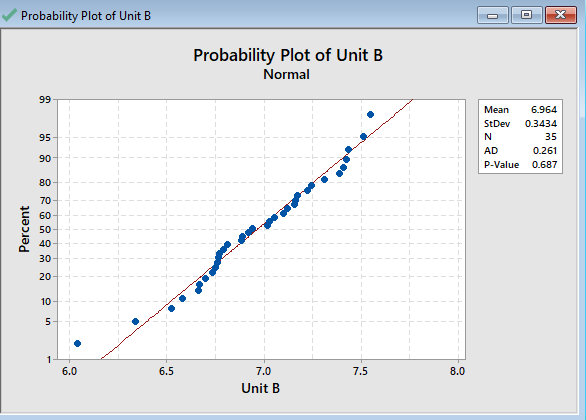
P-value is >0.05.

P High Ho Fly. So, data is Normally distributed

**Create hypothesis for Unit B:**

Ho= Data is normally distributed

Ha=Data is not normally distributed



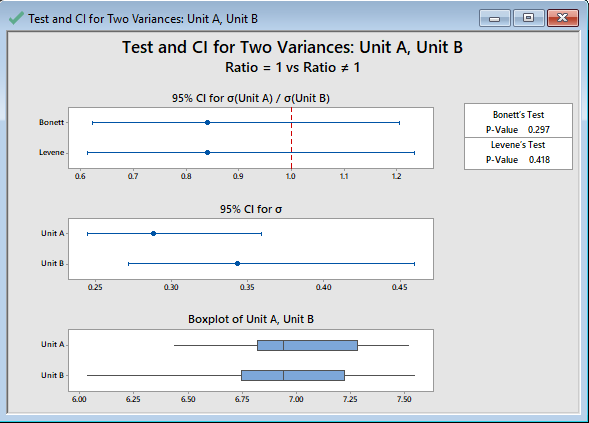
P-value is >0.05

P High Ho Fly.So, data is Normally distributed

**Create Hypothesis for variances of Unit A and Unit B**

Ho= Variance of diameters of Unit A is equal to the variance of diameters of Unit B

Ha= Variance of diameters of Unit A is not equal to the variance of diameters of Unit B



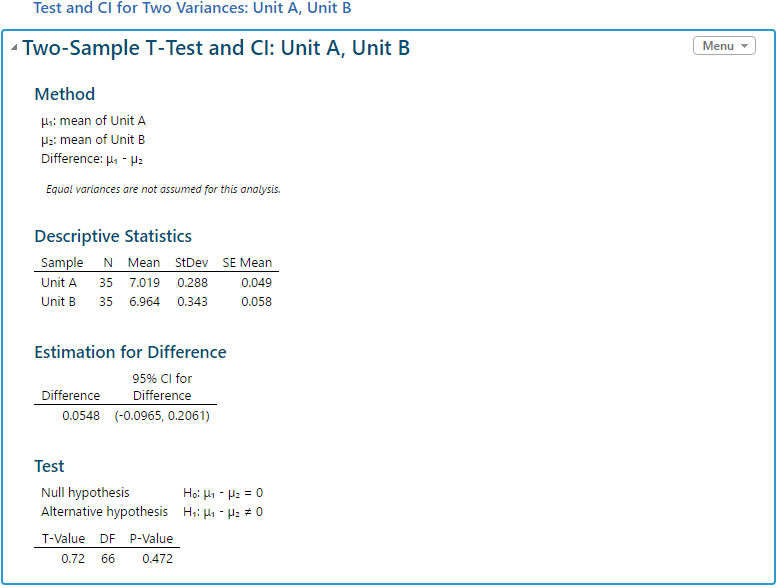
P-value>0.05.

P High Ho fly. We fail to reject Null hypothesis. So we will accept it and hence Variances of A is equal to Variances of B.

**As Unit A and Unit B are 2 Discrete variables and output variable diameter is a continuous.**

Ho= Averages of diameters of Unit A is equal to Averages of diameters of unit B

Ha= Averages of diameters of Unit A is not equal to Averages of diameters of unit B



You can see the detailed summary displayed.

P-value>0.05 and hence P High and Ho Fly.

Inference is that **there is no significant difference in the diameters of Unit A and Unit B**

A hospital wants to determine whether there is any difference in the average Turn Around Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports of 4 laboratories. TAT is defined as sample collected to report dispatch.

Analyze the data and determine whether there is any difference in average TAT among the different laboratories at 5% significance level.

**Sol: dataset -> LabTAT.csv**

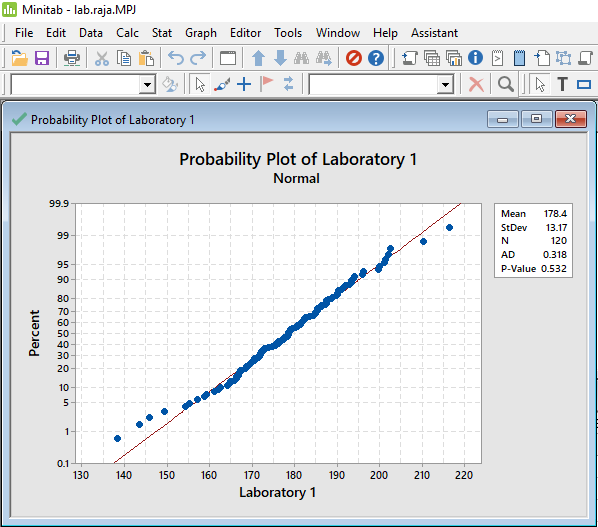
|  |  |  |  |
| --- | --- | --- | --- |
| Laboratory 1 | Laboratory 2 | Laboratory 3 | Laboratory 4 |
| 185.35 | 165.53 | 176.7 | 166.13 |
| 170.49 | 185.91 | 198.45 | 160.79 |
| 192.77 | 194.92 | 201.23 | 185.18 |
| 177.33 | 183 | 199.61 | 176.42 |
| 193.41 | 169.57 | 204.63 | 152.6 |
| 179.45 | 197 | 181.51 | 161.12 |
| 191.37 | 166.36 | 214.21 | 154.02 |
| 166.81 | 169.6 | 183.43 | 163.25 |
| 158.81 | 175.36 | 191.6 | 152.79 |
| 165.88 | 198.68 | 208.43 | 161.98 |
| 174.75 | 189.12 | 231 | 171.22 |
| 193.37 | 140.55 | 198.37 | 183.67 |
| 184.75 | 160.44 | 226.62 | 142.95 |
| 178.54 | 167.03 | 214.44 | 152.37 |
| 180.19 | 182.67 | 159.69 | 163.81 |
| 172.17 | 155.73 | 214.14 | 156.06 |
| 172 | 183.07 | 212.29 | 176.44 |
| 184.92 | 177.7 | 209.25 | 173.68 |

We know[ Outpur(Y)-Continuous & Input(X)-Discrete in more than two categories] then, we should to go with : **ANOVA-One way test**

**Create hypothesis for Lab 1**

Ho= Data is normally distributed

Ha=Data is not normally distributed

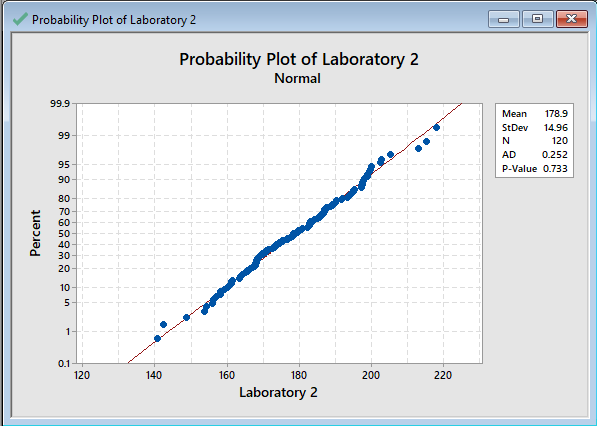


P-value is >0.05. P High Ho Fly.So data is normally distributed

**Create hypothesis for Lab 2**

Ho= Data is normally distributed

Ha=Data is not normally distributed



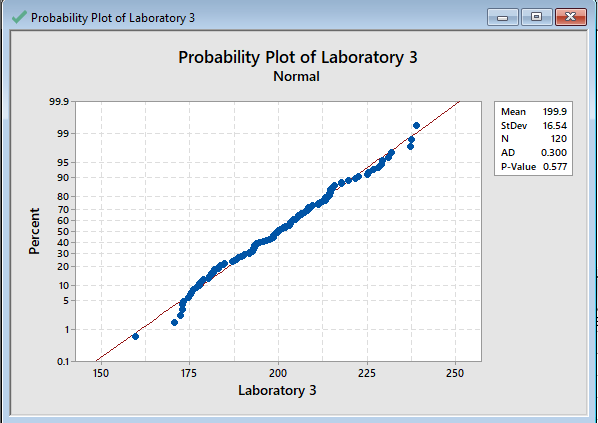
P-value is >0.05. P High Ho Fly.

So data is Normally distributed

**Create hypothesis for Lab 3**

Ho= Data is normally distributed

Ha=Data is not normally distributed



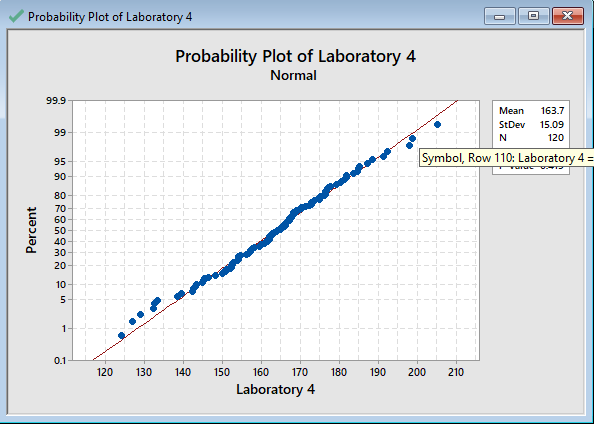
P-value is >0.05. P High Ho Fly.

So, data is normally distributed

**Create hypothesis for Lab 4**

Ho= Data is normally distributed

Ha=Data is not normally distributed



P-value is >0.05. P High Ho Fly.

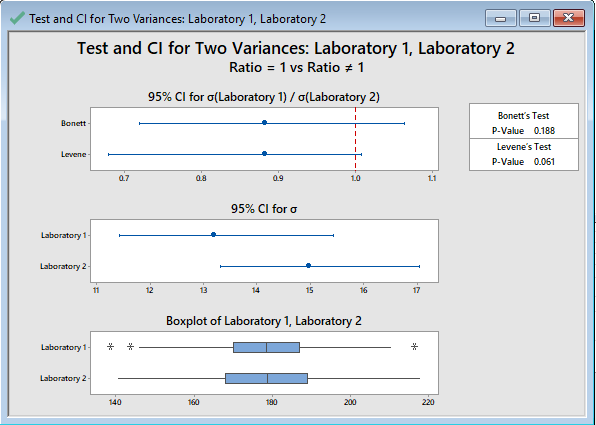
So, data is normally distributed

**Variance Test**

**Hypothesis for variances of Lab 1 and Lab 2**

Ho= Variance of TAT of Lab 1 is equal to variance of TAT of Lab 2

Ha= Variance of TAT of Lab 1 is not equal to variance of TAT of Lab 2



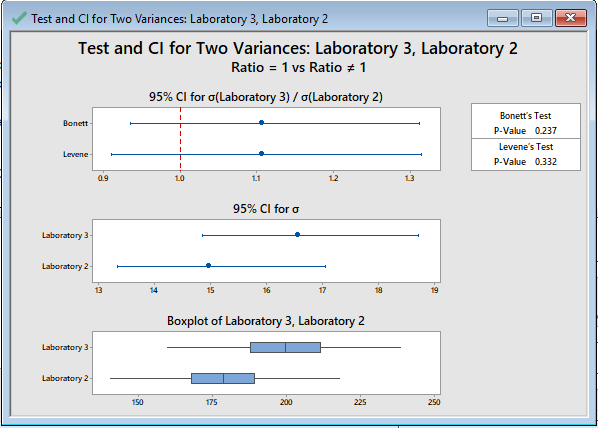
P-value>0.05. P High Ho fly. We fail to reject Null hypothesis.

So, we will accept it and hence Variances of 1 is equal to variances of 2

**Hypothesis for variances of Lab 2 and Lab 3**

Ho= Variance of TAT of Lab 2 is equal to variance of TAT of Lab 3

Ha= Variance of TAT of Lab 2 is not equal to variance of TAT of Lab 3



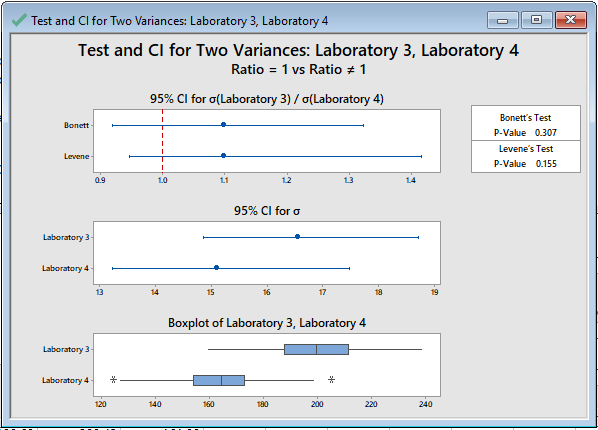
P-value >0.05. P High Ho fly. We fail to reject Null hypothesis.

So, we will accept it and hence Variances of lab 2 is equal to variances of lab 3.

**Hypothesis for variances of Lab 3 and Lab 4**

Ho= Variance of TAT of Lab 3 is equal to variance of TAT of Lab 4

Ha= Variance of TAT of Lab 3 is not equal to variance of TAT of Lab 4



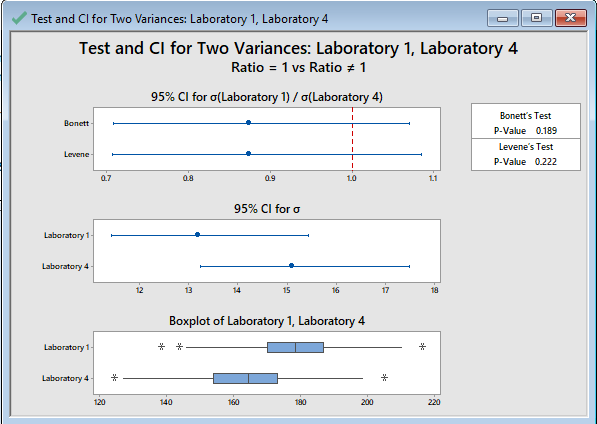
P-value>0.05. P High Ho fly. We fail to reject Null hypothesis.

So. we will accept it and hence Variances of lab 3 is equal to variances of lab 4.

**Hypothesis for variances of Lab 4 and Lab 1**

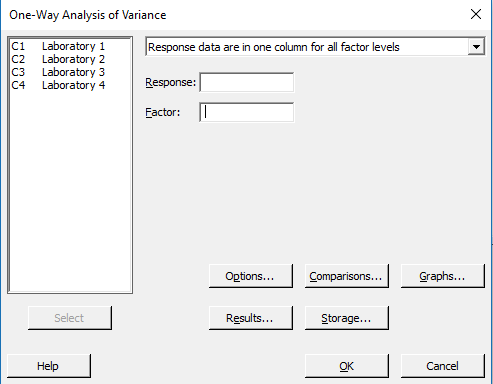
Ho= Variance of TAT of Lab 4 is equal to variance of TAT of Lab 1

Ha= Variance of TAT of Lab 4 is not equal to variance of TAT of Lab 1



P-value>0.05. P High Ho fly. We fail to reject Null hypothesis. So we will accept it and hence Variances of lab 4 is equal to variances of lab 1

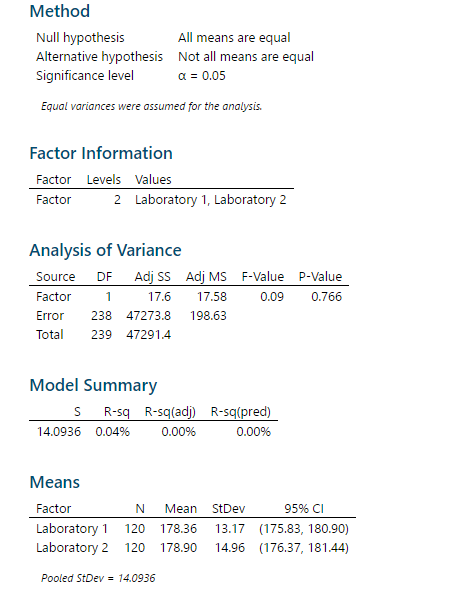
As there are more than 2 discrete variables and output variable TAT is a continuous variable. Hence we will go with **Anova one way test**.



Ho= Average TAT for all the samples is same

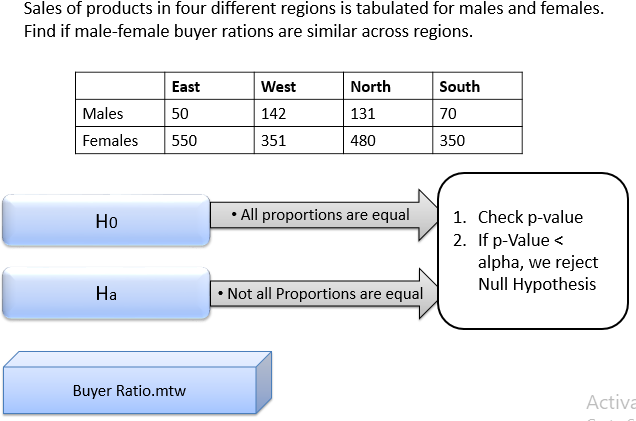
Ha= Averages TAT for all the samples is not same





SImilarly by doing for different lab combinations you can see that P -value is > 0.05. P High and Ho Fly.

Hence, **there is no significant difference in the average TAT for all the labs.**



**Sol: dataset -> BuyerRatio.csv**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observed Values | East | West | North | South |
| Males | 50 | 142 | 131 | 70 |
| Females | 435 | 1523 | 1356 | 750 |

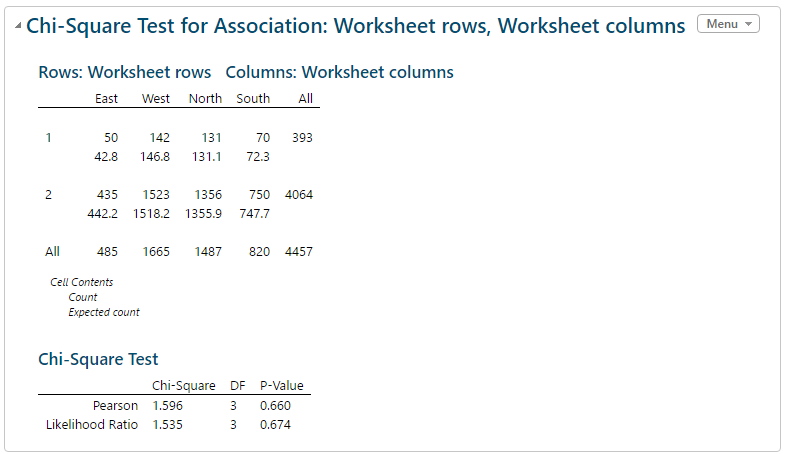
We know[ Outpur(Y)-Discreate & Input(X)-Discrete in more than two categories] then, we should to go with : **Chi-square test**

We are trying to find out if proportions of male and female are similar or not across the regions

**Create hypothesis**

Ho= Proportions of Male and Female are same

Ha= Proportions of Male and Female are not same

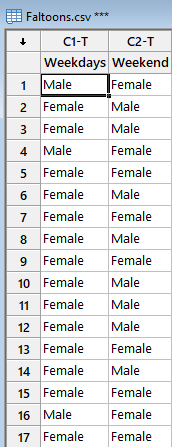


P-value>0.05.Hence we fail to reject Null.

**Hence, proportion of male and female across regions is same.**

Fantaloons Sales managers commented that *%* of males versus females walking in to the store differ based on day of the week. Analyze the data and determine whether there is evidence at *5 %* significance level to support this hypothesis.

**Sol: dataset -> Faltoons.csv**

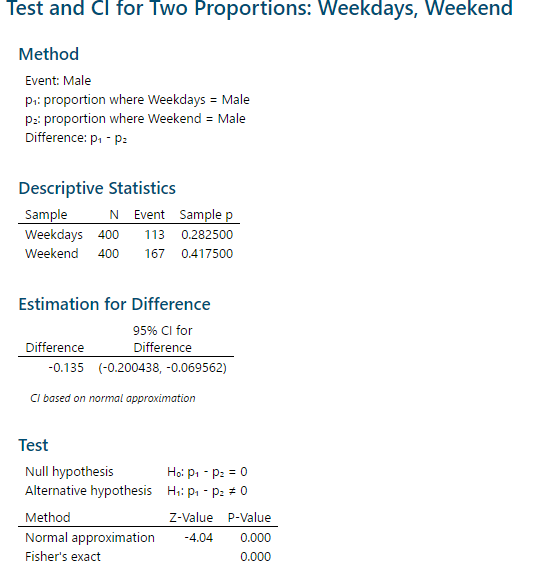


We know[ Outpur(Y)-Discreate & Input(X)-Discrete in two categories] then, we should to go with : **2-proportion test**

**Output**is **Discrete**as we are trying to find out if proportions of male and female walking in to the store is same or not.

Ho= Proportions of Male and Female are same

Ha= Proportions of Male and Female are not same

****

P-value is less than 0.05 and hence we fail to reject Null.Hence proportions of Male and Female are not same

Now we will try to find out whose proportion is higher. We create another hypothesis

Ho= Proportions of Male is less than or equal to Female

Ha= Proportions of Male is greater than Female

P-value < 0.05 and hence we reject null.**Hence, proportion of Male is greater than Female.**

TeleCall uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in order form renders it defective and has to be reworked before processing. The manager wants to check whether the defective % varies by centre. Please analyze the data at *5%* significance level and help the manager draw appropriate inferences

**Sol: dataset -> Costomer+OrderForm.csv**



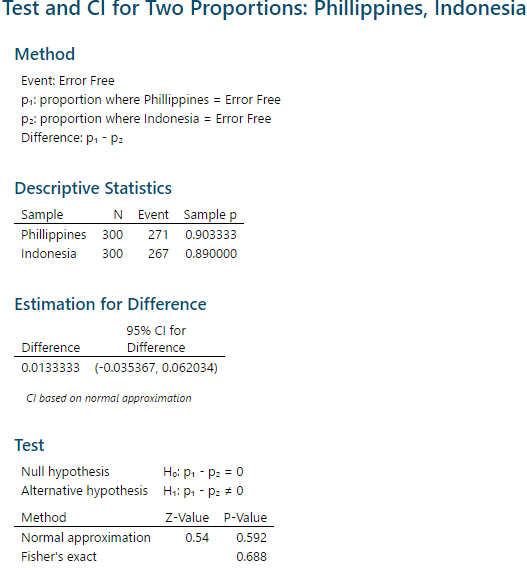
We know[ Outpur(Y)-Discreate & Input(X)-Discrete in two categories] then, we should to go with : **2-proportion test**

**Output**is **Discrete**as we are trying to find out if proportions of Error Free and Defective reworked before processing.

**Hypothesis tested between Phillippines & Indonesia**

Ho= Proportions of Error Free and Defective are same

Ha= Proportions of Error Free and Defective are not same



P-value is greater than 0.05 and hence we fail to reject Null . Hence Proportions of Error Free and Defective are same

Now we will try to find out whose proportion is higher. We create another hypothesis

Ho= Proportions of Error free is less than or equal to Defective

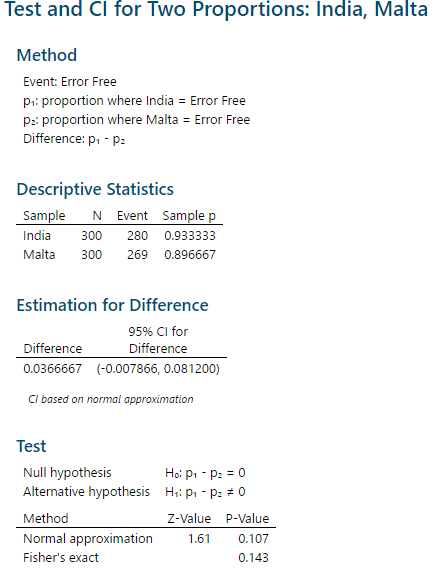
Ha= Proportions of Error free is greater than Defective

P-value > 0.05 and hence we reject null.**Hence, proportion of Error free is less than or equal to Defective.**

**Hypothesis tested between India & Malta**

Ho= Proportions of Error Free and Defective are same

Ha= Proportions of Error Free and Defective are not same



P-value is greater than 0.05 and hence we fail to reject Null . Hence Proportions of Error Free and Defective are same

Now we will try to find out whose proportion is higher. We create another hypothesis

Ho= Proportions of Error free is less than or equal to Defective

Ha= Proportions of Error free is greater than Defective

P-value > 0.05 and hence we reject null.**Hence, proportion of Error free is less than or equal to Defective.**

**Consicutively**, we can make hypothesis between Phil-Malta, Indo-India, Malta-Indonesia. Results will come as before.