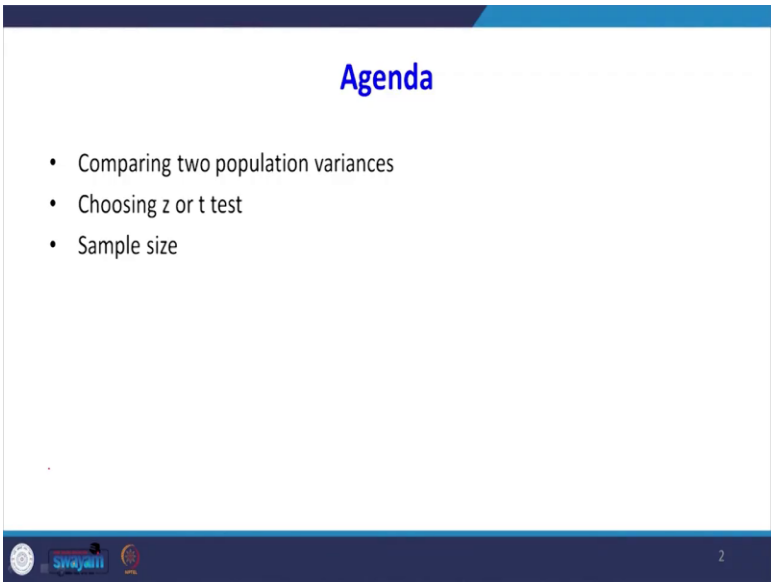
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

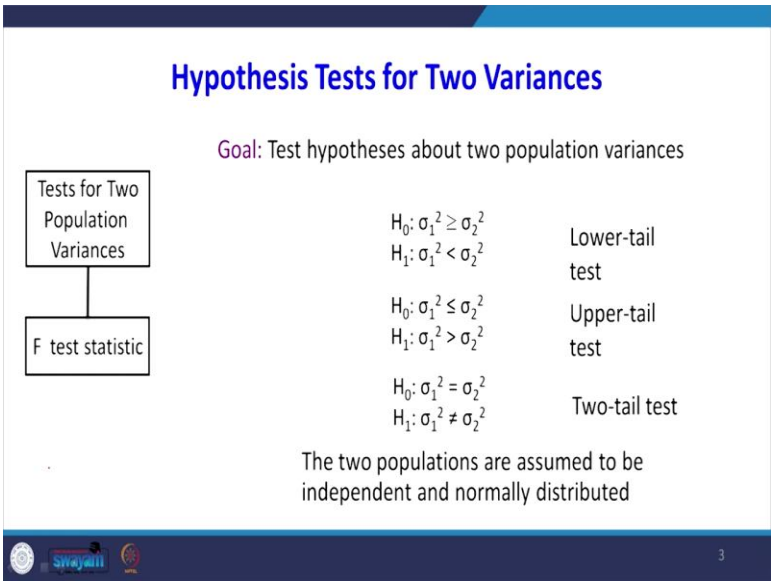
Lecture – 22
Hypothesis Testing: 2 Sample Test-III

(Refer Slide Time: 00:37)



Welcome students in this class we will continue with the 2 sample hypothesis testing. In this class we will see how to compare population variance of 2 population so agenda for this class is comparing 2 population variances then many times student may have this doubt when to go for z test went to go for t test I will clarify that when to go for z test. The third one is it is the most important that what should be the sample size for doing any statistical analysis.

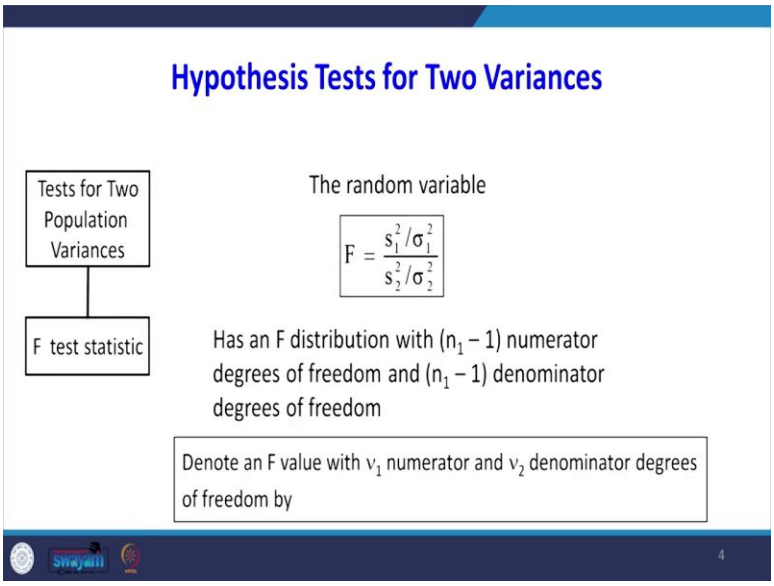
(Refer Slide Time: 00:58)



Now we will test the hypothesis test for 2 variances the goal is to hypothesis about population variances you see that the H0: $\sigma_2^2 \leq \sigma_1^2$, the H 1: $\sigma_2^2 > \sigma_1^2$ it is over left tailed test otherwise the lower tail test it may be this way right it will be right skewed one what will happen this is the here also there is a left side this is left side test this is right side test this is 2 tailed test.

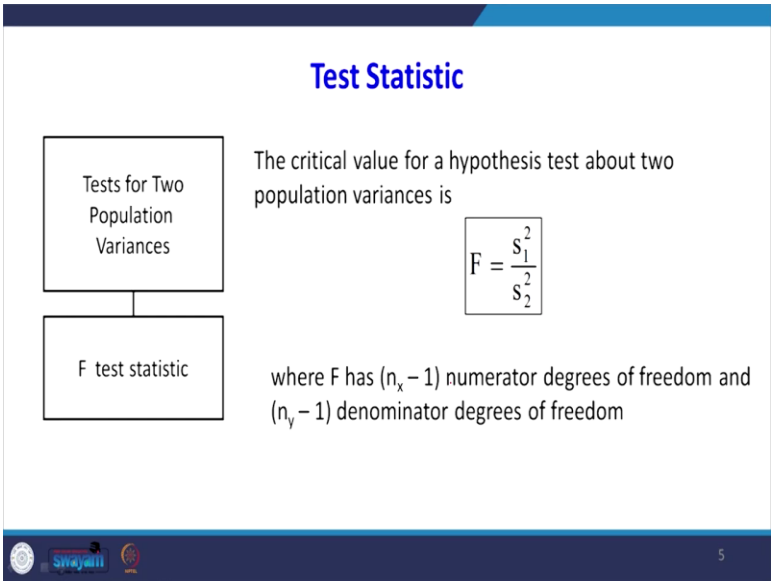
You have to remember that I did not draw the normal distribution this is a right skewed distribution this distribution is called F distribution. So, in the F distribution we have to find out the F statistics that F statistics will decide will help us to accept or reject null hypothesis. As usual here also there may be a left-tail test right tailed test or 2-tailed test but very important assumption which are which here to remember that the 2 populations are assumed to be independent and normally distributed.

(Refer Slide Time: 02:15)



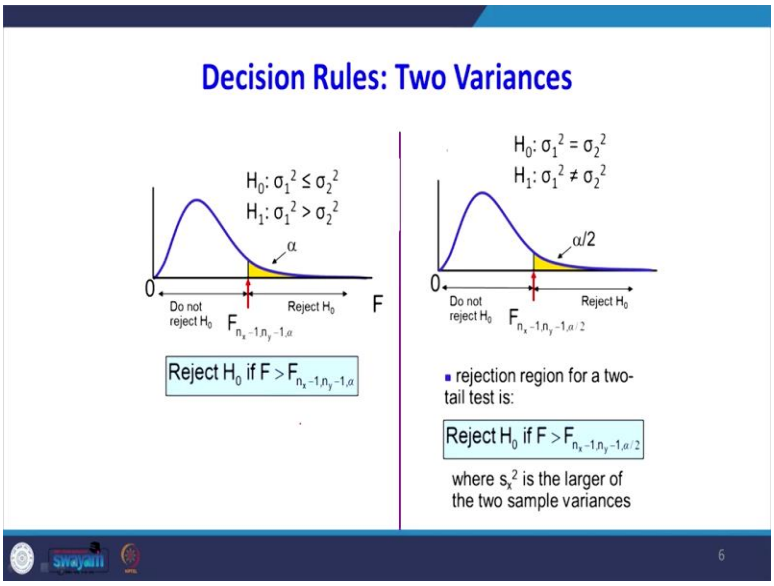
The test statistics for comparing 2 population variances F is nothing but $(s_1^2 / \sigma_1^2)/(s_2^2 / \sigma_2^2)$, the F statistic here is (s1 by Sigma1)/ (square s2 by Sigma2) whole square. If you are assuming both the populations having equal variance the F will become s1 square by s2 square so it the F, n 1 - 1 numerator degrees of freedom and n 2 - 1 this is the n2 - 1 denominator degrees of freedom.

(Refer Slide Time: 02:55)



Yes I told you the critical value for a hypothesis test about 2 population variances you F equal to s 1 square by s 2 square what we are assuming here both population have equal variance where F was n1-1 numerator degrees of freedom and this is n2 - 1 denominator degrees of freedom.

(Refer Slide Time: 03:17)

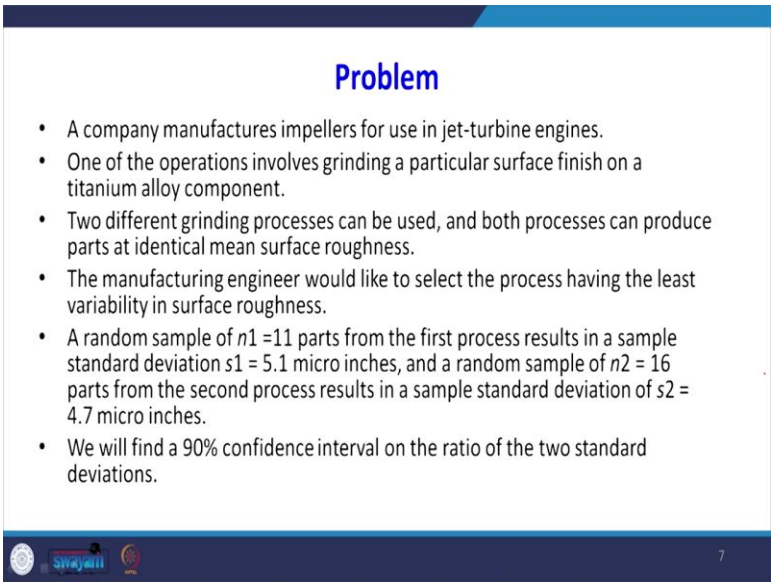


The decision rule for 2 variances for example most of the time the F test is only a right tailed test because whenever there is a variance so only we are bothered about only the upper limit of the variance the lower limit you will not bother about. Because taking lower limit is there is no meaning, so the what we have to assume that what is more important for is it should not exceed the upper limit of the variance.

It is like you see the bus has to come 9 o'clock if it has come 9 :05 or 9 :10 you will get bored but if it is coming early that there would not be any problem like that we have to bother about only the upper limit if there is a lower limit is not that much important. So, the degrees of freedom here is this is n1 - 1 this is n 2 this is n 1 this is n 2 and other things for comparing 2 tailed test right as I told you for comparing 2 tailed test you see that this is Sigma 1 square equal to Sigma 2, square Sigma 1 square not equal to Sigma 2 square the rejection region for your 2 tailed test is see that this is n 1 – 1, n 2 – 1.

So what we have to do while finding the value of F right this is 1 while finding the value of F we have to maintain that the higher variance should be in the numerator. So, what we have to assume this s1 square is greater than s 2 square so where s 1 square is the larger of 2 sample variance that should go to in the numerator. If you take larger of 2 variants in the numerator you need not bother about the lower limit of the 2 tail test only we have to compare only upper limit of 2 tailed test for at accepting or rejecting null hypothesis.

(Refer Slide Time: 05:12)

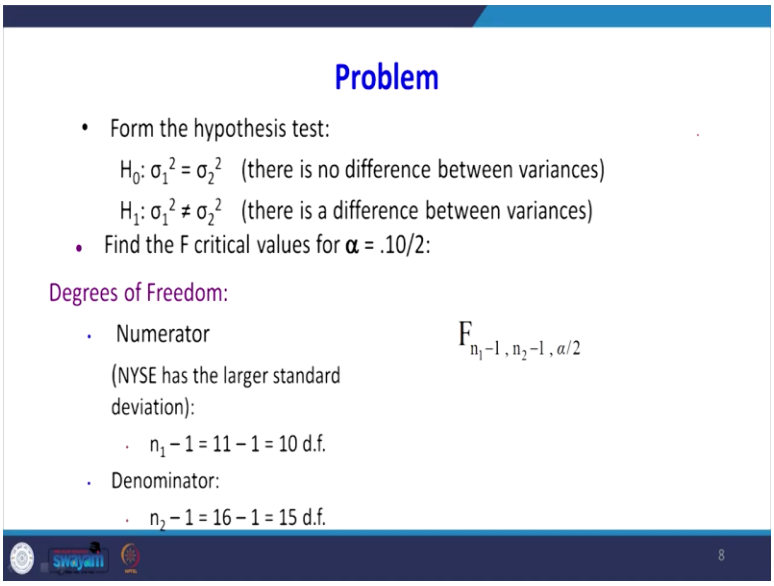


We will take one problem a company manufactures simpler for use in jet turbine engines one of the operations involves grinding a particular surface finish on a titanium alloy component 2 different grinding processes can be used and both processes can produce parts at identical means surface roughness. The manufacturing engineers would like to select the process having the least variability that is a point least variability in the surface roughness.

When you say generally the surface references measured by this way surface roughness is measured by this way suppose the surface roughness is this one so it is not good the surface roughness it cannot be perfectly smooth covered there should be a smaller variations. So, for the manufactures are also interested would like to select a process having least variability in the surface roughness a random sample of n1 11 parts from first 2 processes result a sample standard deviation of 5.1 micro inches and random sample of into 16 parts from the second processes result in a sample standard deviation of s2 = 4.0 micro inches.

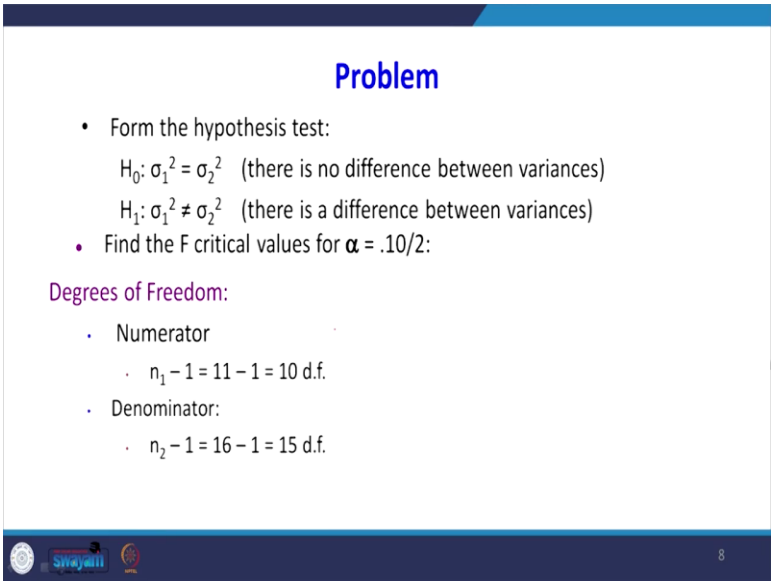
We will find here 90% confidence interval on the ratio of 2 standard deviation then we will compare whether it is these variances are equal or not equal.

(Refer Slide Time: 06:42)



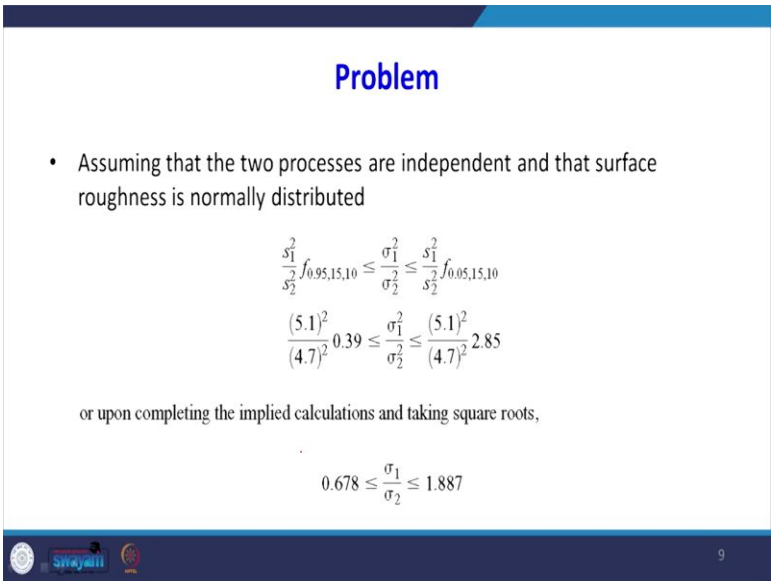
As usual first we have to form the null hypothesis the null hypothesis is $\sigma_2^2 = \sigma_1^2$ what is the meaning of this both the process are having equal variances. Alternative hypothesis is $\sigma_2^2 \neq \sigma_1^2$ there is there is a difference between variances so find the critical value alpha equal to 10%.

(Refer Slide Time: 07:06)



So the first thing is you have to find out the numerator degrees of freedom that is n1 - 1 so 11 - 1 10 is the numerator degrees of freedom it is the denominator degrees of freedom n 2 - 1 so 16 - 1 15 is denominated degrees of freedom.

(Refer Slide Time: 07:24)

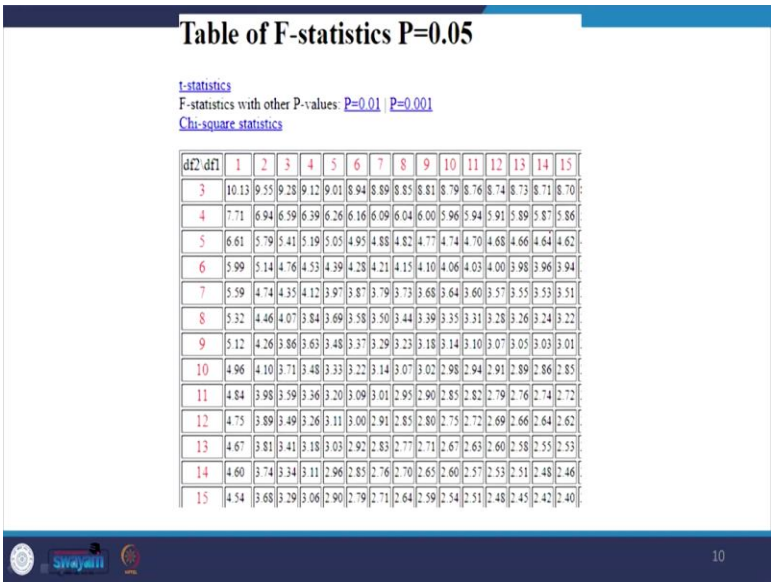


Assuming that the 2 process are independent and the surface reference is normally distributed we are going to find out the confidence interval of the ratio of their variance. So, what is the logic is suppose Sigma 1 square equal to Sigma 2 square, so it will in that interval 1 will be captured so F distribution is like this so here I am going to find out the confidence interval if you look at the F table the area is given only from right to left so when area equal to 0.05 because we have told it is a 10% see if the right side is 0.05 left side is 0.05.

If you look at the F table we can read the for 0.025 significance level what is the corresponding F value right if you want to know the left side 0.05 so you have to read 0.95 significance level then only you can find out the lower limit of the F. So what we see we write $(\sigma_1^2 / \sigma_2^2) \le (s_1^2 / s_2^2)$, first time finding the upper limit upper limit is when it is a 0.05 see that when F equal to 0.05 numerated degrees of freedom is 15 denominated degrees of freedom is 10 that will be my upper limit.

The larger value of the variance should go to numerator. The left side the area the left side critical value is f 0.95 right if we want to because the F table read right to left 0.95 15, 10 degrees of freedom okay so s1 square 5.1, 5.1 divided by 4.7 then 0.95, 10 , 15 deals of freedom this value we are to read from the table. First we will read the upper limit so when the degrees of freedom is 0.05 numerator degrees of freedom is 15 and the denominator degrees of freedom is 10 we will see what is the F value.

(Refer Slide Time: 09:48)

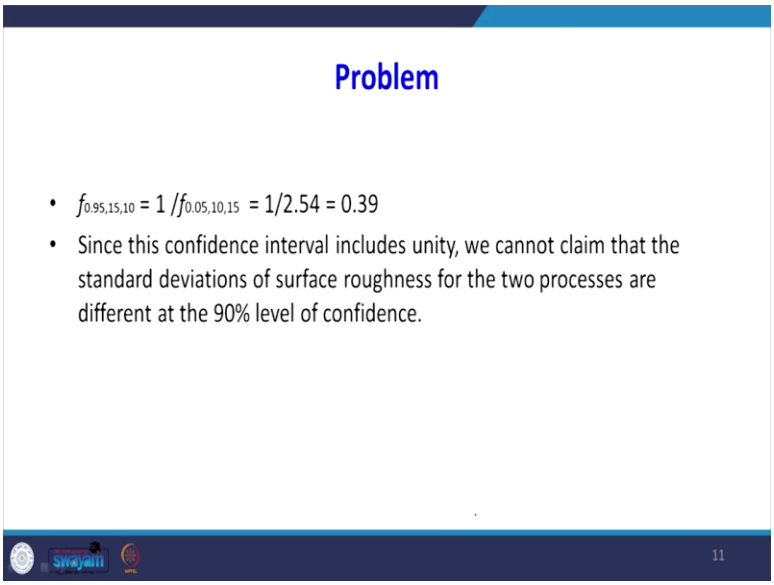


df2\df1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40

So numerator degrees of freedom is 15 denominator degrees of freedom is 10 see that this value this value 2.85 that is where F value now what we are to do we have to know that lower limit for that when significance level is 0.95 when 15, 10 degrees of freedom we have to find out what is the F value for that purpose what you are to do we are to reverse the degrees of freedom so 10, 15 here 10, 15 is this point 2.54 because DF 1 do column says the numerator degrees of freedom the rows is denominated liters of freedom.

So this is 10, 15 so this 1 is 15, 10 because the column says the numerator this one is 10, 15 degrees of freedom when alpha equal to 0.05. So, if you want to know lower limit of their F so what you have to do?

(Refer Slide Time: 11:07)

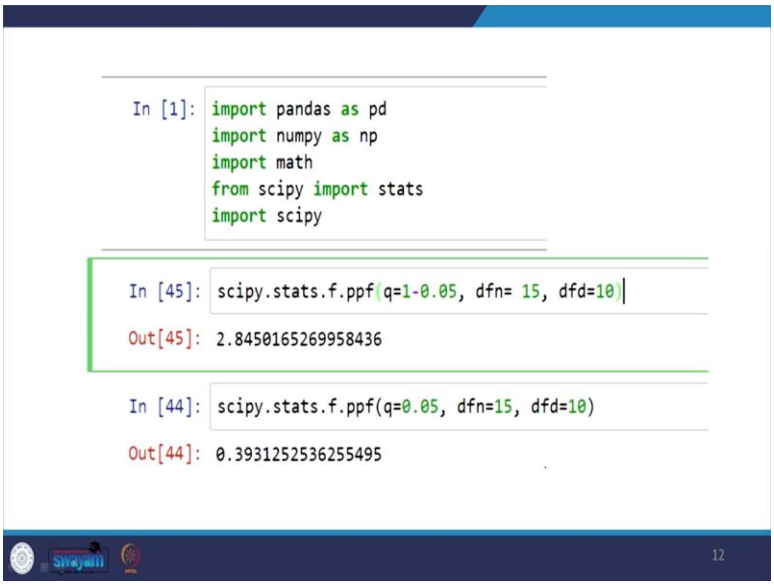


You see that if you want to know the 0.95 for 15, 10 so we had to find the first we had to reverse the degrees of freedom 10, 15 then we had to find out alpha equal to 0.05 then we had to find the inverse of that so that is 1divided by 2.54 how we got 2.54 this value 2.54 so 2.4 we got 0.39 so and going back again so that is why we got this 0.39 when you simplify this the lower limit is 0.678 the upper limit is 1.887 actually after taking the square root because upon completing the implied calculation take a square root of that we are getting this one.

You see that the range is capturing one that means there is a possibility Sigma 1 equal to Sigma 2 because there is this ratio implies that there is a possibility it may take one also so we have to accept our null hypothesis that is a Sigma 1 equal to Sigma 2. Since this conference interval includes unity we cannot claim that the standard deviation of surface roughness for the 2 process are different at the 90% level of confidence.

Going back in case the unity is not coming here so we cannot say both variants are equal since we are able to capture unity because lower limit is 0.6 upper limit is 1.8 there is a possibility the value of the ratio of Sigma 1 by Sigma 2 will become 1 so Sigma 1 by Sigma equal to 1 so that there is a possibility Sigma 1 will become equal to Sigma 2.

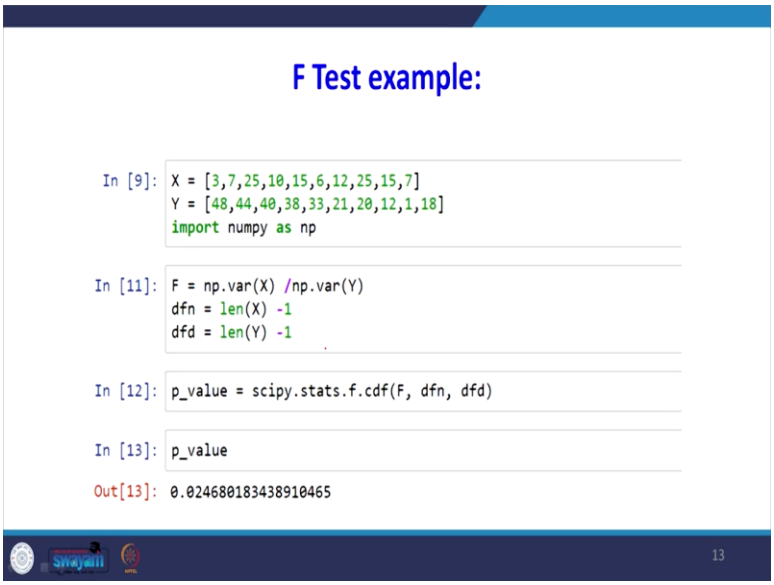
(Refer Slide Time: 12:57)



Now we will use a Python for finding the F table so for that you have to import pandas as pd import numpy is np, import math, from scipy import stats okay you can import scipy so what you have to do scipy.stats.f.ppf you have to give the probability so 1 - 0.05 means 0.95, so 0.95 numerated degrees of freedom is 15 and 10 we are getting 2.84 so that value is nothing but 10, 15 2.84 scipy.stats.f.ppf, q equal to say 1 - 0.95 numerator dfn is nothing but numerator degrees of freedom 15 dfd denominator degrees of freedom 10 we are getting 2.84.

So, if you want to know the lower limit here we can directly read from the F table scipy.stats.f.ppf q 0.05 numerator degrees of freedom is 15 denominated degrees of freedom is 10 you are getting 0.39 so this was our lower limit this was the our upper limit which previous problem.

(Refer Slide Time: 14:10)

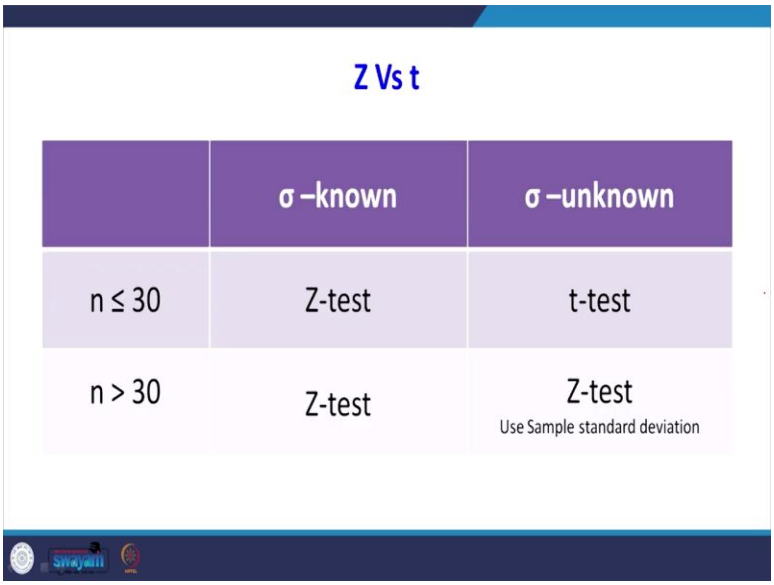


So, for what you have done there are 2 group of population the standard deviation that is the variance of that 2 populations are given instead of that there is a possibility there is a population 1, X will be given Y will be given you have to find out the p value for that. Suppose I am assuming this is the population 1, I am going to call it as capital X there is another population 2, I am going to call it is capital Y, so my null hypothesis H0: I am going to assume $\sigma_x^2 = \sigma_y^2$. Alternative hypothesis is $\sigma_x^2 \neq \sigma_y^2$ right I am going to take alpha equal to 5 % that is 0.05.

So, declare variable X declare variable Y, import numpy as np then find out the ratio that is variance of X divided by variance of Y then you find numerator degrees of freedom len function that will tell you how many element is there in that array 1, so number of element - 1 that is a degrees of freedom for numerator, number of element - 1 there is a degrees of freedom for denominator. So, to get the p value equal to scipy.stats.f.cdf this is the syntax.

First declare what is F we have found this here then say what is it degrees of freedom numerator degrees of freedom denominator when you enter p value we are getting 0.024 so what will happen this your F distribution okay this is 0.05 this 0.024 see that is 0.024 is lesser than 0.05 so we have to reject a null hypothesis. When you reject it we are accepting that Sigma X square not equal to Sigma Y square. This is easiest way for testing variance of 2 population.

(Refer Slide Time: 16:27)



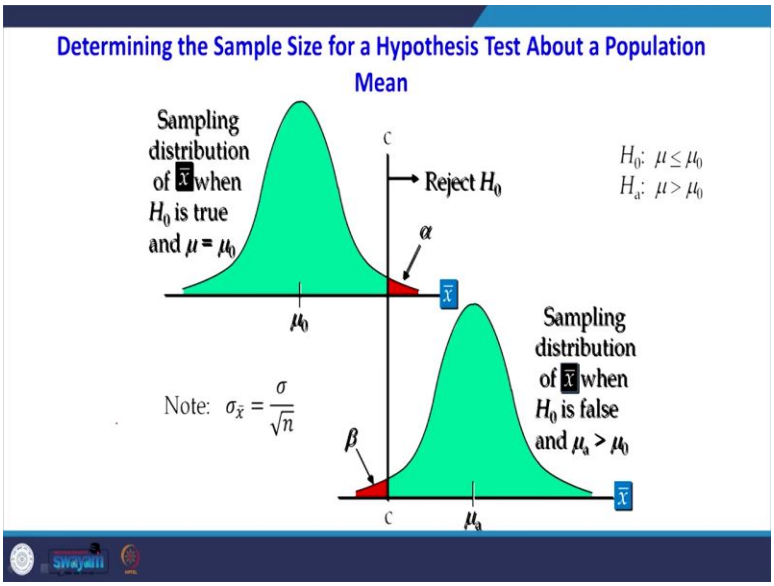
Many students may have doubt when we will go for Z test when you go t tests before going this I will summarize what you have done we have done one sample Z-test we have done t-test then we have done Z proportion test then we have done 2 sample z test then we have done 2 sample t-test. In the 2 sample t-test we have heard assumption population variance are equal are not equal. Then we have done 2 sample Z proportion test then later we have compared to sample variance test comparing we have compared to population variance.

After completing this then people may have doubt when should we go for Z test a t-test. You see that whenever the Sigma is known what is the meaning of the Sigma is known as whenever the population standard deviation is known you should go for Z test. So, to decide when should go for Z test there are 2 criteria one is the sample size and other is whether the Sigma is known or unknown. So, without considering the sample size whenever the Sigma is known whether the sample size is less than 30 or greater than 30 you should go for Z test.

So, when you look go for t test whenever Sigma is unknown and n is less than 30 you should go for t test. There may be a possibility Sigma is unknown but n is greater than 30 so instead of t test you can go for Z test because as we have studied previously the t distribution is the special case of your Z distribution. Whenever the degrees of freedom for example t distribution will be flat whenever the degrees of freedom is increasing, is increasing it will behave like your Z distribution.

So, whenever the Sigma is unknown n is greater than 30 you can go for we Z test that is why in many statistical package there would not separate tab for running Z test there will be a tab only for t test.

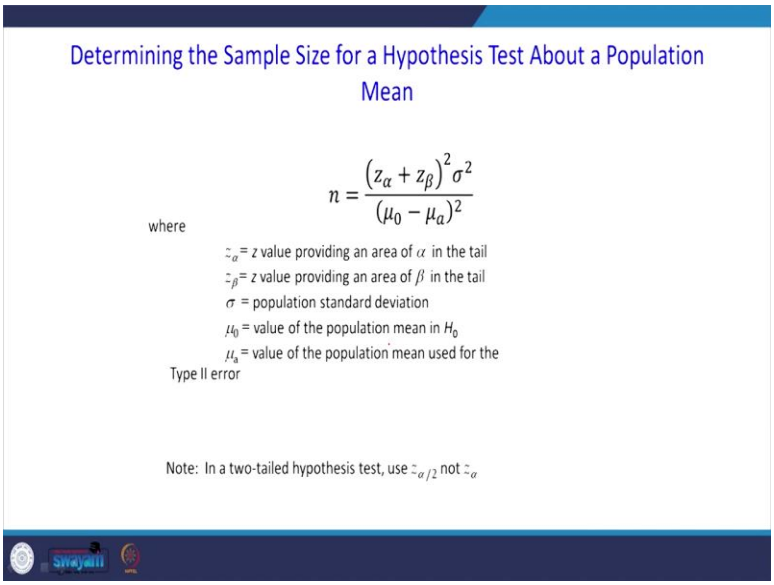
(Refer Slide Time: 19:01)



The another important question students will have is how to choose the sample size determining the sample size for a hypothesis the test about the population mean you see that there are 2 population mu0 is the base population there is a mu_a is alternative mean, so this is a right tailed test where is a right tailed test if the mu greater than mu0 you will reject it. Now what is happening any point which goes right hand side we will reject it.

This much portions which are falling acceptance side of my distribution I have accepted. Now this is beta there is a false acceptance this is alpha in correct rejection. By considering the Alpha and Beta you see in this point the value of X bar is same X bar for this population X bar for this population same.

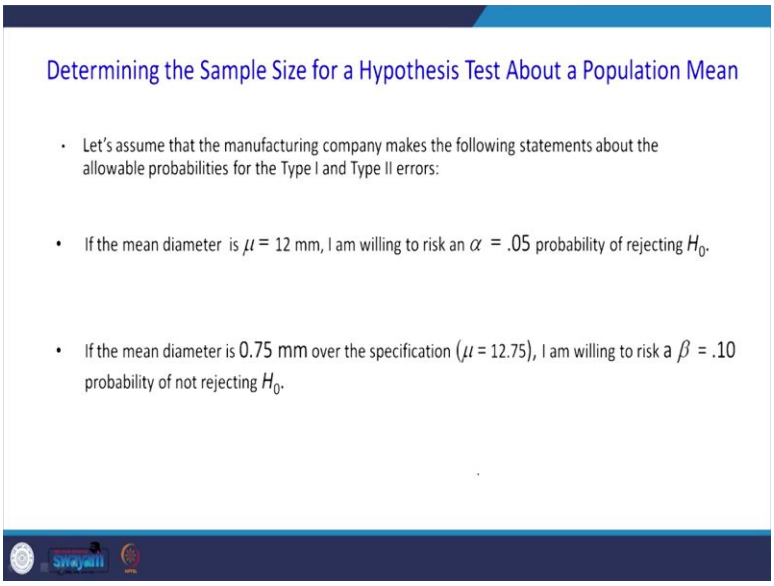
(Refer Slide Time: 20:00)



So, by equating this X bar for both populations we can derive a formula for knowing the sample size considering Alpha and Beta this we can derive it, it is very simple derivation. So, what how to derive this we have to X bar because this line is same for both the population. So, here what will happen Z alpha is nothing but (X bar - mu0) divided by Sigma by root n for here Z beta equal to (X bar - mu_a) divided by Sigma by root n.

You see that the value of this Z beta will become negative because this is lower value minus upper value, so when you equate this when you from these 2 from these 2 equation 1 and 2 when you simplify you can Sigma by root n we can get the value of n. So, when you the value of n is (Z alpha + Z beta)^2 multiplied by $\sigma_2^2 \neq \sigma_1^2$ by (mu0 - mu_a) whole square okay because it is a 2 tailed test if it is a 2 tailed test we have to use Z alpha by 2 not Z alpha.

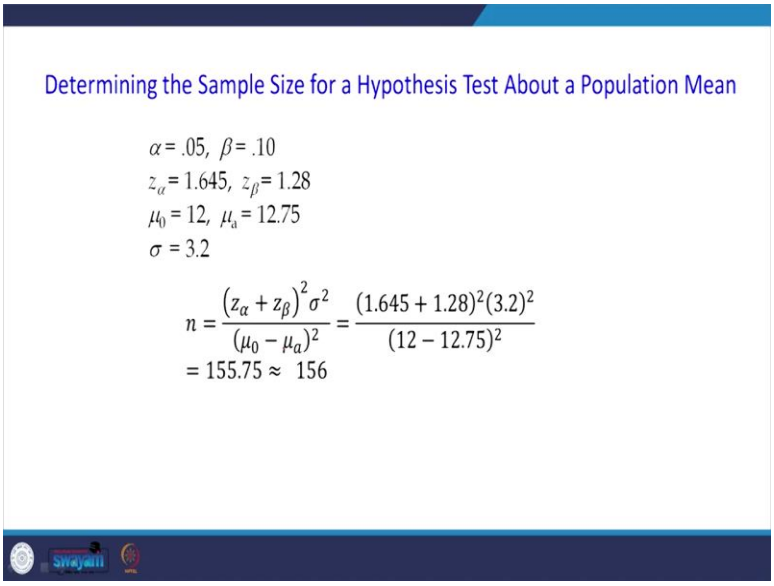
(Refer Slide Time: 21:24)



We will do a small problem for this let us assume that the manufacturing company makes the following statement about the allowable probability for type 1 type 2 error. Suppose somebody is manufacturing a shaft whose diameter say 50 mm. If the mean diameter is 50 mm I am willing to risk an alpha equal to 5% of a probability for rejecting null hypothesis if the mean diameter is 0.75 mm over the specification that is if the mu equal to 12.75.

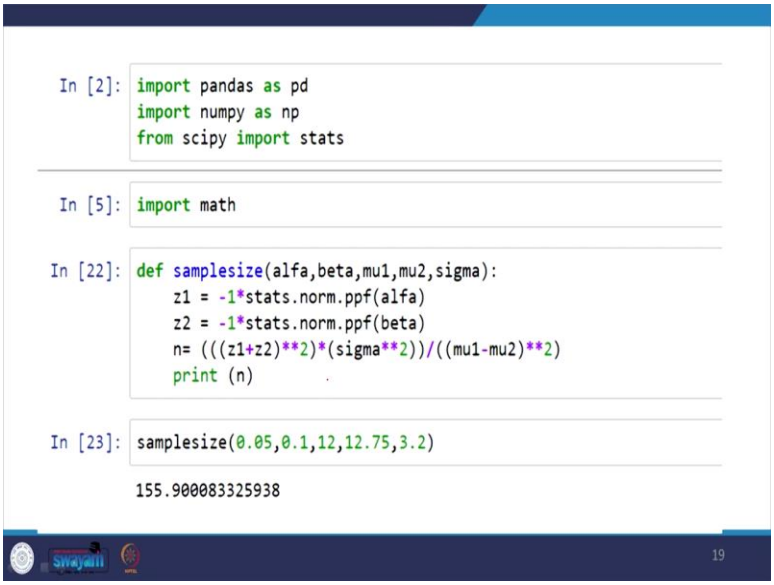
I am willing to take a risk beta equal to 10% probability of not rejecting there is a false acceptance. Now what is happening alpha is given beta is given, alpha is your type 1 error, beta is a type 2 error, see actual mean is given an alternative mean is given.

(Refer Slide Time: 22:27)



So, if that is the case what should be the sample size so alpha equal to 5% beta equal to 10% when alpha equal to 5% we can get it is 1.645 the Z_{beta} we can find out directly how to find out the Z_{beta} I am going back this value Z_{beta} is X bar - mu a divided by Sigma by root n all will be given. So, $Sigma_{Beta}$ is 1.28, mu 0 is 12 mu_a is 12.75 Sigma is 3.2 just substitute this value so it should be 156.

(Refer Slide Time: 23:07)



We will use a Python for solving this problem import pandas pd import numpy as np from scipy import stats import math so we have to define a function samplesize (alpha beta mu 1 mu 2) Sigma so Z1 = - 1 * stats.norm.ppf (alpha), Z 2 = - 1* stats.norm.ppf (beta), n equal to Sigma 1 + Sigma 2 whole square that is the same formula if you substitute it you can print the n. So, when you supply alpha value beta value mu 1 value mu 2 value Sigma value we are getting 155.90.

So, far we what we have seen we have compared to 2 population then we have compared variance of the 2 population we have tested whether the populations are variants are equal or not then we have seen when to go for is a testament to when to go for t test after that by considering the Alpha and Beta value we have found your formula how to decide or how to arrive the value of your sample size it would take an old sample problem they have conducted then we found what was the value of the sample size, thank you.