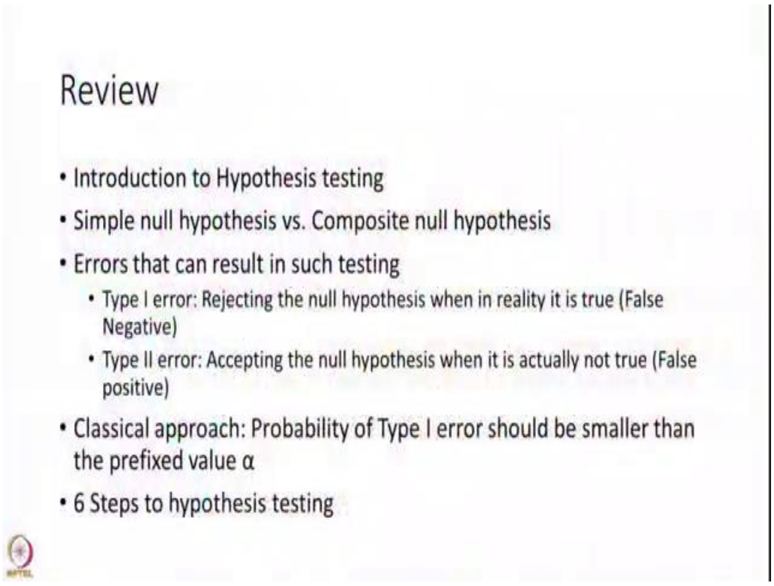
Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 69 - Hypothesis Testing 2

Hello and welcome to the course on Dealing with Materials Data. In the process of learning statistical techniques, we are presently going through sessions on hypothesis testing. We introduced in the previous session what is called hypothesis testing.

(Refer Slide Time: 0:34)



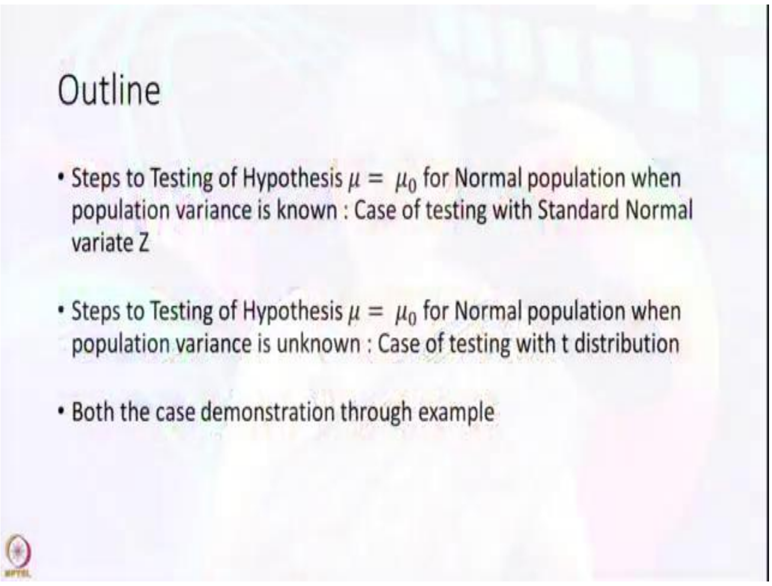
We defined the null hypothesis and we said that if the null hypothesis is such that it defines if it is true and it defines the whole population then it is called a simple null hypothesis. If the null hypothesis is true but still does not define the full population then it is called a composite null hypothesis.

Then we defined, what is called a critical region. We said that we have a sample from the population. And if we can find a region, that when this sample falls in that region, we say that hypothesis is to be rejected, then that region we are going to call critical region.

When we do this, when we make such a decision that null hypothesis is true or null hypothesis is not true, in any such situation, we commit two kinds of errors. The first error is called rejecting the null hypothesis when in reality it is true. In medical parlance, it is known as false negative. And the type 2 error is called accepting the null hypothesis when in actual in reality it is not true. This is called false positive.

Our idea in classical approach is to minimize the type 1 error and later on study the type 2 error, which also gives us by subtracting it from one the power of the test. And then we went through three steps of classical hypothesis testing. In the present session, we want to consider two cases under normal population assumption that is our population comes from a normal distribution. One case is where the variance is known.

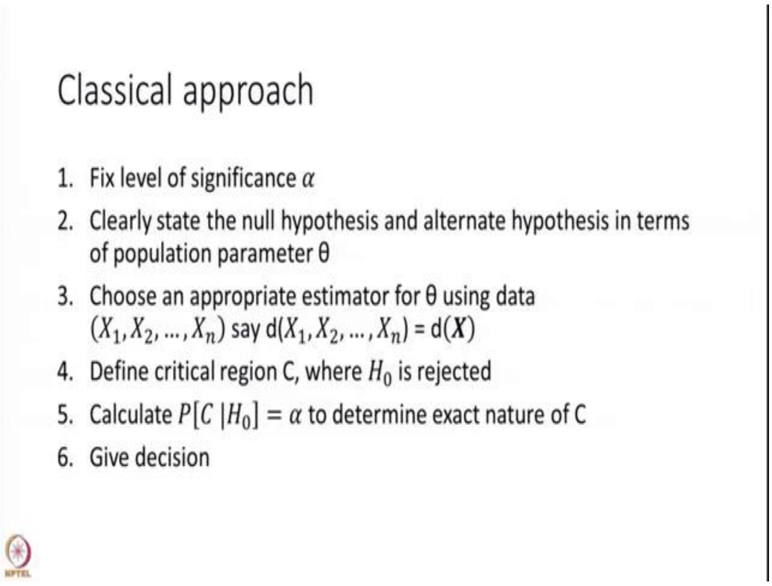
(Refser Slide Time: 2:46)



In both the cases we accept that the mean is not known. One case is when we say that variance is known, and when and we will follow the classical 6 steps to do, derive the test statistic and then we will assume that the variance is unknown and we will follow the same 6 steps to derive the test statistic and the testing procedure.

We will go through one same example for both the cases so that we understand how it the test really applies. I would like to bring it to your notice one thing, you please observe very carefully. The procedure that we are following right now assumes normality. But look at the details, I will bring it out to your notice that what we call test statistic, I will we will define it pretty soon. That actually is a standalone. It is, stands by itself and it does not really need a normal assumption, we will have one such case also in not in this session, but in future sessions.

(Refer Slide Time: 4:15)

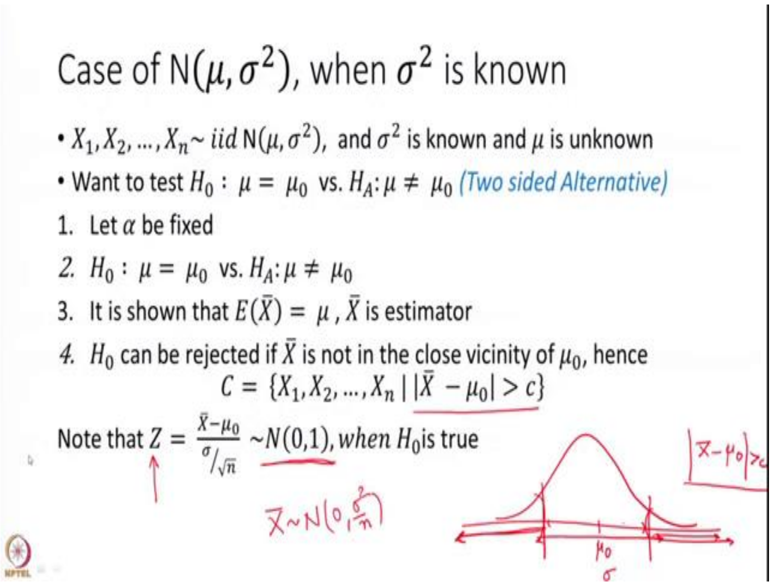


So, let us start with the classical approach of 6 steps quickly we have a fix, first we fix a level of significance alpha, which is a small value typically chosen at 1 percent, 5 percent, 10 percent kind of levels. Then we clearly state what is a null hypothesis and what is its alternate hypothesis, that is critical region false for what condition is what we have to define. Then we choose an appropriate estimator for theta, we call it, say d(X), and then we define a critical region C using d(X) where H naught or the null hypothesis can be rejected, then we calculate the error probability that the data falls in critical region when actually null hypothesis is true and we fix it to a level alpha.

This way we find out exact nature of critical region C, once we have a nature of critical region C, its final point is to make a decision whether the data you have taken, the sample value that you have got falls within the critical region or not, if it does, your decision is hypothesis is to be rejected. If it does not, then you say that you do not have sufficient evidence to reject the null hypothesis.

As I said, we are going to take a model case of normal distribution with the mean mu variance sigma square.

(Refer Slide Time: 6:02)



And here we are going to assume that sigma square is known. So, we have a data which is a sample random sample independent and identically distributed normal mu sigma square sample

$$X_1,\ X_2,\ \ldots,\ X_n \sim iid \text{ N}(\mu,\ \sigma^2), \text{ and } \sigma^2 \text{ is known and } \mu \text{ is unknown}$$

Want to test $H_0 : \mu = \mu_0$ vs. $H_A: \mu \neq \mu_0$ *(Two sided Alternative)*

So, it could be less than mu naught, it could be larger than mu naught but it is not mu naught.

So first, we have to fix an alpha, so let us say that alpha is fixed, we state the null hypothesis and alternate hypothesis

$$H_0 : \mu = \mu_0 \text{ vs. } H_A: \mu \neq \mu_0$$

In the third step, we have to find a test statistic. We are going to perform all the test to test whether the mean is equal to a fixed value mu naught, therefore mean is our point of focus, and therefore, we find a test statistic in terms of the sample value X1, X2, X3, Xn, which is X bar, which says that expected value of X bar is mu and therefore, X bar is a good estimator.

It is shown that $E(\bar{X}) = \mu$, $\bar{X}$ is estimator

Now, when we say that we want to reject the null hypothesis if it is not in close vicinity of mu 0, what do we mean by that? Say this is your mu 0, and you have some normal distribution because that is what we have assumed. And some one sigma limit is this much. This is one sigma limit, this is unknown to you, I mean presently it is known to us. Now, what I would like to say is that I should, once I find a X bar, once I find a X bar, the question is, if my X bar is not exactly mu 0, I cannot expect it to be exactly equal to mu 0, but even if it is in some vicinity of mu 0, I should be happy.

But if it is way beyond the mu 0, then I can say that my null hypothesis does not seem to be true. So, this region where I said that it is way beyond the mu naught value is called a critical region. So that critical region I am defining is that if somewhere here I have X bar, and if I take X bar minus mu 0, and I take its absolute value positive or negative, it should be larger than some predefined value, some value not pre-decided but some value.

See, if it is larger than C, this is what I have written. If my absolute value of X bar minus mu 0 is larger than some value C, which says that it is not in vicinity of mu 0, then that should be my critical region.

H_0 can be rejected if $\bar{X}$ is not in the close vicinity of μ_0, hence

$$C = \{X_1, X_2, \ldots, X_n \mid |\bar{X} - \mu_0| > c\}$$

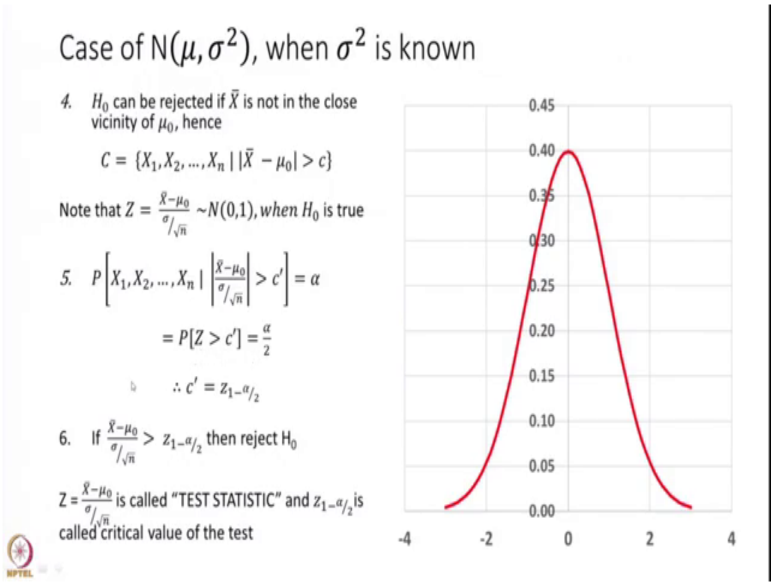
Now what I find is that, if I take

$$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \ when\ H_0 \text{ is true}$$

then it is distributed as normal (0, 1). If $H_0 : \mu = \mu_0$

So if this is the case, so then I actually change my test statistic to Z, because now it is distributed in an area where it neither depends on mu nor depends on sigma square, it is a parametric independent distribution that I have got. And therefore, Z is what I choose as my test statistic or X bar minus mu not divided by sigma over square root n is what I choose as my test statistic and I move forward.

(Refer Slide Time: 11:02)



This step I have just repeated here for the continuity sake. So the next step is that probability that I have to find the type one error.

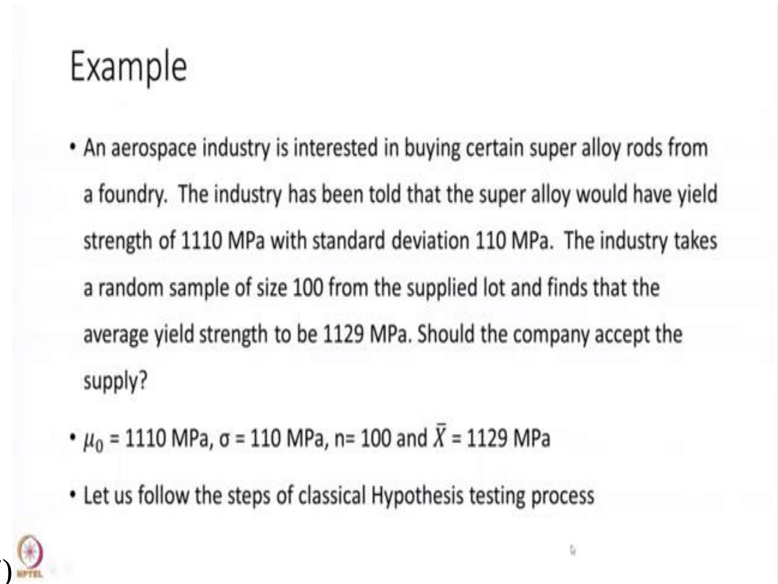
So

$$P\left[X_1,\ X_2,\ \dots,\ X_n \mid \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > c'\right] = \alpha$$

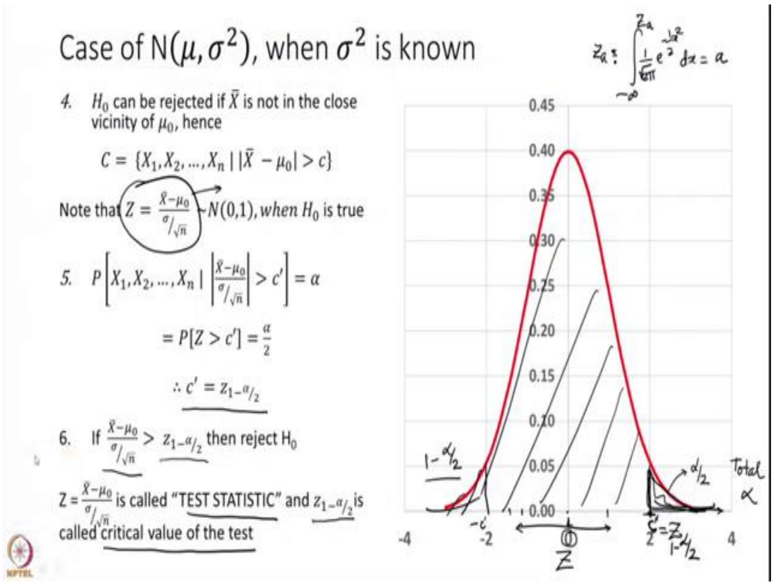
$$= P[Z > c'] = \frac{\alpha}{2}$$

$$\therefore c' = z_{1-\alpha/2}$$

So you remember that this is now the distribution of Z, I have shown.



(Refer Slide Time: 11:47)



See here, this is the, this is the distribution of Z, okay. So it is mean and its one sigma value is here. This is one sigma limits of it. This is one sigma limit and its mean value is 0. Now, when $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > c'$, it means that I am looking at some c dash here and minus c dash here and the value should lie here and this total should be alpha, total value should be alpha.

So it means that I am saying that, that if you take only this much area, we say that it is a Z greater than c dash, so

$$P[Z > c'] = \frac{\alpha}{2}$$

So, now it says that your c dash has to be c dash has to be such that probably of Z greater than c dash has to be alpha by 2 and therefore, c dash is small z, 1 minus alpha by 2. Now these are the small z values, if I put the values remember Z is the capital Z is a random variable, small z is the values that it takes.

So if I take any value, what is the value of c dash? Well, it has to be this has to be alpha by 2, it means that all the area on this side of the curve has to be 1 minus alpha by 2 and remember the Za value is defined as

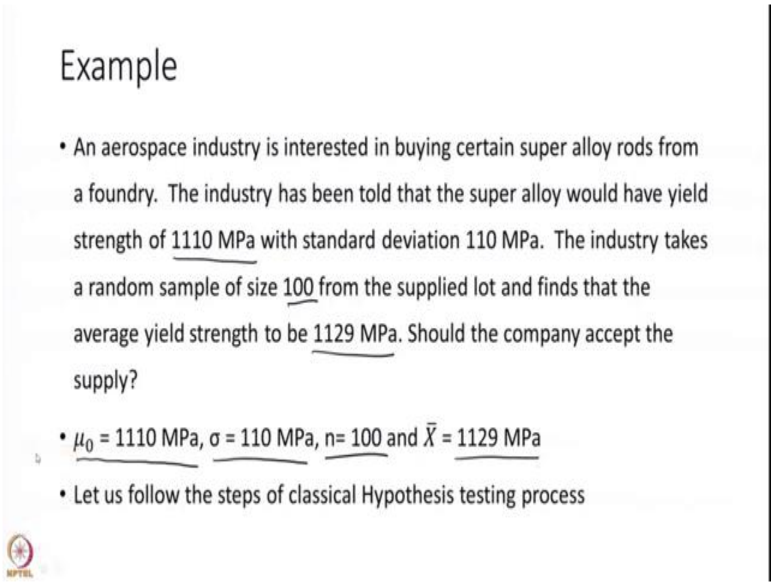
$$Za = \int_{-\infty}^{Za} \frac{1}{2\pi} e^{-\frac{1}{2}x^2} dx = a$$

So you have this this probability is 1 minus alpha by 2, that also brings you to the value of Z sub alpha by 1 minus alpha by 2. And therefore, it says that c dash is Z 1 minus alpha by 2 and therefore, we get the critical region that

$$\text{If } \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \text{ then reject } H_0$$

In this, once again I say that the test statistic, Z, X bar not as a value, but as an estimator. This is called a test statistic. As I have mentioned here and small z 1 minus alpha by 2 is called the critical value of the test, because that is the value which decides the critical region.

(Refer Slide Time: 15:43)

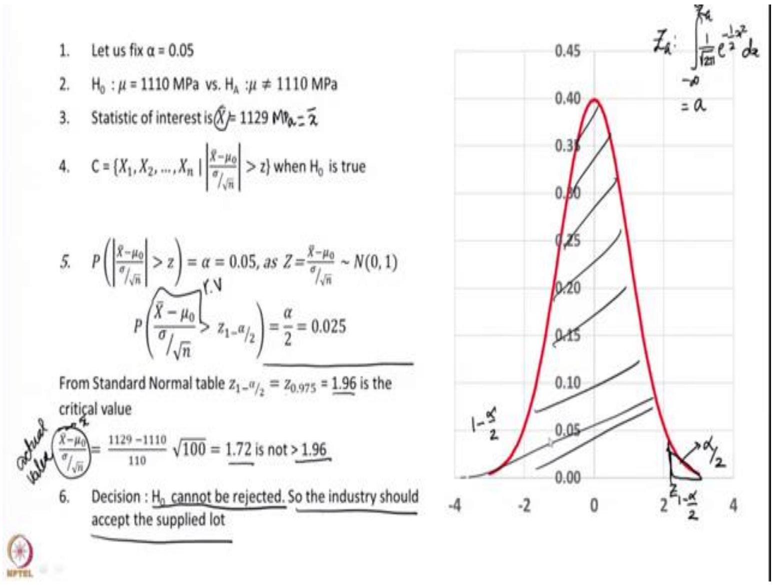


So now, to understand it better let us have go through a example. So here I have an example, it is a completely a made up example, an aerospace industry is interested in buying certain super alloy rods from a foundry and the industry has been told that the super alloy would have yield strength of about 1110 MPa with a standard deviation of 110 MPa. The industry takes, the aerospace industry takes a random sample of 100 from the supplied lot.

Now it wants to find out whether this lot whatever the mean it has, does it satisfy, does it fall in the vicinity of 1110 MPa? So, it wants to, what it does, he finds that actually the average yield strength is 1129 MPa. Now, company has to take a decision whether it should accept the supply.

So, let us first understand in our terminology, what all has been given to us. Number one, mu 0 is what the average company has claimed, foundry has claimed. So, it is 1110 MPa sigma square variance is no, standard deviation is 110 MPa the sample of size 100 is taken. So, n is equal 100 and it has found the mean yield strength of 1129 MPa.

(Refer Slide Time: 17:38)



So, let us follow the classical steps.

$$\text{Let us fix } \alpha = 0.05$$

$$H_0 : \mu = 1110 \text{ MPa vs. } H_A : \mu \neq 1110 \text{ MPa}$$

$$\text{Statistic of interest is } \bar{X} = 1129$$

Let us fix the alpha at 5 percent, so it is alpha is 0.05. Null hypothesis is that the mean value should be 1110, and alternate hypothesis is we are taking only one sided alternate and we say that it has to be, it is not equal to 1110 MPa because it is 1129 MPa, well 10 and 29, how much difference does it make? We do not know, the statistic of interest is X bar, which takes a value of 1129 MPa, I am sorry, I forgot to write the unit. Let us write it down. It should be MPa, okay.

It is very important to realize that, I just thought that when we say that mu is equal to 1110 MPa, what we really mean is that, mean lies in the vicinity of 1110 MPa. The reason being it has a variance of 110, sorry standard deviation of 110 MPa. So, it can vary within that range.

And therefore, the question comes whether this is outside the range or not and that range is what we are trying to find and that is the range which is acceptance region, what we are looking for is the rejection region.

$$C = \{X_1,\ X_2,\ \ldots,\ X_n \mid \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > z\} \text{ when } H_0 \text{ is true}$$

So, here we take the next step, we say that the test statistic, which is $\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > Z;$ And therefore, this is the rejection region when null hypothesis is true therefore, you have written mu is equal to mu 0 that should be alpha which is 0.05. So, and Z is as we said before it is a parameter less distribution, the distributed variable it is distributed as normal (0,1) not even parameter less, it is a unit less because we have already divided MPa with MPa.

$$P\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > z\right) = \alpha = 0.05, as\ \ Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,\ 1)$$

$$P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \ z_{1-\alpha/2}\right) = \frac{\alpha}{2} = 0.025$$

So, it is a unit less, parameter less, random variable which has a normal 01 it is fully defined random variable. So, then probability of that Z greater than, we know that when my Z of small set sub 1 minus alpha by 2 has to be points 0.025. This has to be alpha by 2 and if you look into it, as we say, please, this is something which we must remember. So I am going to repeat it several times in 2 days and 3 other sessions this and other sessions. We are looking for Z value here so that this region is alpha, but we always calculate Z of any ‘a’ is such that minus infinity to Z of ‘a’ 1 over square root 2 pi exponential minus 1 half x squared dx is equal to ‘a’.

So it takes this region as its probability, this interval is this region and therefore this value,

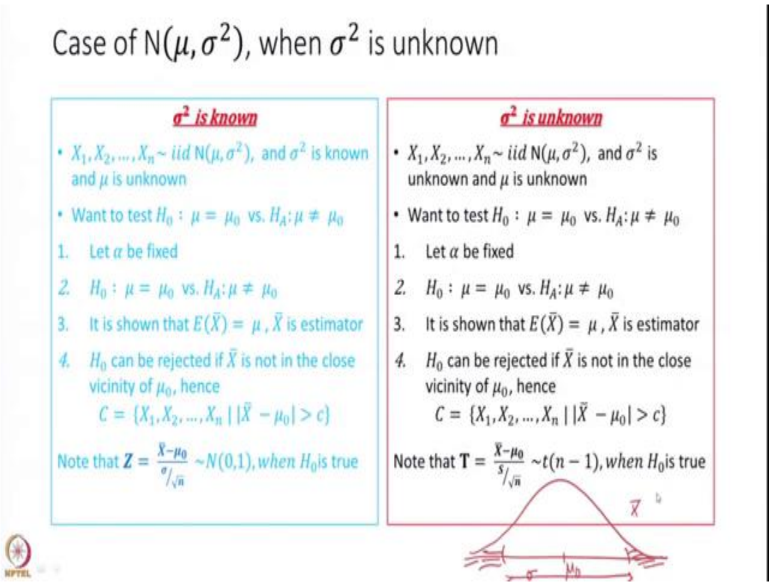
From Standard Normal table $z_{1-\alpha/2} = z_{0.975} = 1.96$ is the critical value

$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{1129-1110}{110}\sqrt{100} = 1.72 \text{ is not} > 1.96$$

we cannot reject the null hypothesis, and so the industry should accept the supplied lot. This is the decision. This is how the 6 steps are to be followed.

What is to be remembered is that up to this point you are talking about this value as a random variable and at this point in reality, it is actual value. Maybe, we should make a change here, we should instead of calling X bar we should call it a small x bar and this X bar may be replaced by a small x bar that would make the, that would bring more clarity, I believe. So, we find our final decision in this case is that we accept the lot.

(Refer Slide Time: 23:42)



Suppose sigma square is not known. Here I have shown both the cases parallel, we have already derived the 6 steps for the sigma known. You will find that they are going absolutely parallel when sigma square is unknown. Your X1, X2, X3, ... , Xn, is same,

$$X_1,\ X_2,\ \dots,\ X_n \sim iid\ \text{N}(\mu,\ \sigma^2),\ \text{ and } \sigma^2 \text{ is unknown and } \mu \text{ is unknown}$$

$$\text{Want to test } H_0: \mu = \mu_0 \text{ vs. } H_A: \mu \neq \mu_0$$

$$\text{Let } \alpha \text{ be fixed}$$

$$H_0: \mu = \mu_0 \text{ vs. } H_A: \mu \neq \mu_0$$

So, X bar is the estimator.

It is shown that $E(\bar{X}) = \mu$, $\bar{X}$ is estimator

H_0 can be rejected if $\bar{X}$ is not in the close vicinity of μ_0, hence

$$C = \{X_1, X_2, \ldots, X_n \mid |\bar{X} - \mu_0| > c\}$$

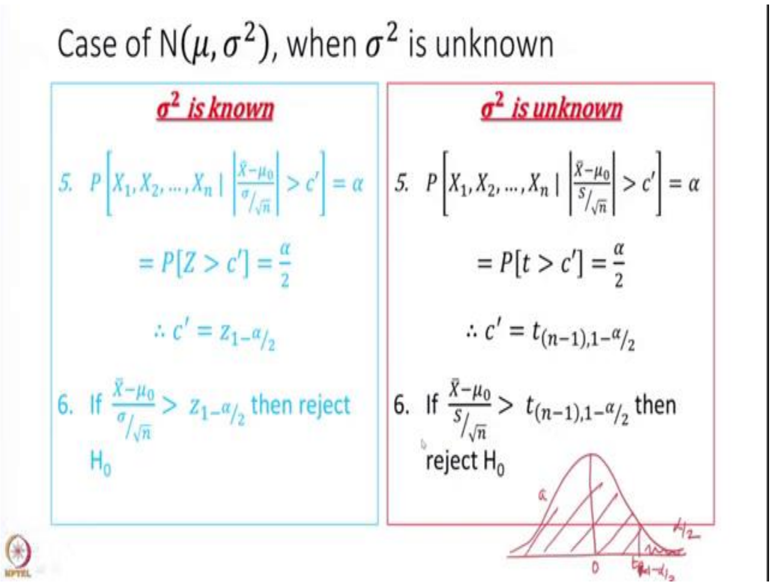
Once again we are saying that X bar has to be in the vicinity of mu naught, may not be a bad idea to repeat what we had said before. So, let us do it that we have this is mu naught and this is the distribution of the population and the variance is sigma, sorry standard deviation is sigma and then we are saying that the X bar has to be in some vicinity of mu naught so that we can say that if it is beyond if it is too large, it is not acceptable.

Please remember that we are accepting the fact that, it can never be dot mu 0. We are never saying that it should be dot mu 0, we are saying that it will fall in the vicinity of mu 0. And when it is very far away when it is too far away from mu 0, that is when we define our rejection region. So, here we have to define a rejection region that X bar minus mu is greater than some quantity c and therefore now, we have a test statistic just as we did, what we did? We neutralized, we removed the units and we brought a test statistic to a level which had a distribution without any unknown parameters.

So, here also what we find is, it is a T, which is equal to X bar minus mu 0. Instead of sigma we have a sample standard deviation divided by square root n. And we know that under normality,

$$\boldsymbol{T} = \frac{\bar{\boldsymbol{X}} - \boldsymbol{\mu_0}}{\boldsymbol{S}/\sqrt{\boldsymbol{n}}} \sim t(n-1), \; when \; H_0 \text{ is true.}$$

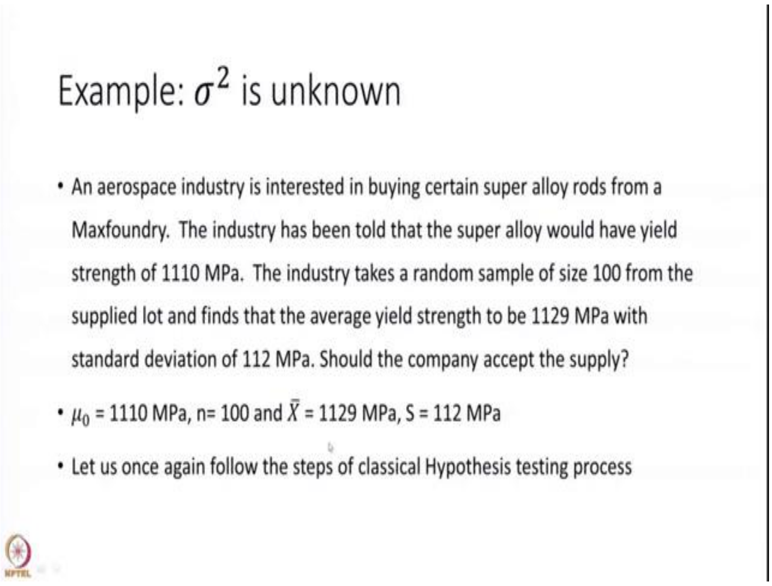
(Refer Slide Time: 26:39)



And therefore, on the similar line instead of a normal table, you will be looking into the t table. Please remember, that the t table, t distribution is also a symmetric distribution but it has a thicker tails. It does not have a thin tail like normal, here also the value 0 is in the centre. And therefore, here you are looking once again, if the t value is defined t of a is defined if this is a, then this is the probability define t a, then this is called a probability a. We would like to have this to be alpha by 2, then this has to be 1 minus alpha by 2, your a has to be 1 minus alpha by 2.

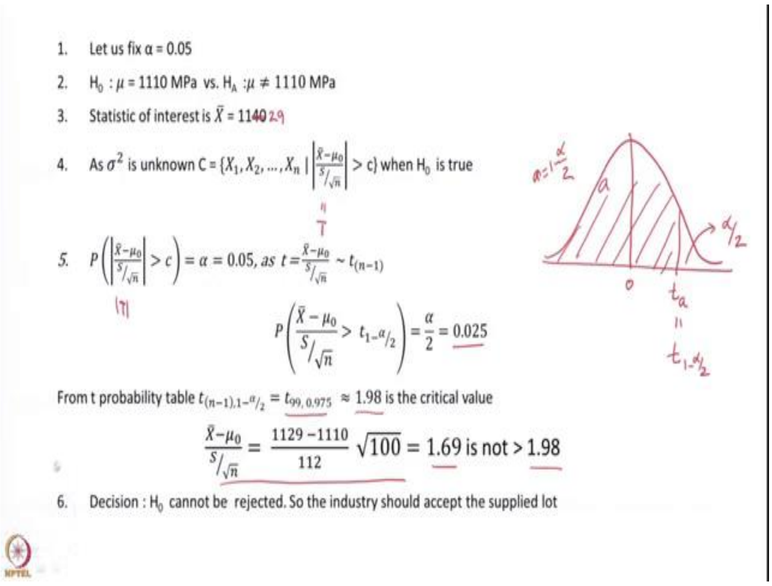
So, accordingly, it is shown here and we say that, if your t statistic which is X bar minus mu 0 divided by sample standard deviation by square root n is greater than p 1 minus alpha by 2 for the n minus 1 degrees of freedom of t distribution then you reject H 0.

(Refer Slide Time: 28:03)



Let us again take an example, the same example I am taking, now I am not saying that it has a variance of 110 MPa, we do not know the variance, 100 samples have been taken. The mean yield strength is found to be 1129 MPa and the standard deviation is found to be 112 MPa. Should the company accept the supply? So, here we have mu 0 is equal to 1110 MPa, which is what the company has claimed and n is 100, X bar is 1129 MPa and sample variance sample standard deviation is 112 MPa.

(Refer Slide Time: 28:58)



So, we follow the 6 steps,

1. Let us fix α = 0.05
2. H_0 : μ = 1110 MPa vs. H_A :$\mu \neq$ 1110 MPa
3. Statistic of interest is $\bar{X}$ = 1140
4. As σ^2 is unknown C = $\{X_1, X_2, \ldots, X_n \mid \left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > c\}$ when H_0 is true
5. $P\left(\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > c\right) = \alpha = 0.05, \, as \;\; t = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$

$$P\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > \; t_{1-\alpha/2}\right) = \frac{\alpha}{2} = 0.025$$

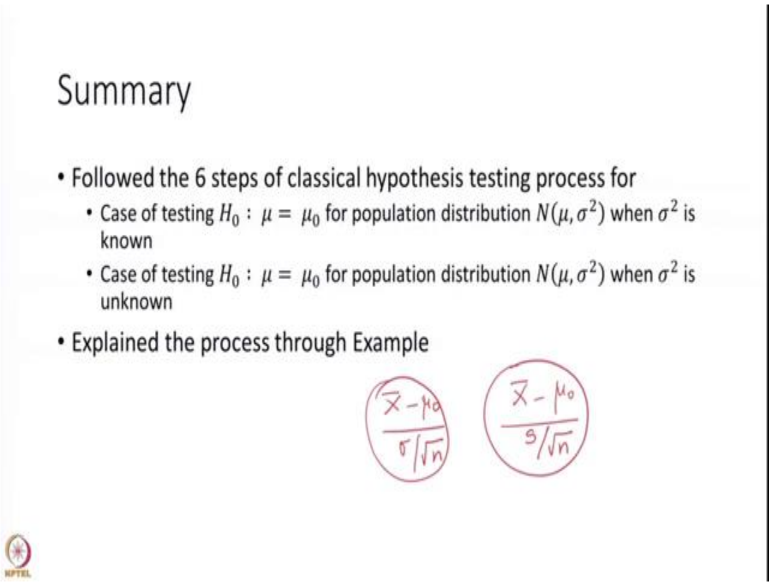
From t probability table $t_{(n-1),1-\alpha/2} = t_{99,\;0.975} \approx 1.98$ is the critical value

$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{1129-1110}{112}\sqrt{100} = 1.69$ is not > 1.98

6. Decision: H_0 cannot be rejected. So the industry should accept the supplied lot

So, again we cannot reject the hypothesis and therefore, we company has to accept, industry has to accept the lot.

(Refer Slide Time: 31:44)



So, let us summarize it. In this session, we explained two examples or two methods assuming normal population, we followed 6 steps of classical hypothesis testing process. And both the methods where sigma square is known and sigma square is unknown we gave an example to demonstrate what really we have to do. The few things to remember is that please follow the 6 steps one after the other to avoid any kind of confusion. Second thing, please notice that the underlying statistic is actually X bar minus mu 0 over sigma square root n or X bar minus mu 0 over sample standard deviation square root n.

Please remember that these are the two central statistic we are working with and they are very important. They do not really require a normal assumption. You can deviate from it, only the procedure changes. Thank you!