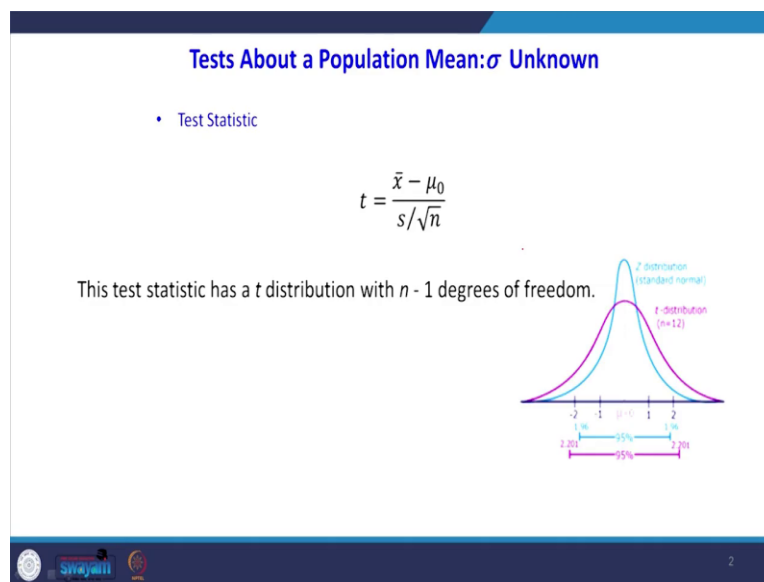


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 18
Hypothesis Testing- III

Welcome students in the previous lecture we have seen hypothesis testing when Sigma is known so that test we call it as Z test. Now we will go to be another category of hypothesis testing procedures where Sigma is not known. Most of the time the population standard deviation is not known to us so that time we should go for another test that is called t test.

(Refer Slide Time: 00:49)



I will explain what is the connection between the Z test and t test? So, previously you remember that we have used the $Z = (\bar{X} - \mu) / (\sigma/\sqrt{n})$. So, in this case since Sigma is not known to us instead of Sigma that Sigma is replaced by s that is a sample standard deviation. The another relation between Z and t is when the n is increasing look at this picture here the blue one is Z distribution the pink one is t distribution.

So, when n is increasing when the sample size is increasing the behavior of Z test and t test is same that is why in many software packages there would not be separate tab for doing Z test there will be a tab only for doing t test for example hypothesis testing and in SPSS. When you go

for even in Minitab also when you go for that there would not be a column to do the Z test but there will be a column for t-test.

For doing Z test the t test is enough, so what is the t, $t = (\bar{X} - \mu) / (s / \sqrt{n})$. here the degrees of freedom will come into place because the shape of the t distribution it is affected by the degrees of freedom when the degrees of freedom is increasing so the behavior is same. So, the test statistic has a t distribution with $n - 1$ degrees of freedom.

(Refer Slide Time: 02:22)

Tests About a Population Mean: σ Unknown

Rejection Rule: p-Value Approach
 Reject H_0 if p-value $\leq \alpha$

Rejection Rule: Critical Value Approach

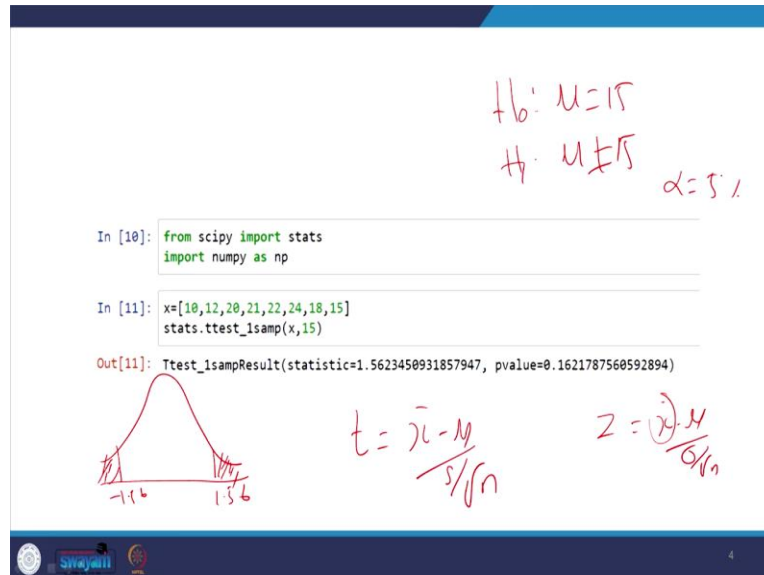
$H_0: \mu \geq \mu_0$	Reject H_0 if $t \leq -t_{\alpha}$
$H_0: \mu \leq \mu_0$	Reject H_0 if $t \geq t_{\alpha}$
$H_0: \mu = \mu_0$	Reject H_0 if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

What is the rejection rule the same whatever we have seen for the previous lecture that is Z test that rule is applicable for here also so reject H_0 if the p-value $\leq \alpha$, if it is a critical value approach suppose if $\mu \geq \mu_0$ so alternative hypothesis will $\mu < \mu_0$ so it is a left tailed test. So, in the left tailed test, test statistic t is less than your $-t_{\alpha}$, you have to reject it.

For example see here this is $\mu \leq \mu_0$ so the signs are complementary so this right tailed test so what is this left tailed test. So, the left tailed test will be like this so this is t_{α} , so if the calculated t value is lying on this side we are to reject it what does this right tail test I am writing this side so if it is this way this is right if the t is like this is t_{α} the t, calculated t is lying on the right side we have to reject it if $\mu = \mu_0$ what will be the alternate hypothesis here $H_1: \mu \neq \mu_0$.

So, this will be this way there will be $t_{\alpha/2}$ on the right side $-t_{\alpha/2}$ on the left side if the t value calculated t value is lying on any of this side we have to reject our null hypothesis.

(Refer Slide Time: 03:55)



For doing t test in Python we have to import stats, from scipy import stats then you have to import numpy else import numpy as np, now X equal to 10 12, X is an array 10 12 20 21 22 24 18 and 15 the function for doing t test is start stat t test underscore one sample here the X represent this array the mu represents our as you would mean. So, for this problem here the null hypothesis is $H_0: \mu = 15$, $H_1: \mu \neq 15$.

So when you run this you are getting this test statistics what is this value does $t = (\bar{X} - \mu)$ divided by (s/\sqrt{n}) . So, what has happened the Python calculated the value of \bar{x} from this given array and the value of s sample standard deviation with the help of \bar{X} μ this is 15 n is the sample size it is taken care. So, what is happening the previously what you are then when you are doing Z test we are using $(\bar{X} - \mu)$ divided by (σ/\sqrt{n}) the \bar{X} has to be formed with from the sample.

But here you need not do that one just you mention that array name the built-in function will take care. So, this p value what we are getting is the two-sided p-value, two-sided p-value me in the sense suppose if it is a two-tailed test, so this side when t value what is the t value when t is 1.56 so plus 1.56, - 1.56, so the total area that is the right side area and this left side area that the area


is 0.16 assume that my alpha equal to say 5% what is happening the p value is exceeding the 5%. So we have to accept our null hypothesis this is the way to do the t test in Python.

(Refer Slide Time: 06:08)

One-Tailed Test About a Population Mean: σ Unknown

Example: Ice Cream Demand

- In a ice cream parlor at IIT Roorkee, the following data represent the number of ice-creams sold in 20 days
- Test hypothesis $H_0: \mu \leq 10$
- Use $\alpha = .05$ to test the hypothesis.

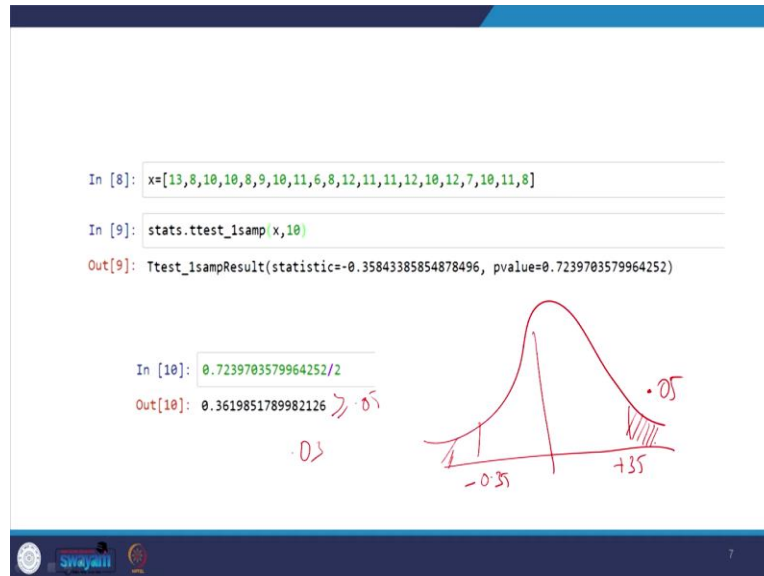


Day	No. of Ice-cream Sold	Day	No. of Ice-cream Sold
1	13	11	12
2	8	12	11
3	10	13	11
4	10	14	12
5	8	15	10
6	9	16	12
7	10	17	7
8	11	18	10
9	6	19	11
10	8	20	8

It will take an example for this in an ice cream parlor at IIT Roorkee the following data represents the number of ice cream sold in 20 days. So, here n equal to 20, the 20 days data were surveyed the shop owner want to test that the average is less than 10 by taking alpha equal to 5% so what is happening there are 20 data set is there the number of ice cream sold on day 1 is 13 day 2 is 8 and so on so for day 20, 80.

So, if you are doing manually that means with the help of statistical table we have to find out the sample you have to find out the X bar then we have to find out the sample standard deviation then you have to use this formula $(\bar{X} - \mu) / (s/\sqrt{n})$.

(Refer Slide Time: 06:58)

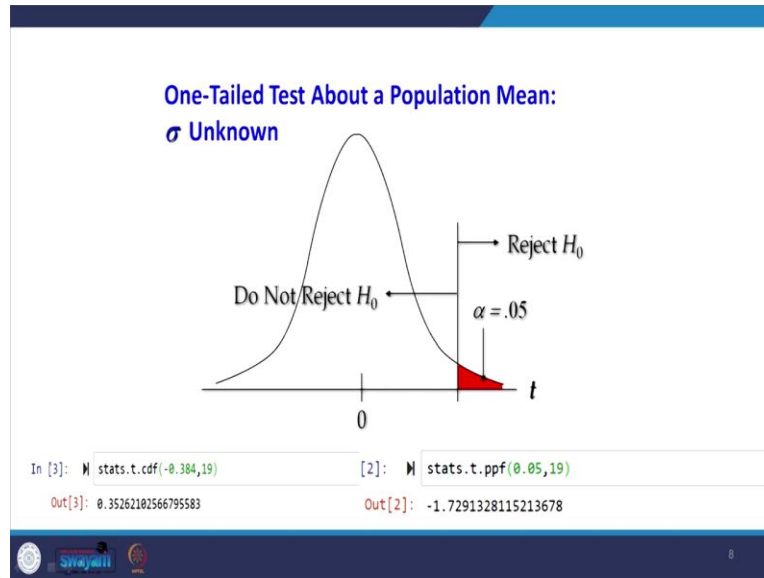


But in Python it is helping us very easily so we will go back, so what is happening the H_0 : is μ less than or equal to 10 so $H_0: \mu \leq 10$. So, what will be over alternative hypothesis: μ greater than 10, so it is a right tailed test. So, right tailed test I have to shade to this side okay what is alpha is 0.05 first you will saw 0.05 so I have stored this given value into an object X so X is 13 8 10 all the values which I was shown in the previous table that I have stored in this one. Then stats dot t-test underscore 1 samp in bracket X ,10 the here the 10 is our assumed mean assumed population mean.

This X is this array, so what we are getting we are getting the t value is - 0.35 so I am drawing this distribution now, so what kind of yeah this is right tailed test right, right tailed it is 0.05 but when we do the sample statistics we are getting that is - 0.35 but this 0.72 is when Z equal to - 0.35 what is the left side area when Z equal to plus 35 what is the right side area? So, that added area is this 0.72, 0.723. Now since it is a one tailed test we have to divide by 2 when you divide this one it is 0.36 so this 0.36 is greater than 0.05.

So, we have to accept null hypothesis. In case if the p-value is for example 0.03 so we will be stand, will be landing on here so we have to reject the null hypothesis yeah this is right tailed test this is the right tailed test.

(Refer Slide Time: 09:20)



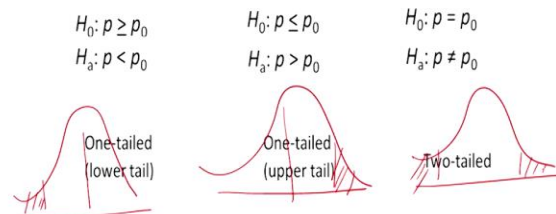
What was the t value here what was the t value one tailed test about the population mean when Sigma is unknown the previously the value of t is -0.384 , so when it is -0.384 the corresponding p -value suppose if it is -3.84 it will be here this will be -0.384 , -0.384 . so, this left side area is 0.3526 because in Python you see that from the t value you can find out the area `stats dot t dot cdf`, you have to find out the way to enter the t value and corresponding degrees of freedom.

Previously our sample size is 20 so the degrees of freedom is 19 so when the, this is calculated t value the corresponding area is 0.35 . so, what is happening this 0.35 is greater than this is left side left side the area similarly if you put plus 0.384 you will get the area towards left that has to be substrate from -1 , so we will get the right side area. So, the right shady area will be 0.35 approximately 0.35 will be right side area will be here 0.35 . So 0.05 is less so the p -value is more than the Alpha so we have to accept the null hypothesis.

(Refer Slide Time: 10:57)

Null and Alternative Hypotheses: Population Proportion

- The equality part of the hypotheses always appears in the null hypothesis.
- In general, a hypothesis test about the value of a population proportion p must take one of the following three forms (where p_0 is the hypothesized value of the population proportion).



Next we will go for hypothesis testing for proportion similar to null and alternate hypothesis for mean. So, here the Equality part of the hypothesis always appear in the null hypothesis in general a hypothesis test about the value of the population proportion P this is P population proportion must take one of the following 3 forms right. So, for example the $P \geq P_0$ so this is this is a situation like this; what is happening this is this is a left tailed test example for this.

There is another possibility it may appear this way this is your right tailed test. How I am naming left tail or right tail, I'll write a test by looking at the sign of your alternate hypothesis. Now it is a two tailed test what would be two tailed test? This way if anything below all, so will reject it anything above it is, similar to what you have seen previously.

(Refer Slide Time: 12:03)

Tests About a Population Proportion

Test Statistic

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}}$$

where: $\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$

assuming $np \geq 5$ and $n(1-p) \geq 5$

Here the test statistic raised $Z = (\bar{P} - P_0)$ divided by $\sigma_{\bar{p}}$, here the \bar{P} is your sample proportion this is your assumed a population proportion this is a standard error for the proportion. So, the standard error is $\sqrt{(pq)/n}$ that is $\sqrt{(p_0(1-p_0)/n)}$, assuming one assumption here is that the value of np and npq should be greater than or equal to 5 because this is this is this follow binomial distribution. If you if you want to approximate binomial distribution to the normal distribution. So, the assumption is we're np and npq should be greater than or equal to 5

(Refer Slide Time: 12:48)

Tests About a Population Proportion

Rejection Rule: p -Value Approach

Reject H_0 if $p\text{-value} \leq \alpha$

Rejection Rule: Critical Value Approach

$H_0: p \leq p_0$ Reject H_0 if $z \geq z_{\alpha}$
 $p > p_0$

$H_0: p \geq p_0$ Reject H_0 if $z \leq -z_{\alpha}$
 $p < p_0$

$H_0: p = p_0$ Reject H_0 if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$
 $p \neq p_0$

What is the rejection rule the same rejection rule if the p -value is less than or equal to α we have to reject it, for the critical value approach the same thing if it is Z this is less than or equal

to so p is greater than equal to p_0 this is a right tailed test so this is p less than equal to p_0 left-tail test. So, there right tailed test the Z value is more than Z_{α} , reject it. For the left tailed test if the Z value is less than minus α , reject it.

When it is a two-tailed test p equal not equal to p_0 so whether it lies on either side of the minus α by 2 and plus α by 2 we have to reject it.


(Refer Slide Time: 13:26)




Two-Tailed Test About a Population Proportion

Example: City Traffic Police

For a New Year's week, the City Traffic Police claimed that 50% of the accidents would be caused by drunk driving.

A sample of 120 accidents showed that 67 were caused by drunk driving. Use these data to test the Traffic Police's claim with $\alpha = .05$.



13

We will take an example suited traffic police that is the example for the New Year's week the city traffic police claimed that 50% of the accident would be caused by drunk driving. So, the sample of 120 accident showed that 67 were caused by drunk driving, use the data to test the traffic police claimed that α equal to 0.05. Similar to that here also the sample data is given and population proportion is given.

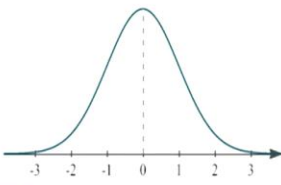
What are the sample data first you will solve this p-value approach what are the sample data is given we go back see n is 120 okay the probability actually here the success is 67. So, we have to find out \bar{p} \bar{p} equal to 67 by 120 this is a sample data even. So, the population proportion capital P equal to 0.5 and α equal to 0.05. So, now we will use this data we will test the claim that the 50% of accident Road caused by drunk driving.

(Refer Slide Time: 14:49)

Two-Tailed Test About a Population Proportion



1. Determine the hypotheses.

$H_0: p = .5$
 $H_a: p \neq .5$
2. Specify the level of significance. $\alpha = .05$
3. Compute the value of the test statistic.



$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.5(1-.5)}{120}} = .045644$$

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{(67/120) - .5}{.045644} = 1.28$$



15

So, first we will go for p-value approach the first step in the hypothesis testing is determine in the hypotheses. So, what is the null hypothesis $H_0: P = 0.5$, alternative hypothesis: $P \neq 0.5$ this is given in the problem. The next step is specifying the level of significance alpha equal to 5%. The third step is compute the value of the test statistics that is the value of Z, so for finding the value of Z we need to know the standard error of the population proportion.

So $\sigma_{\bar{p}} = \sqrt{(p_0 (1 - p_0) \text{ divided by } n)}$ the P_0 it is nothing but the assumed population proportion that is 0.5, $1 - 0.5$, n is the sample size 120 so the standard error of the population proportion is 0.045644. Now we will use our traditional said formula, $Z = (\bar{P} - P_0) \text{ divided by } \sigma_{\bar{p}}$ here the \bar{p} is nothing but our sample proportion. So, what is the sample proportion out of 120, 67 accidents due to drunk driving.

So \bar{p} is 67 divided by 120 minus this is our assumed a population proportion that is 0.5 the $\sigma_{\bar{p}}$ already we got it 0.045644 so the Z value is 1.28. when alpha equal to 0.05 because it is a two-tailed test why we are calling it a two-tailed test when you look at the sign of over alternative hypothesis it is not equal to if it is not equal to it is a two-tailed test if it is greater than that is a right tailed test if it is less then, then it is a left tailed test.

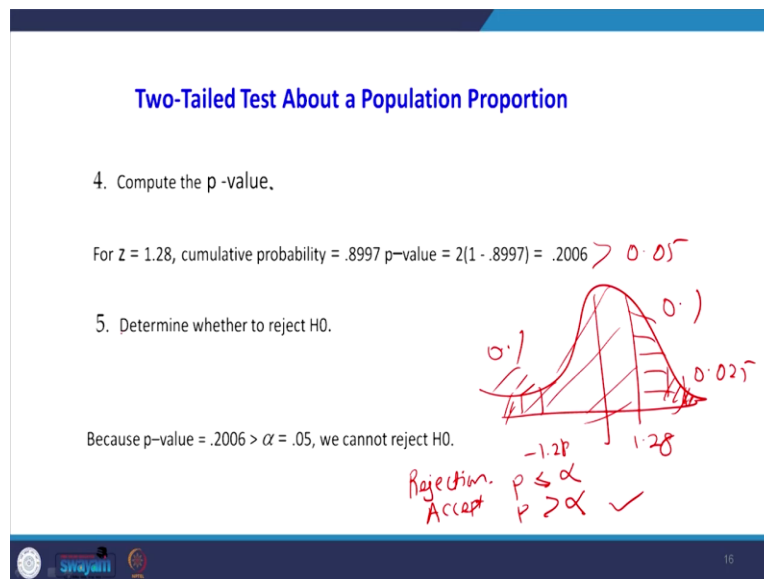
Now since the sign of our alternative hypothesis is not equal to type it is a two-tailed test. So, when a alpha equal to 0.05 when you divide by 2 so, this side area is 0.025 this side area is 0.025

when it is a point zero to five the corresponding Z value is 1.96 left hand side is - 1.96 now what has happened our calculated Z value is 1.28. so, 1.28 will be approximately here, 1.28 now when you compare 1.96 and 1.28 this 1.28 is lying on the acceptance region, so we have to accept null hypothesis.

But the methodology for testing the hypothesis what we are using is the p-value approach so the decision of accepting or rejecting our hypothesis not with respect to 1.96 and 1.28. Here we are going to compare the probability so what is that probability is see this left hand side 0.025 the right hand side also 0.025, so what we are going to do when the calculated Z value is 1.28 we are going to look at what is this side area.

Similarly because it is a two tail test when it is - 1.28 we are going to look at this side area by adding this side area plus this side area then after adding the two side area if it is exceeding 0.05 we are going to accept the null hypothesis otherwise we are going to reject it. Now what has happened when it is z-values 1.28.

(Refer Slide Time: 18:54)



The corresponding area is the corresponding area is it is 1.9887 as I told you the fourth state fourth step is compute the p-value so when Z equal to 1.28 the cumulative probability from left to right 0.8997 so when we go back here so this side area when Z values 1.28 this side area is

0.8997 so the right side area this side area will be $1 - 0.8997$ that is approximately 0.1, so what happened but here the area is 0.025.

Now we have to take the decision you can compare 0.025 and 0.1 when they compared 0.1 and 0.25 so the 0.1 is lying on the acceptance side. So, we are about to accept null hypothesis that is the one way to take the decision otherwise the right side area is 0.1 similarly when Z value is -1.28 this left side area is 0.1 so when , when you add this $0.1 + 0.1$ that will be 0.2006 so that value is greater than ever 0.05. So, this added value is greater than 0.05 so we have to accept our null hypothesis.

So, what is the simple rule if the p-value the p-value is less than alpha reject a null hypothesis less than or equal to if the p-value is less than or equal to alpha reject null hypothesis you the p-value is greater than alpha except a null hypothesis. So, now here the p-value that is 0.2006 is more than this so this is the condition for rejection what is the condition the p-value is less than alpha reject it, if the p-value is greater than alpha accept it.

Now what happened now is the p-value is the second condition is satisfied that is the p-value 0.2006 that value is greater than 0.05. So, we are accepting our null hypothesis now we will go back to the step compute the p-value for Z equal to 1.28 the cumulative probability is 0.8997 then the p-value is nothing but $1 -$ of 0.8997 because the two-tailed test so material by two outer multiplying that you are getting 0.2006 that is greater than 0.05.

So, next we go to next step determine whether to reject H_0 or not because the p-value this 0.2006 is greater than alpha that is a 0.05 we cannot reject it that means we have to accept null hypothesis.

(Refer Slide Time: 22:07)

```

In [13]: from statsmodels.stats.proportion import proportions_ztest

In [14]: count=67

In [16]: samplesize = 120

In [17]: P=0.5

In [18]: proportions_ztest(count, samplesize,P)

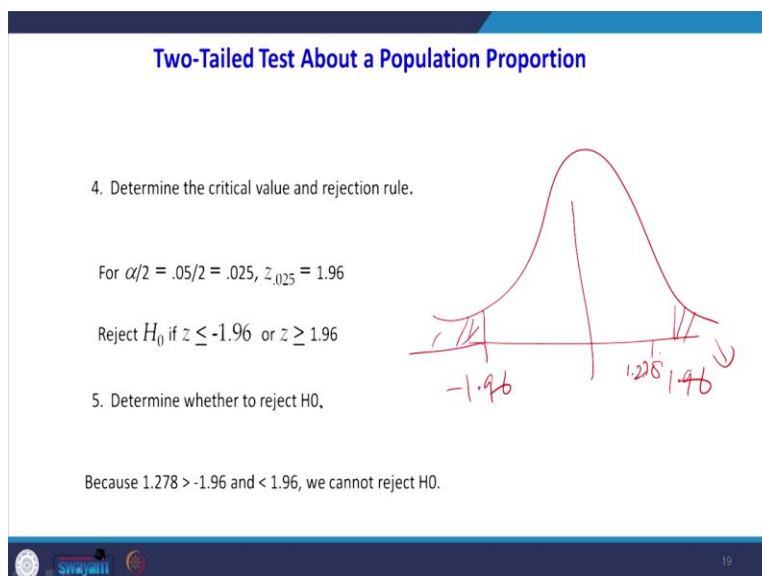
Out[18]: (1.286806739751111, 0.1981616572238455)

```

In Python there is an inbuilt function is there so what to do for that from statsmodel dot stats dot proportion import proportions underscore Z test the proportion test always the Z test there would not be proportion t-test that is number that is not available. So, only if the proportion test means the Z test. So, count is equal to 67 that is the number of success. The sample size is 120, so this capital P is the population mean what we assume it is 0.5, so proportion underscore Z test the syntax is count come on sample size comma capital P.

So, we will get the Z value is 1.28 and the p value is 0.19 so 0.19 because we previous slides at 1.20 it is the address of approximation.

(Refer Slide Time: 23:02)



The same proportion test will solve the help of critical value approach determine the critical value and rejection rule because for when it is alpha by 2 that is the for area 0.25 they said well is 1.96 and the left side is - 1.96 if the calculated it is this way 1.96 - 1.96 if the calculated Z value is going this side or this side we are to reject it. So, what has happened the Z value is 1.278 so 1.278 is in the will be here 1.278, so you have to accept the null hypothesis we have to accept the null hypothesis.

Dear students previously we have solved this population proportion test with help of p-value approach then we have solved with the help of critical value approach both a time we have accept a null hypothesis that is the P equal to 0.5. That is a 50% of accident is due to drunk driving. So, I will conclude that in this lecture what you have seen first you have seen t test when you will go for t test when Sigma is not known when the sample size is below 30 we should go for t test.

We have solved one problem for hypothesis testing, we have solved with the help of p-value approach and critical value approach. After that we have solved a problem using a population proportion mean so that means the population proportion is given we have tested whether population proportion can be accepted or not accepted, thank you very much. In the next class we will see different types of error while doing hypothesis testing, thank you very much.