Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 17
Hypothesis Testing- II

Welcome students in the last class we have seen how to formulate the hypothesis then we have seen some theory when the hypothesis should be accepted when the hypothesis should be rejected. In this class we will take some practical examples then we will solve then we will understand the concept of hypothesis in detail. So, the class objective is when the population standard deviation is known how to do the hypothesis testing.

(Refer Slide Time: 01:00)

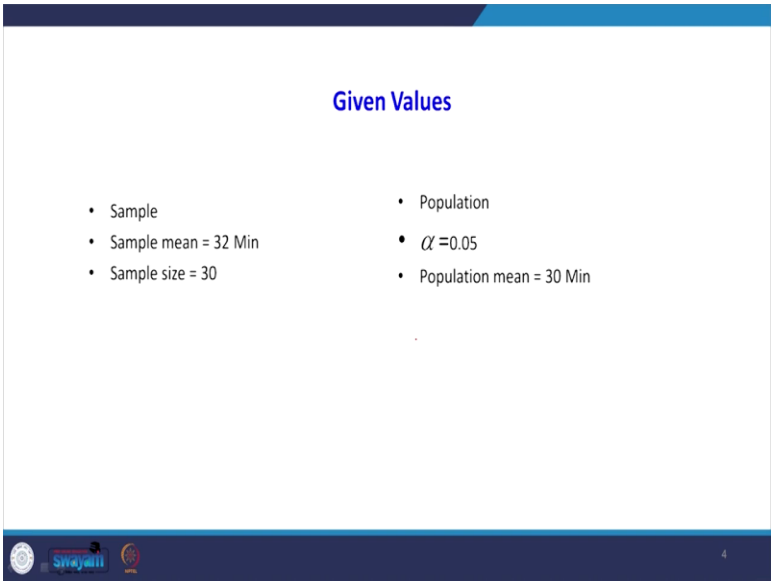
Here hypothesis testing is we are going to check the population mean will take own problem this is example of one tailed test about the population mean when Sigma is known Sigma means population standard deviation. The assumption is on population mean so the problem is the mean response times for the random sample of 30 pizza deliveries is 32 minutes. So, they conducted a sample survey in that they have sample sizes 30 out of 30 they found that the mean delivery time for Pizza is 32 minutes.

The population standard deviation is believed to be 10 minutes Sigma is known this population standard deviation this population standard deviation is nothing but your Sigma 10 minutes this

is nothing but over Sigma. The pizza delivery services director wants to perform a hypothesis test when alpha equal to 0.05 level of significance to determine whether the service goal of 30 minutes or less is being achieved.

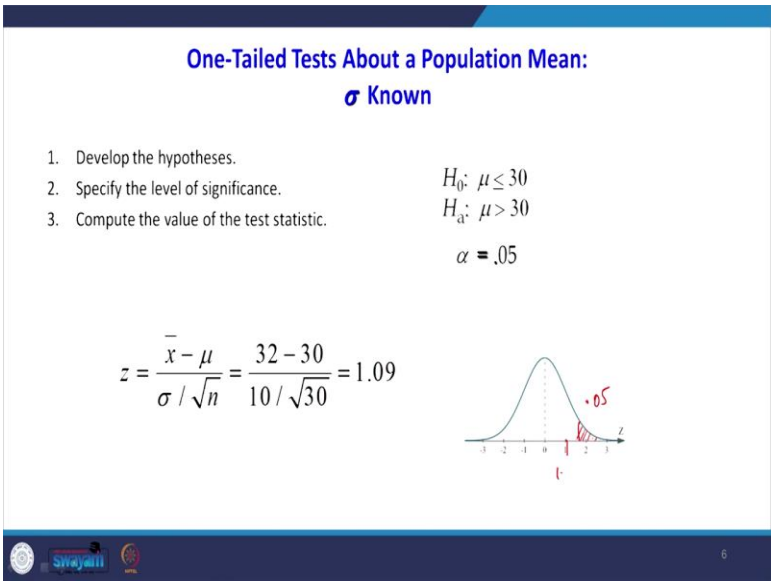
Manager of that store or that shop wanted to verify whether the survey's goal of 30 minutes or less is being achieved or not. So, now the status quo first thing is formulating the hypothesis the status quo is the Pizza is delivered within 30 minutes.

(Refer Slide Time: 02:19)



Before that we will see what are the values are given there are any hypothesis testing there is a two kind of data some data from sample some data from population. So, in the sample, sample mean is 32 minutes sample sizes n. so, this sample mean this is nothing but your x-bar this is nothing but your n, so with respect to population, population mean which we have to assume is the 30 minutes what is the population standard deviation also has to be given, what is the population standard deviation? Sigma equal to going back yeah the population standard deviation is 10 minutes.

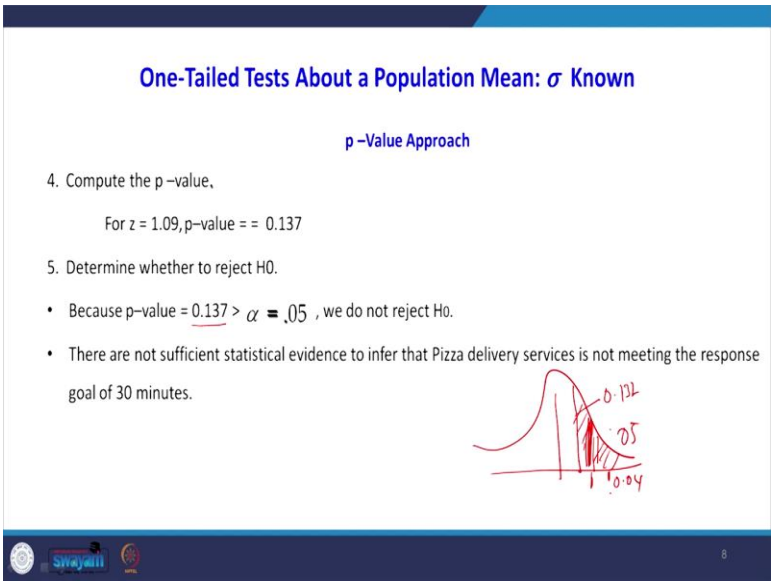
(Refer Slide Time: 03:05)



Now first we will solve this problem using p-value approach what is the step 1 develop the hypothesis develop the hypothesis in the previous class also I given some hint the status quo, status quo should go to null hypothesis. What is the status quo currently the pizza is delivered and the average of 30 minutes, so mu is less than or equal to 30. After you write the null hypothesis then you should go for alternative hypothesis the clue used that these signs are complementary when you write for null hypothesis it is a less than or equal to for alternative episodes it should be greater so greater than 30.

The step 2 is specified the level of significance alpha is given 5% now we have to decide whether it is a right tailed test or left tailed test. As I told you by looking at the sign off your alternate hypothesis it is greater than 30 so it is a right tailed test for example this is right tailed test. What is alpha it is 0.05. The next one compute the value of test statistic for the test statistic Z = (X bar -μ) /(σ/√n), X bar is given 30 to μ (mu) is as you would mean divided by 10 it is the population standard deviation/ root of n, n is your sample size, z = 1.209. So when you mark 1.09 approximately it will be here 1.09.

(Refer Slide Time: 04:40)



So, what we have to do then the calculated Z value are the test statistics is 1.09 we have to find out what is the right side area that is whatever p-value, so what is the meaning is when it is Z values 1.09 so we are to suppose this is 1.09 we have to mark this side area that side area is nothing but p value p value. So, with the help of Python when you go for 1 minus stat.norm.cdf 1.09 because cdf is we are finding minus infinity to Z value when you want the right side area that has to be divided by 1 that is nothing but if I draw here one more time in Python we can get area when Z values 1.09.

For example approximately here we have to find out what is the right side the area. So, for finding the right side area if you put stat.norm.cdf of 1.09 because the Python is giving area from minus infinity to here Z value. So, you will get the left side area but we want to know the rights area. So, since you know the area is 1, so 1 minus this left side area will give you the right side area. So, the right side area this one is 0.137 you have to compute the p-value for Z equal to 1.09 the p-value is 0.137.

Now we have to determine whether to reject H0 or not what has happened since the p-value 0.137 is greater than alpha value because alpha is 0.05 so the p-value is greater than alpha value so we do not reject null hypothesis. So, what is the meaning is, again I am drawing one more time even though there are many places and drawing normal distribution that will be for our

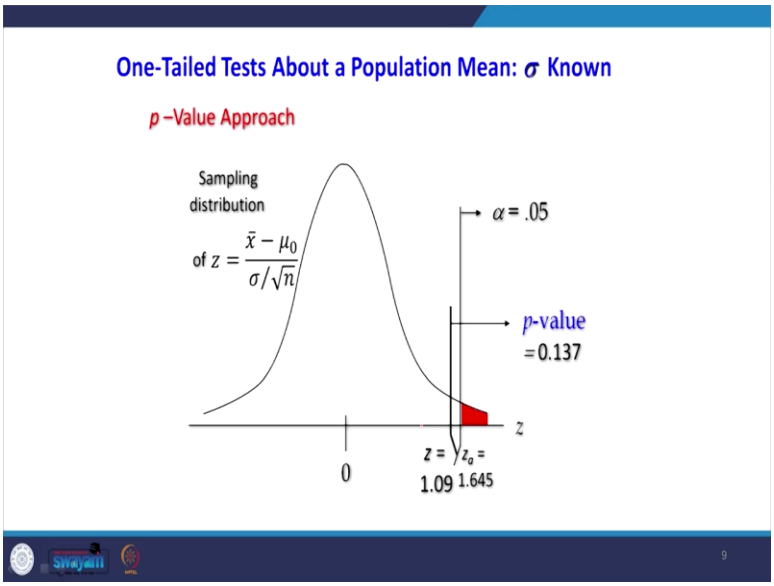
understanding purpose, this is 0.05. So, whatever value which is on the right side of this 0.05 will be rejected what happened the area which I found this area is 0.137.

So, this area is 0.137 so now what happened now I have entered into the acceptance region. So, I have to accept the null hypothesis. In case with the p-value is 0.04 so I will be standing here because this is up to this I am writing this is 0.05 when you say 0.04 I will be standing here so that means I am standing in the rejection region I have to reject it. Now what has happened that we are crossing that boundary of 0.05 so we have entered into the acceptance region.

So we have to accept the null hypothesis but generally we would not say accept do not reject null hypothesis. So, what is the conclusion there are not sufficient statistical evidence to infer that Pizza delivery service is not meeting the response goal of 30 minutes. So, when you say accept null hypothesis so what we say that the μ is less than or equal to 30 minutes that means the Pizza is delivered before 30 minutes.

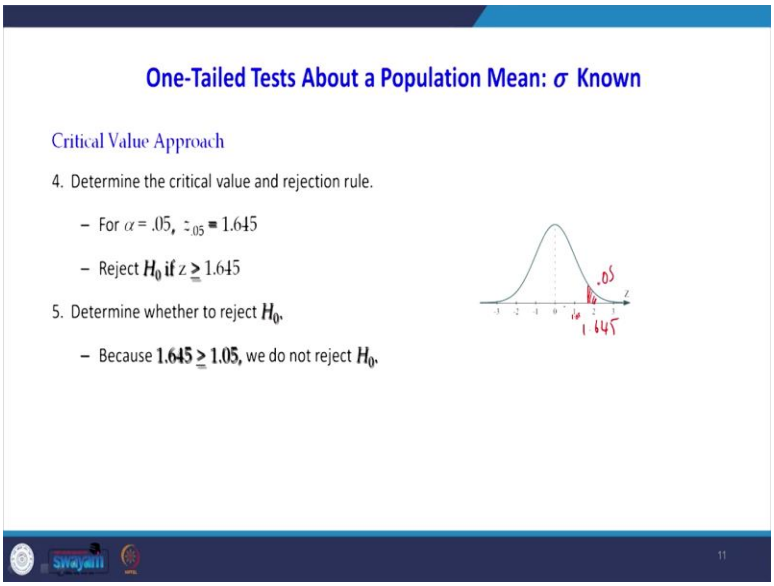
You know that offer is there if it is not delivered within 30 minutes the Pizza is free for you. So, they make sure that all deliveries are delivered within 30 minutes.

(Refer Slide Time: 08:12)



The same example you see when Z equal to 1.09 corresponding p value is 0.137 but the Alpha is 0.05 the red region represents the rejection region. So, what has happened we cross into the acceptance region so we have to accept the null hypothesis.

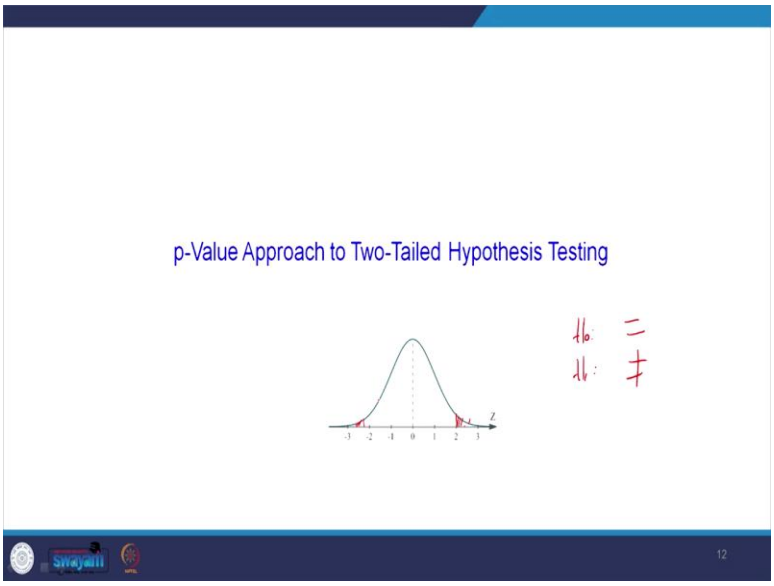
(Refer Slide Time: 08:32)



Then the same problem will do with the help of critical value approach in both approach we have to get the, we have to get the same answer. So, we will continue from the step 4 first determine the critical value and rejection rule. So, what is the rejection rule is when alpha equal to 0.05 we have to find out what is the Z value Z value is 1.645. Then you see that our calculated Z value is one 0.05 that is nothing but our test statistic. So, test statics will statistic will be 1.05 here 1.05.

So, what is the logic if the calculated Z value are the test statistics is, if it is falling on the rejection reason we have to reject it but here it is falling on the acceptance region so we have to accept the null hypothesis.

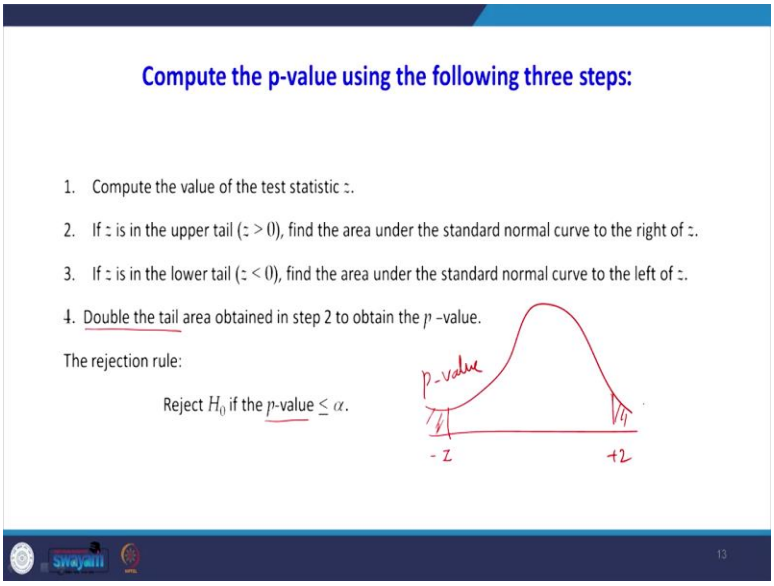
(Refer Slide Time: 09:31)



The previous example was the p value for one tailed test now we will see how to do hypothesis testing for a two-tailed test as I told you how to know it is the two-tailed test in alternative hypothesis if the sign is for example H0 is this one H0 if the sign is not equal to then it is a two-tailed test. Generally two tailed test what will happen there will be an upper limit there will be a lower limit.

If any values any test statistics if it is false on the above this upper limit of the acceptance region we will reject it for this falls below the below the acceptance region we will reject it.

(Refer Slide Time: 10:16)

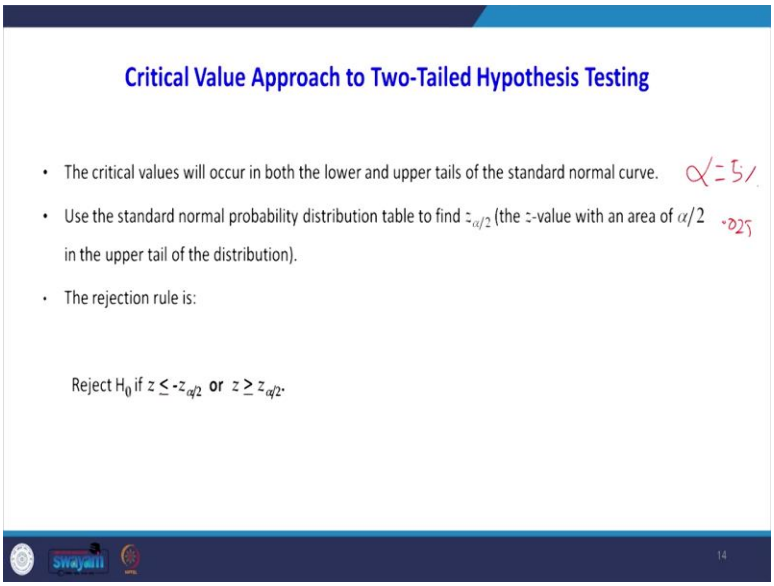


Now computing the p-value using the following 3 steps one is compute the value of test statistic Z if Z is an upper tail find the area under the standard normal curve to the right of the Z. If the Z is a lower tail that is if the less is less than 0 find the area under the standard normal curve to the left of Z. So, double the tail area obtained by in step two that is a logical why what we are going to do, since it is a two tailed test.

So, whatever area which were found left side or right side that has to be doubled to obtain the p-value. The rejection rule is if the double devalue double the p-value is less than or equal to alpha reject it otherwise accept it so what is the what it say is that you go this way for a test statistics, for the test statistics for example Z you find what is the area you multiply this left side area this is a p-value multiplied by 2 times because it is a two-tailed test.

And not only that it is a symmetric if after multiplying if the p-value is still less than or equal to alpha we have to reject it. Otherwise what you can do instead of multiplying you when it is a minus Z you find out what is the p-value when it is the plus Z what is the find of p-value you add it the added p-value should be less than or equal to; if it is less than or equal to alpha we have to reject it.

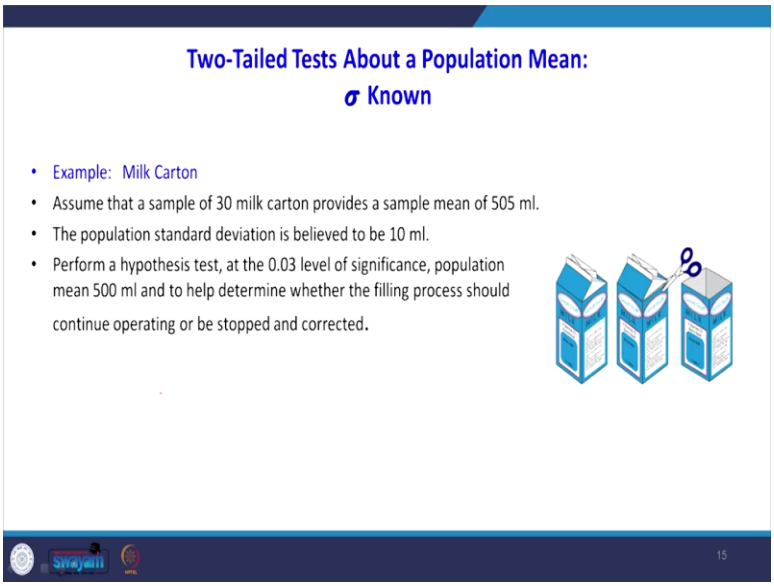
(Refer Slide Time: 11:53)



The critical value will occur in both lower and upper tail of the standard normal curve use the standard normal probability distribution table to find out $Z_{\alpha/2}$. Why we are doing $Z_{\alpha/2}$ because a

2 tail if alpha equal to 5% for example so alpha/2 is 2.5% it is 0.025 so when alpha by 2 is 0.025 we have to find out corresponding Z value on left side and right side. So, the rejection rule is if Z is less than the lower limit reject it or if Z is above the upper limit reject it.

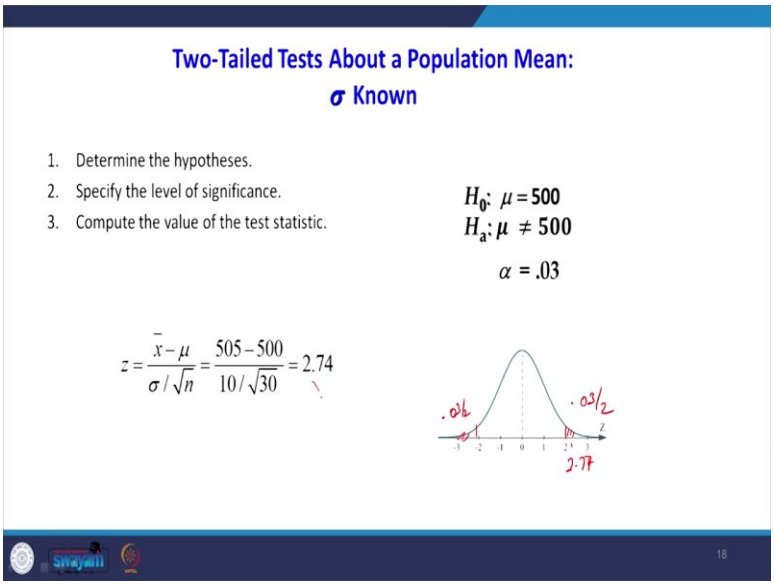
(Refer Slide Time: 12:45)



Here the Z means sample statistic we will do an example for that his example is for doing hypothesis testing for the two-tailed test when Sigma is known. The example is a milk carton assume that a sample of 30 milk carton provides a sample mean of 505 ml, the population standard deviation is believed to be 10 ml, perform a hypothesis test are at 0.03 level of significance when the population mean is 500 ml. To help to determine whether the filling process should be continued to operating or it has to be stopped and corrected.

So, what is happening there is a assume that it is assembly line so the in the assembly line that the bottles are filled with 500 ml, what is happening generally if it is over filling also there is a problem if it is under filling also there is a problem that is why if it is H0 mu equal to 500 ml, Ha: mu $\neq$ 500 ml. The logic why we did not go for left tail or right tail test is because we have to go for not equal to because even over filling and under filling is the problem for us that is why we should go for two-tailed test.

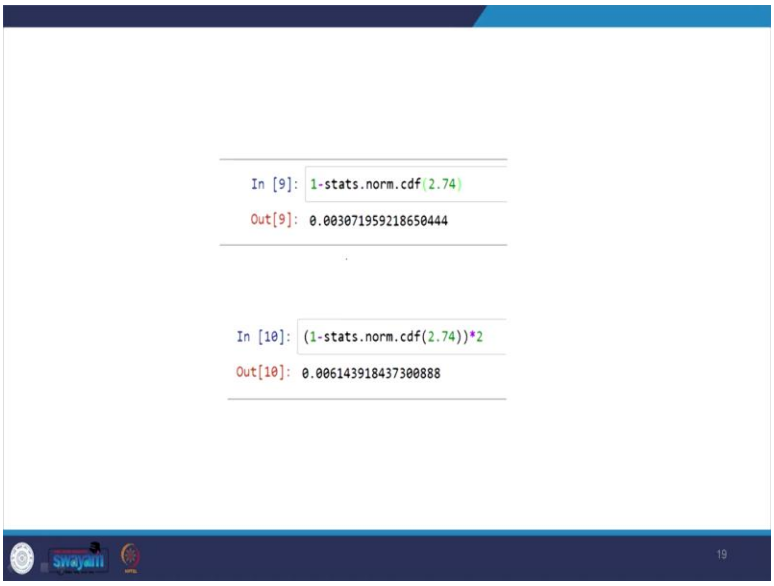
(Refer Slide Time: 14:08)



So, what are the data is given here as usual data will be given for sample and population n equal to 30 this sample mean is file 505 ml with respect to population what kind of data is given we are assuming that mu equal to 500 and standard deviation Sigma equal to 10 ml and alpha equal to 0.03. This problem that is a two tailed problem will solve with the help of p-value approach. First you have to determine the hypothesis.

You see that mu equal to 500 why this as I told you because both overfilling and under filling will cause the problem for the company. So, the hypothesis is formulated specify the level of significance alpha it is given in the problem it is 0.03% it is a two-tailed test, so I have to mark this side 0.03 by 2 left side also it is 0.03 by 2. Next I have to compute Z statistic, Z statistic is (X bar -μ) divided by (σ/√n), X bar is 505 mu is assumed mean 500 divided by 10 is given root of 30, so 2.74 so you have to mark this 2.74.

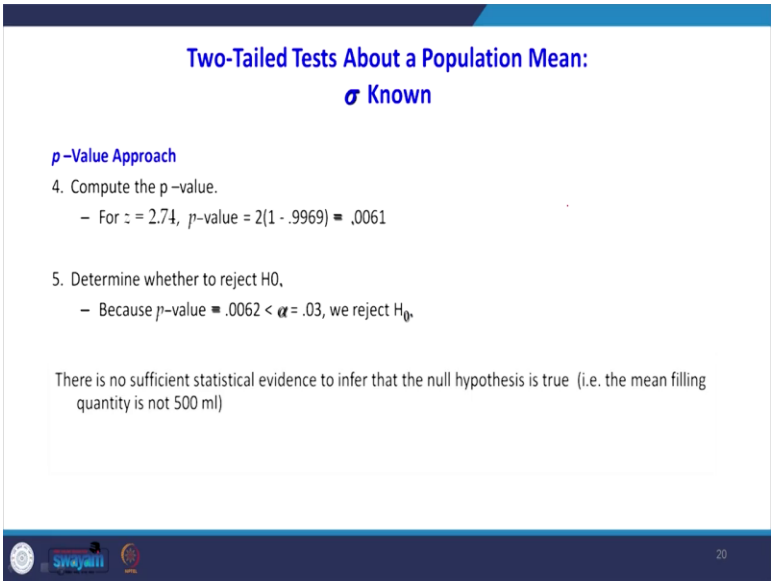
(Refer Slide Time: 15:50)



For example assume that the 2.74 is here so what I have to do when the Z value is 2.74 we have to find out the area towards the right in Python if you type this 1 – stat.norm.cdf of 2.74 the right side area is 0.003. If you multiply this both side because multiply two times because it is symmetric. So, this what this meaning is in Python it is this way so when Z equal to 2.74 corresponding right side the area is point 0.00307.

This side also when Z equal to minus 2.74 the area is 0.00307 you mu add both you will get 0.0061.

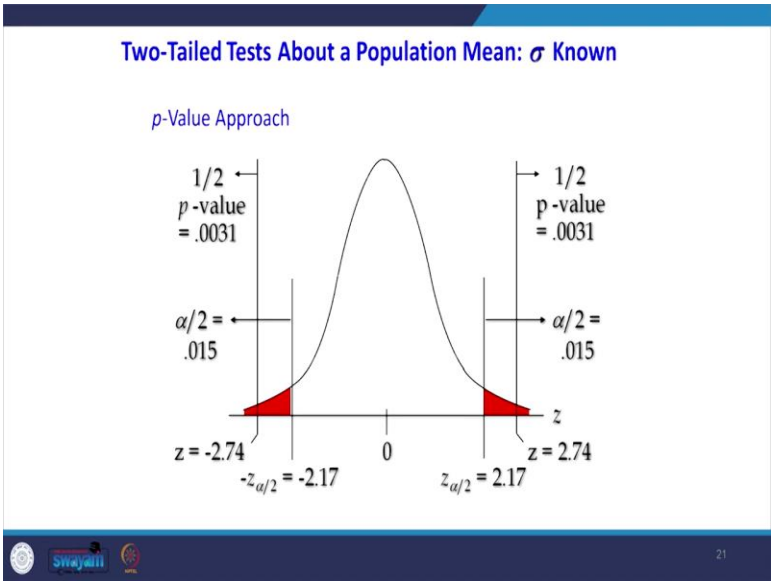
(Refer Slide Time: 16:42)



So what is happening that 0.0061 is less than your 0.03 see this 0.0061 this was 0.0062 by after approximation the Alpha is 0.0 there is still it is that less than the alpha value so we have to reject H0. When we reject H0 there is no sufficient statistical evidence to infer that the alternative hypothesis that means the mean filling quantity we are rejecting null hypothesis so when we reject null hypothesis what was our null hypothesis $\mu = 500$, H1: $\mu \neq 500$, when you reject it there is no sufficient statistical evidence to infer that the alternative hypothesis is true.

So we have found that the p-value is 0.002 that is less than alpha 0.03, so we have to reject null hypothesis, when you reject a null hypothesis we are accepting our alternative hypothesis that there is no sufficient statistical evidence to infer that the null hypothesis is true. So, the mean filling quantity is not 500ml so immediately they have to stop the assembly line they had to make the corrective action that is the inference.

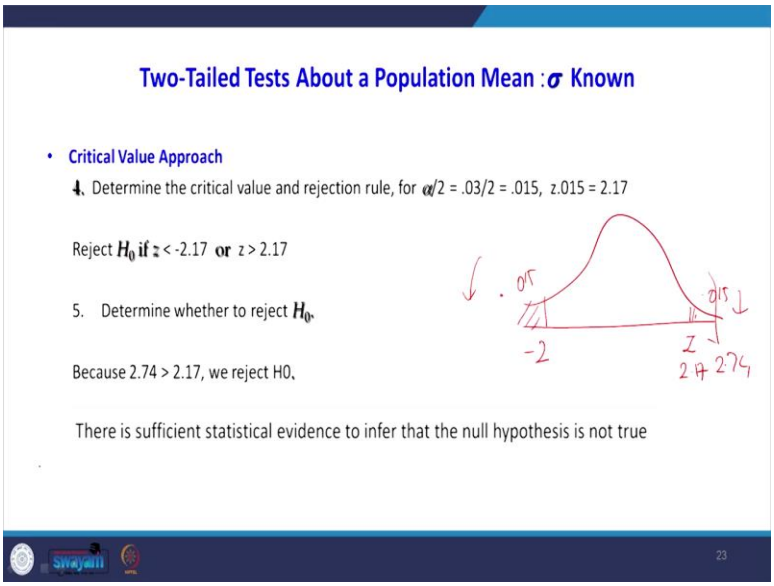
(Refer Slide Time: 17:57)



Yes that was shown here in the picture form Z equal to 2.74 is the test statistics. So, the right side area is 0.0031 when test two statistics is -2.74 the left's idea is 0.003,1 after adding that still it is less than or equal to alpha we have to reject it otherwise we can compare this 0.0031 versus 0.15 the p-value this is half of the significant value, the half of the p-value is 0.03 that is lesser than the 0.015 so we can reject it. But many software packages you may not give this half of the p-value and half of the Alpha value.

You will get the added value that means this 0.0031 is added with another 0.0031 then this alpha on the 0.015 is added with another 0.015 so the added p-value is compared with added alpha value then we take the decision if the p-value is less than alpha we are to reject it.

(Refer Slide Time: 19:08)

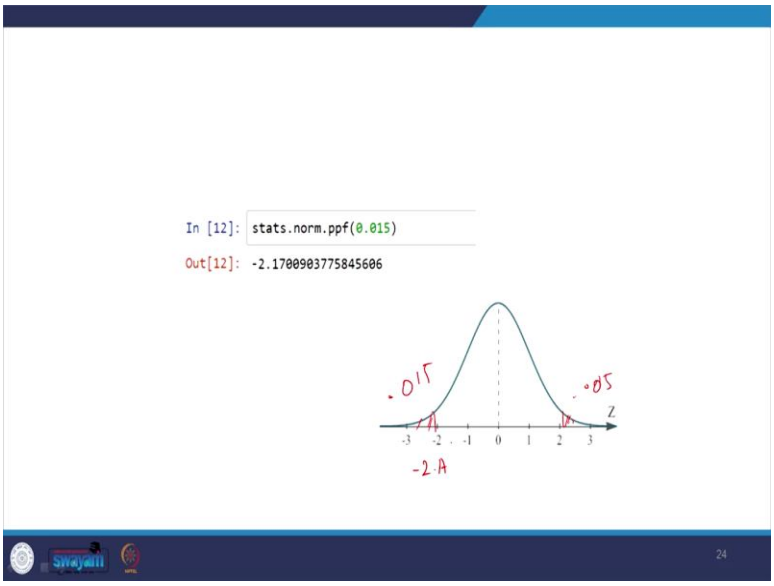


So, here we are rejecting I will go for critical value approach the critical value approach will continue from the fourth step determine the critical value and the rejection rule for alpha by 2 0.015 so what is the meaning is when alpha is 0.015 we have to find out this critical value for the right side when this side area is 0.015 we were to find out - Z critical value. So, if the calculated Z value is lying on right side we are projected to what is lying on the left side we have to reject.

Because what happened the 2.74 is our calculated Z value otherwise test statistic sample statistic this value is 2.17, so the 2.74 will be on this side 2.74 will be on the rejection side. So, we have to reject it so there is a sufficient statistical evidence to infer that the alternative hypothesis is not true. Now test statistics 2.74 lying on the rejection side we have to reject our H0, so the conclusion is there is sufficient statistical evidence to infer that the null hypothesis not true.

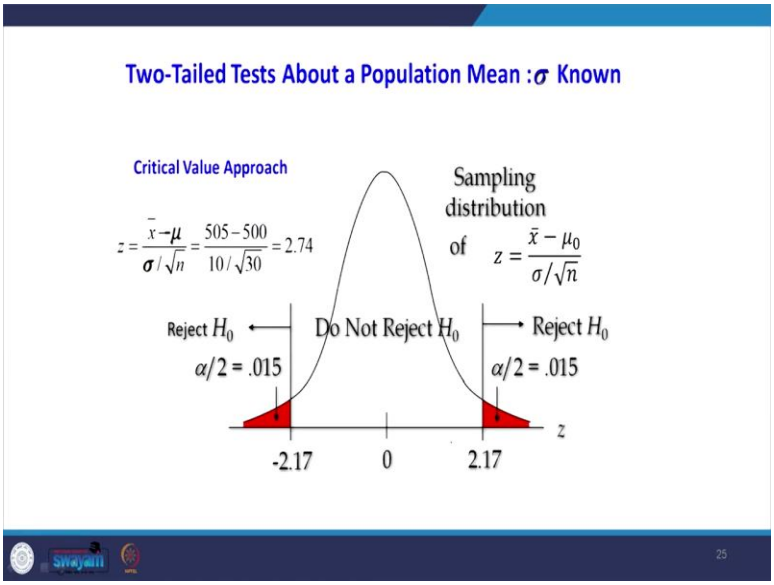
So we have to accept our alternative hypothesis that means the assembly process is not filling average value of 500 ml. So, we have to stop that assembly line then we have to make corrective actions.

(Refer Slide Time: 20:50)



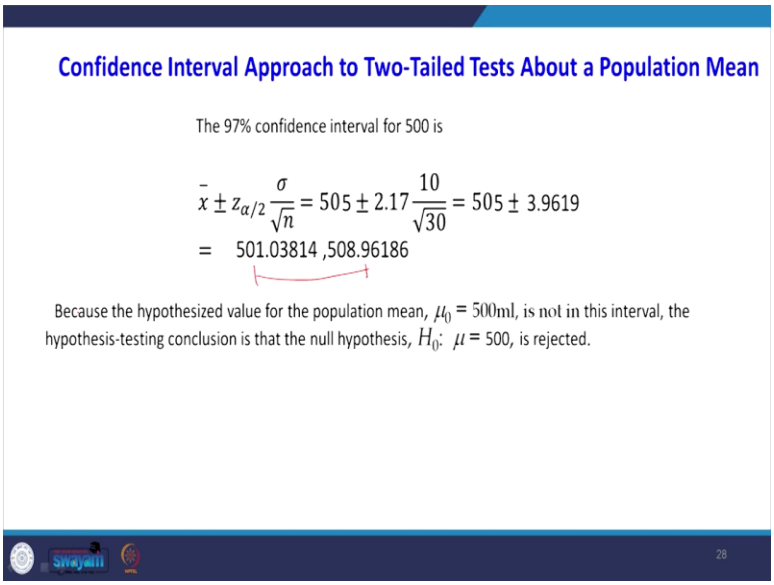
So, what is the step here as I told you when this value is alpha by 2 that is a 0.015, so the corresponding, this is a positive side the left side 0.015 so the left side -2.17 how can you get this one when you type stats.norm. cdf of in Python when you put 0.015 you will get lower limit of our critical value this is symmetric. So, right side, it is also will be same value.

(Refer Slide Time: 21:26)



This also same thing what has happened when alpha by 2 is 0.015 the lower limit is -2.17 it is 0.15 on the right hand side the upper limit is 2.17 this Z value that is we calculated 2.74 so this 2.74 will be this side 2.74 will be on the rejection side you have to reject it. In case for example the Z value, say 2 for example 1.5 say 1.5 will be here so we have to accept it.

(Refer Slide Time: 22:02)



We will solve the same problem with the help of confidence interval approach confidence interval approach for two tailed test about the population mean. So, select the simple random sample from the population and use the value of the sample mean X bar here X bar to develop your confidence interval for the population mean mu (μ), if the confidence interval contains hypothesis value of 500 do not reject it.

So how we are going to develop this conference interval is we know this very familiar formula (xbar - μ)/(σ/√n) when you readjust this so we can find out in terms of x-bar the upper limit of mu and lower limit of so is, μ + Z (σ/√n) will be the upper limit when you put μ - Z (σ/√n) will be the lower limit. So, what we are do with the help of x-bar we have to express the upper limit lower limit of mu so that formula has come from this Z = (X bar - μ)/(σ/√n).

So the upper limit of mu is X bar + Z Sigma by root n lower limit will be X bar - Z Sigma by root n this value by adjusting this Z equations we got this one in this interval. Suppose we have to find out the upper limit say this is lower limit this is upper limit, in this interval if the 500 is the assumed mean is lying we have to accept null hypothesis. Otherwise reject it, actually H0 should be rejected if mu happens to be equal to one of the endpoint of the conference interval.

Now you see that the formula which have explained previous slide x-bar ± $Z_{\alpha/2}$ (σ/√n), and X bar is 505 +, $Z_{\alpha/2}$ is 2.17, so Sigma and n so sample size is 30 so this is 505 ± 3.9619 so the lower

limit is this is lower limit this is upper limit. So, in this interval we are not able to capture 500, so we have to reject null hypothesis, so to see this because the hypothesis value for the population mean mu 500ml is not in this interval so not in this interval the hypothesis testing conclusion is that the null hypothesis H0 mu equal to 500 as rejected.

Dear students what you have seen in this lecture so far we have taken one practical problem for the pizza delivery problem. We have learned how to test one tailed test that is left tail test then we have learnt how to do the two tail test. In one tail test first we you solved with help of p-value approach then we solved with the help of critical value approach. In two tail test also first you solved with the help of p-value then critical value.

Then the third one which ever solved using confidence interval method in all these 3 method the final result is same that we have rejected our null hypothesis. In the next class will start with the t-test so what will happen in t-test so far we the population standard deviation is given there may be a situation where you may not know the population standard deviation that time you should go for t-test that will continue in the next class.