Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture - 23
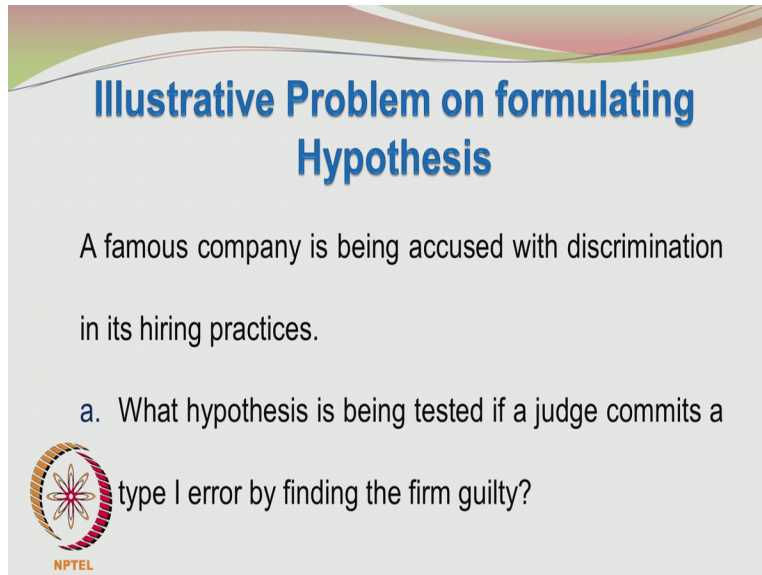**Hypothesis Testing - Part B**

(Refer Slide Time: 00:22)

So coming back to the errors in decision making, we have the table here, the first column says statistical decision, and the true state of null hypothesis is given in the second and third columns. So the true state of null hypothesis maybe H0 is true, H0 is false, the null hypothesis is represented by H0, the original base hypothesis. So when the statistical decision is do not reject H0, well you are implying accept H0 without seeing in so many words, H0 is true you are not rejecting H0 it is a correct decision nothing more has to be said.

H0 is false and you are saying do not reject H0 implying accept H0, then this is a type 2 error, we are letting a criminal go free. When the next statistical decision is reject H0, and H0 is true and you are rejecting H0, the person is innocent but you are sending him to jail, then you are committing a type 1 error. H0 is true, the statistical decision is rejecting H0, then this is a serious error the person is innocent but the person is being sent to jail, so it is a type 1 error.

H0 is false and you are rejecting H0 correct decision, they are inset the matter, so here we have type 1 error and type 2 error depending on whether you rejected H0 or did not reject H0.

(Refer Slide Time: 02:26)
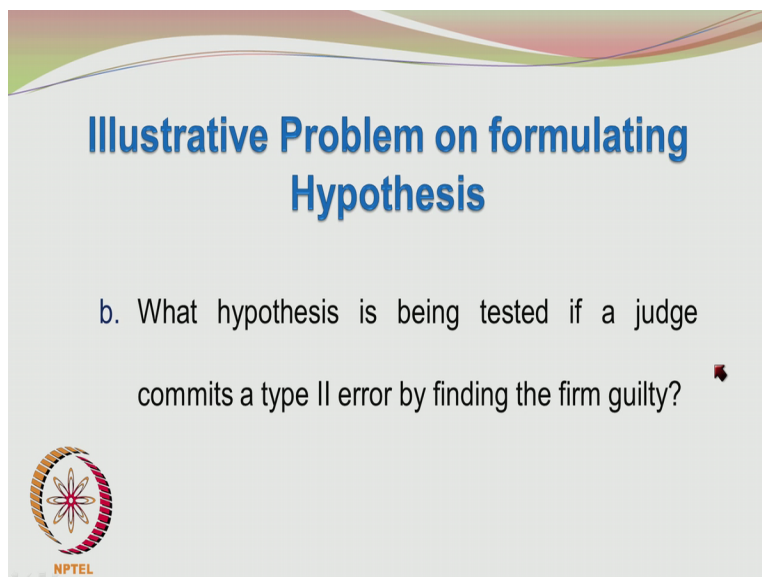
A famous company is being accused for being discriminatory when hiring, so what is the hypothesis that is being tested if a judge commits a type 1 error by finding the firm guilty, so the judge has committed a type 1 error that means he has rejected the null hypothesis. So what could have been the null hypothesis? The null hypothesis in that case would have been the company is fair in its hiring practices.

(Refer Slide Time: 03:18)

In the next situation what hypothesis is being tested if a judge commits a type 2 error by finding the firm guilty okay, so what would have been the original hypothesis statement if by finding the firm guilty the judge is committing the type 2 error. This is simple example and I would request you to think about it and then answer this particular subdivision, part a I already answered. So take a couple of minutes separately to find the answer it is quite straight forward.

So by now you should have got the answer that H0 would have been the firm is indeed discriminatory in its hiring practices, by accepting the null hypothesis the judge has committed a type 2 error by finding the firm as guilty. So the null hypothesis statement in this case is slightly debatable, because the status quo always would be the firm is fair in its hiring practices, so here the null hypothesis has been stated that the firm is discriminatory in its hiring practices okay.

That is not the general case, there may be very bad incidence of companies that are probably discriminatory, but that is not be used for trend. So the null hypothesis itself is subject to debate here, the null hypothesis for part b is the company is indeed discriminatory in its hiring practices, and the judge has committed a type 2 error by finding the firm guilty, that means he has started with the notion that the company was discriminatory, and he did not accept the alternative that the firm was innocent or being fair.
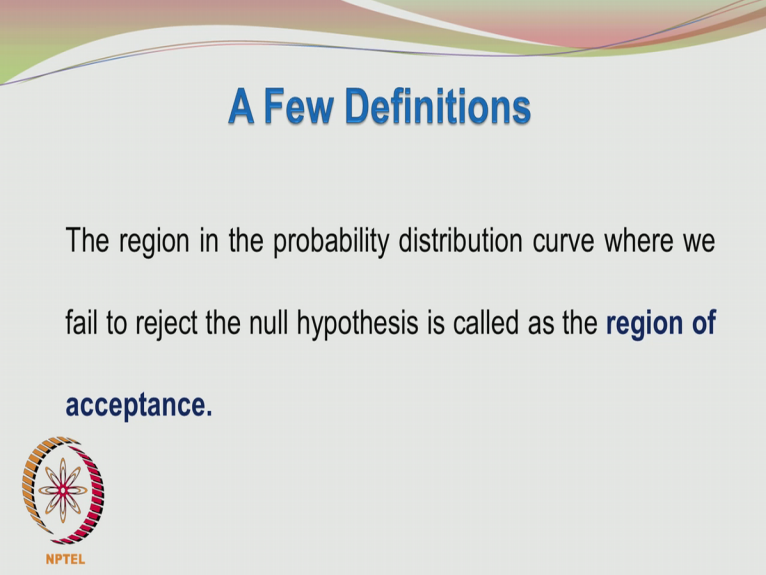
(Refer Slide Time: 05:58)

Illustrative Problem on formulating Hypothesis

a. H_o: Firm is fair in it's hiring practices

b. H_o: Firm is unfair in it's hiring practices

NPTEL

So part a, to summarize H0 firm is fair in its hiring practices. Part b, H0 firm is unfair in its hiring practices. This interesting example, I came across somewhere unfortunately I do not remember from which book or from which source I got this example, so I am unable to provide the reference for this example, I am sorry for that but it is a very nice example.

(Refer Slide Time: 06:36)

A few definitions are in order, the region in probability distribution curve where we fail to reject the null hypothesis is called as the region of acceptance. Well what are we saying now? We are having a sampling distribution of the means, suppose the sample came from a normal distribution, the sampling distribution of the means will also be normal. In other words, the random samples may have different values, when they are taken from a population.

Obviously, the same mean and same variance may not come for different random samples, and so there is a distribution of the sample means. If the population was normal with mean mu and variance sigma square, the sampling distribution of the mean will also be normal with mean mu and variance sigma squared/n, this we have seen repeatedly in the past lectures. Now the sampling distribution of the mean is normal, so we are having a normal distribution of the sample means.
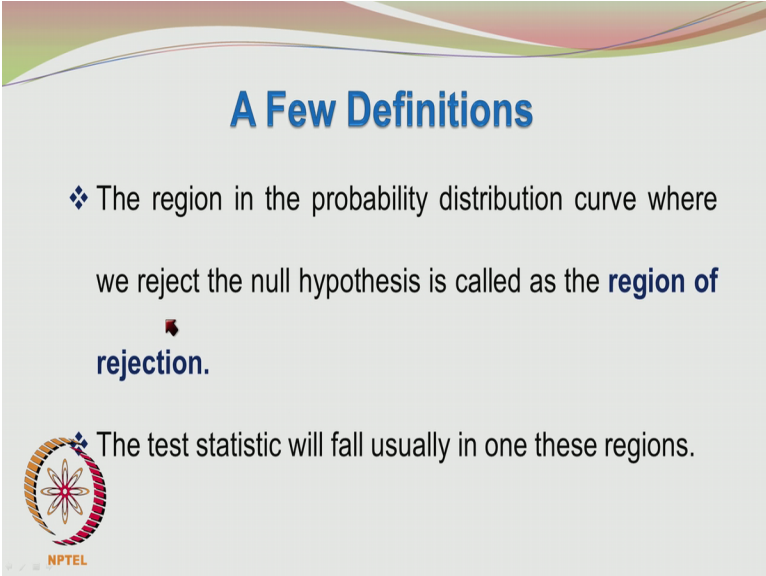
Now we say that if the sample mean value exceeds a particular number, then we can no longer accept the claim that it has come from a population with mean mu=mu 0. For example, the

random sample is giving a value of 90, then we cannot really accept that it has come from mean of 50 okay. Suppose the random sample mean value is 90, we cannot say that the sample came from a population with mean mu=50.

The probability of a random sample taking a value of 90 from a distribution of sample means of 50 is pretty low okay, it is pretty much non-existent because the variance of this sampling distribution of the means actually reduce, the sampling distribution of the means have a lower spread, because the variance is squared/n, so the spread has reduced. So there is even less chance that the probability of a random sample taking a value >90 could have ended come from a population with mean mu=50.

So this is what we have to locate and identify in the sampling distributions of the means, so the region in the probability distribution curve where we fail to reject the null hypothesis is called as the region of acceptance.

(Refer Slide Time: 10:19)

The region in the probability distribution curve where we reject the null hypothesis is called as the region of rejection. The test statistic will usually fall in one of these 2 regions, so we having 2 regions which are complementary, one is the region of acceptance, and one is the region of rejection. Acceptance of what? And rejections of what? The acceptance of the null hypothesis, and the rejection of the null hypothesis.

The region in the probability distribution curve where we accept the null hypothesis is called as the region of acceptance, and in the region in the probability distribution curve where we reject the null hypothesis we called as the region of rejection. So we have to see in the distribution of the sample means where our sample mean values are actually lying, then we have to compare that with the critical value. So if the sample mean lies beyond this critical value then we reject the null hypothesis.

If the sample mean is lying within the critical value, we accept the null hypothesis. Well some of you may still find it difficult to understand, what I am talking about but soon we will see example, and we will also see the probability distribution curve illustrated, and then things will become clear and fall into place. The boundary demarcating the acceptance and the rejection region is denoted as the critical value.

(Refer Slide Time: 12:25)
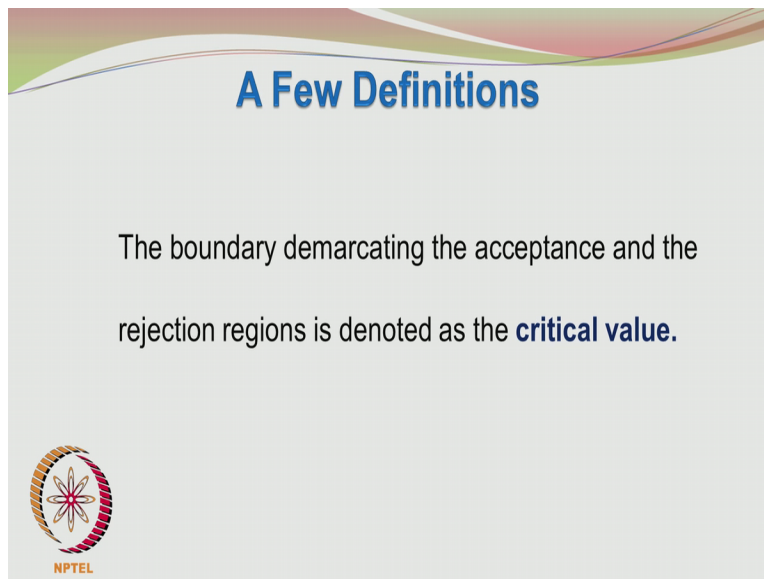
So essentially we are having the region of acceptance and the region of rejection, and the boundary between these 2 is called as the critical value. So the critical value is the value of the sample mean, which divides these 2 regions okay.

(Refer Slide Time: 12:39)

When to reject H_o?

We reject H_o in favor of H_1 if the test statistic falls in the critical region and fail to reject H_o otherwise.

We reject H0 in favor of H1, if the test statistic falls in the critical region, and fail to reject H0 otherwise. So if the test statistic for example x bar is such that it falls in the critical region, then we reject H0. For example, be said critical value of 68, and our test statistic gives a sample mean of 90, so 90>68 and so the sample value is lying well in the rejection region, and so we can reject the null hypothesis. However, if the sample mean value is only 60, then we cannot reject the null hypothesis, we have to accept the null hypothesis.

(Refer Slide Time: 14:02)

Illustration

Let us assume that a shipment received from the vendors is accepted on the design assurance that the mean impurity is only **0.20** ppm. Here the population mean is **0.20** ppm. **10** specimens are selected at random and checked for mean impurity.
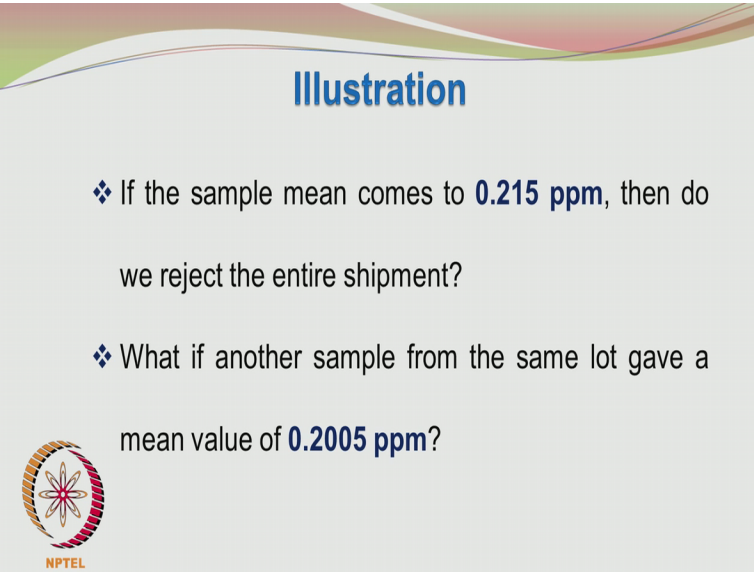
I will give another example, let us assume that a shipment received from the vendors is accepted on the design assurance that the mean impurity is only 0.2 ppm, here the population mean is 0.2 ppm, 10 specimens are selected at random and checked for mean impurity. Obviously, we would

like to take more specimens okay, but we are prevented from doing so, maybe the entire shipment is very valuable one, and there may be only 100 pieces shipped to you.

And you even taking 10% out of them for testing maybe too much too expensive, and these specimens also may be subject to the destructive testing to find the mean impurity, so that they cannot be used again. So our sample size though it is preferable to have a large sample size, we may at times we constrained to take smaller ones. So let us say that we are taking only 10 specimens.

(Refer Slide Time: 15:17)

Illustration

❖ If the sample mean comes to **0.215 ppm**, then do we reject the entire shipment?

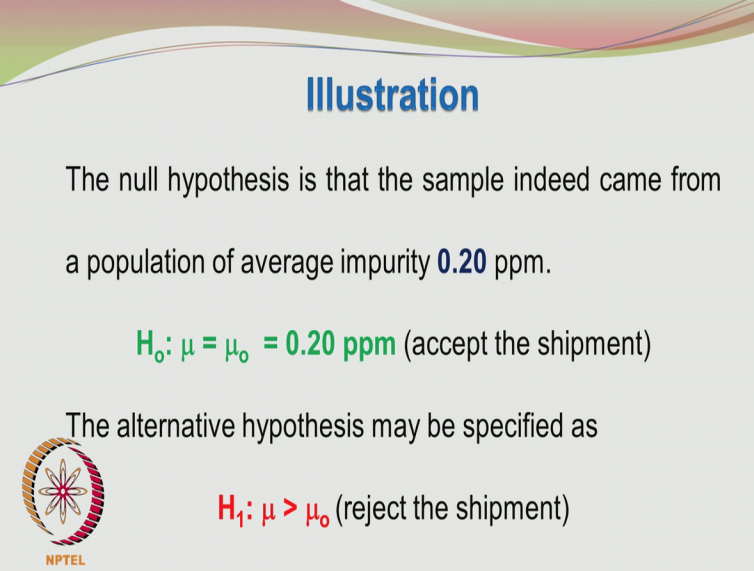❖ What if another sample from the same lot gave a mean value of **0.2005 ppm**?

Well life is always not fair, in the sense even though we say that life is full of random phenomena, we noticed the random phenomena more closely when the situations are not favorable to us, random phenomenal leading to unfavorable outcomes are more noticed by us than those random phenomena, which are favorable to us, we probably do not even notice them. Even though the probability of occurrence of the random phenomena are all the same.

We observed those which are not favorable to us a bit more frequently, anyway enough of this philosophical discourse, we will come back to the problem at hand. So what I am trying to say is the sample mean could have well been 0.19 ppm, but we do not get that we get the sample mean as 0.215 ppm. So let us be consistent with the units, mention the units all the time okay, so I will just add the units here.

It is important that the units are added not only for the sample mean, the population mean. But also for the sample standard deviation, population standard deviation, because you should appreciate that both the mean and standard deviation carry the same units, and it also helps us to relate both of them on an equal footing. If the sample mean comes to 0.215 ppm, then do we get excited angry and reject the entire shipment. And what would have been our thoughts and feelings, if another sample from the same lot we have a mean value of only 0.2005 ppm.

(Refer Slide Time: 17:43)

Illustration

The null hypothesis is that the sample indeed came from a population of average impurity **0.20** ppm.

$H_o: \mu = \mu_o = 0.20$ ppm (accept the shipment)

The alternative hypothesis may be specified as

$H_1: \mu > \mu_o$ (reject the shipment)

So the null hypothesis is that the sample indeed came from the population of average impurity 0.2 ppm, we want to give the benefit of the doubt to the vendor, and we say yes the sample has indeed come from a population with mean impurity of 0.2 ppm as was stipulated in the contractual statements. The null hypothesis H0 mu=mu 0=0.2 ppm, inclination is to accept the shipment.

The alternate hypothesis is the product that is being supplied are coming from population with mean impurity >0.2 ppm, so since the impurity is higher than what was agreed upon in the contract, we reject the shipment. So a decision making is involved here, and we have to make sure that our decision making is not arbitrary, and we have to do it in a fair manner.

(Refer Slide Time: 18:57)

Illustration

H_o: $\mu = \mu_o$ = 0.20 ppm

H_1: $\mu > \mu_o$

You may want to actually ask if μ_o = **0.20 ppm**

what is the probability of picking up a sample of **10** units

with mean impurity of **0.215** ppm?

NPTEL

So H0 mu=mu 0=0.2 ppm, and H1 mu >mu 0 or mu >0.2 ppm, so the question you have to ask is if mu 0=0.2 ppm, what is the probability of picking up a sample of 10 units with mean impurity of 0.215 ppm? If the probability of picking up a sample with a mean impurity of 0.215 ppm is pretty high, because of random phenomena it is highly probable that you may take a sample of mean impurities 0.215 ppm from the sampling distribution of means which are centered around the 0.2 ppm.

Then you give the benefit of the doubt to the vendor, and accept the shipment okay. So our aim now is to find the probability of picking up a random sample, whose mean is 0.215 ppm, if this random sample came from our distribution of sampling means, which are centered at 0.20 ppm okay.

(Refer Slide Time: 20:20)

Illustration 1

If this probability is rather high, you may decide to accept

the shipment.

If it is very low, then reject the shipment.

So if the probability of picking up the sample with the mean impurity of 0.215 ppm is very low, then you have to reject the shipment.

(Refer Slide Time: 20:40)

Illustrative Problem 1 on Inferential Statistics

$H_o: \mu = \mu_o = 0.20$ ppm
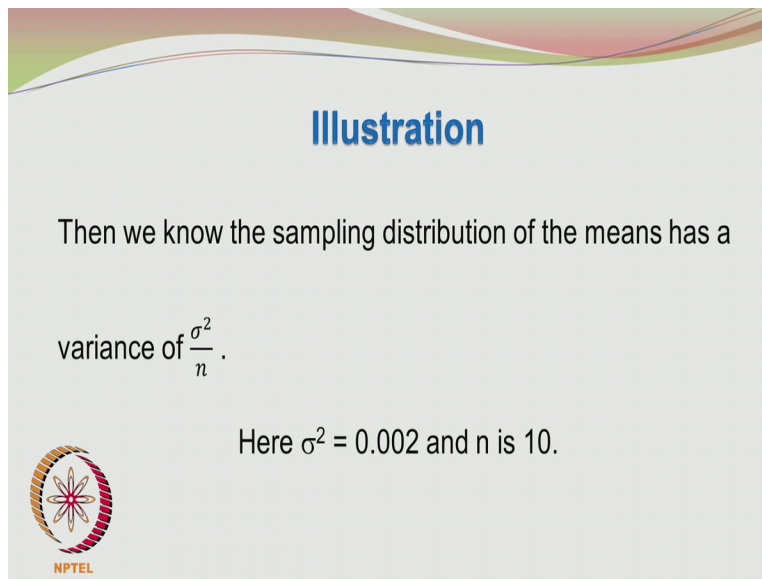
$H_1: \mu > \mu_o$

The question is how to find this probability. Assume that the population variance (σ^2) is known as **0.002 (ppm²).**

So the question is how do we find the probability? Another important information that is available to us is the population variance sigma squared 0.002 ppm squared. Well here now we are given the value of sigma squared, and we are speculating the value of mu 0, so we can use both mu and sigma squared in the appropriate probability distribution okay, and it is not as if we are knowing the value of mu and sigma squared.

If the values of mu and sigma squared were known for certainty, then we do not have to do anything, but we are now speculating the value of mu as mu 0=0.2 ppm. And the variance is given to us, usually the variance is also not known to us, but if the process manufacturer says that impurity levels in the product is subject to a variability of 0.002 ppm squared, then we can take that as sigma squared itself, so this is 1 assumption. So we have speculated value of mu 0, and we have the population variance sigma squared given to be 0.002 ppm squared.

(Refer Slide Time: 22:10)

We know the sampling distribution of the means as a variance of sigma square/n, the sampling distribution of the mean is centered around the population parameter mu, but it is having a lesser spread because the variance of the sampling distribution of the means is given by sigma square/n. Here, sigma square=0.002 ppm squared and n=10.

(Refer Slide Time: 22:35)

Illustration

Let us assume that the concentration of impurities in the parts is normal.

Construct the standard normal variable from the sample mean and find the probability from the tables.

So if it is further assumed that the distribution of impurities in the parts is normal okay, normal assumption is not a very bad assumption okay, deviations from normality are not that serious okay. So we will assume for the sake of illustration that the distribution of impurities in the parts is normal rather than saying distribution of impurities, I should have said the concentration of impurities okay, the distribution of impurities seems to give a different meaning.

So let me change it to concentration of impurities right, so what we do is we are having a speculated mean value of 0.2 ppm, we are having a sigma squared of 0.002 ppm squared, so we essentially have the probability distribution of the means, and we also know the sample size. But in order to use the probability tables okay, we need to construct the standard normal distribution, and we know the transformation, if x bar is known, mu is known, sigma squared/n is known.

We can convert this into the standard normal variable by the transformation z=x bar-mu/sigma square/n, well that is incorrect because x bar and mu have units of ppm, whereas sigma squared will be having units of ppm squared, so you cannot divide a term in the numerator having units of ppm, with a term in the denominator having units of ppm squared. So the concentration of impurities is distributed normally.
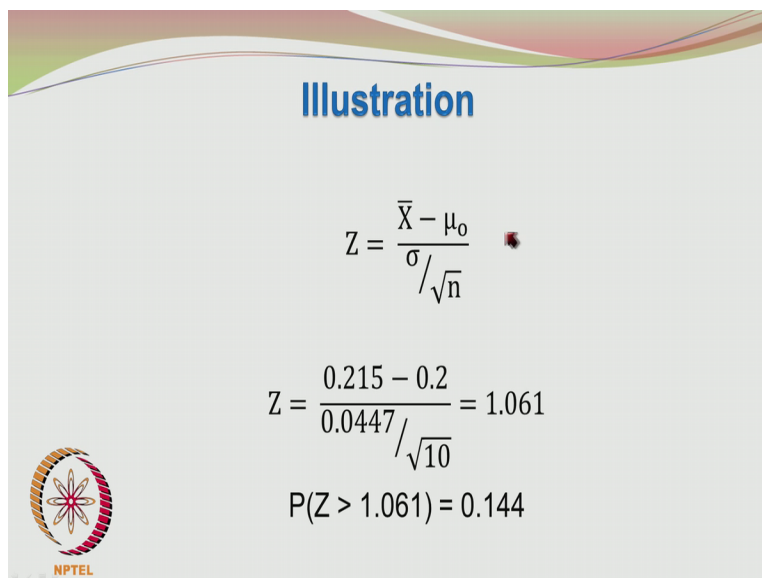
Now we know x bar, we know mu, we know sigma squared/n okay, we have to find the standard normal variable. Why do we need to find the standard normal variable? Reason is we are able to

construct the normal distribution with x bar, mu and the sigma/root n, the normal distribution can be constructed with the parameters mu and standard deviation sigma/root n. But for finding the probabilities we do not have tables for different normal distribution centered at different mean value.

We have only one table or chart where the normal distribution is centered at the value of 0, and having a variance of 1, so we have to convert our present normal distribution into a standard normal form. So what we do is we use the transformation z=x bar-mu/sigma/root n, please do not make the mistake of putting z=x bar-mu whole divided by sigma squared/n that is not correct, because x bar and mu have units of ppm, whereas sigma squared is having a units of ppm squared.

You cannot divide a term in the numerator with units of ppm, with the term in the denominator with the units of ppm square. So even if you are suddenly unsure which should be the correct term to use in the denominator, these kind of dimensional analysis will help you out. Anyway so the standard normal variable z is given in terms of x bar-mu/sigma/root n, the sample size of course is dimensionless. So we have a standard normal variable z=x bar-mu whole divided by sigma/root n. I went a bit fast because all these things should be now very familiar to us.

(Refer Slide Time: 26:48)

Illustration

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$Z = \frac{0.215 - 0.2}{0.0447/\sqrt{10}} = 1.061$$

$$P(Z > 1.061) = 0.144$$

So we have the standard normal variable x bar-mu 0/sigma/root n, so z=0.215-0.2/0.0447/root 10, how did this 0.0447 come I took the square root of the variance, the variance was given to be 0.002 ppm squared, so I took square root of 0.002 ppm squared I got 0.0447 ppm/root 10 the sample size is 10, so there is no mistake here. And here we have to find what is the probability of z>1.061, here the problem statement is quite important.

Given the fact that the population mean value is 0.2 ppm, so the sampling distribution of the means will also have a mean value of 0.2 ppm, what is the probability of picking up a random sample with mean impurity level of 0.215 ppm or higher, so this is very important. What is the probability of picking up a random sample with mean impurity of 0.215 ppm or higher, when the population mean value is 0.2 ppm.

So with this we convert the x bar into z using the transformation x bar-mu 0/sigma/root n, here we use mu 0 the value speculated or postulated in the null hypothesis, so we have z=0.215-0.2/0.0447/root 10 which comes to 1.061, so we find probability of z>1.061 as 0.144. This in my opinion is a pretty high chance okay, or a very high probability, well some of you may say that what is so special about 0.144, it looks like low enough probability for mean.

It depends upon your strictness level. So again it is a matter of application, but the judge sitting in his court may say that if there is a 14 or 15% chance that the person is innocent, then he will of course free him, so what is the typical limit? The probability where you decide against the null hypothesis is usually 0.05 or lower, so if there is a very small probability that the random sample could have come from a population of mean 0.2 ppm, then you reject the null hypothesis.

Here, it is pretty high at 0.144, the typically used value is 0.05, and here it is higher than 0.05, so we have to accept the null hypothesis, and hence accept the shipment.

(Refer Slide Time: 30:32)

Distribution Plot
Normal, Mean=0.2, StDev=0.0141412
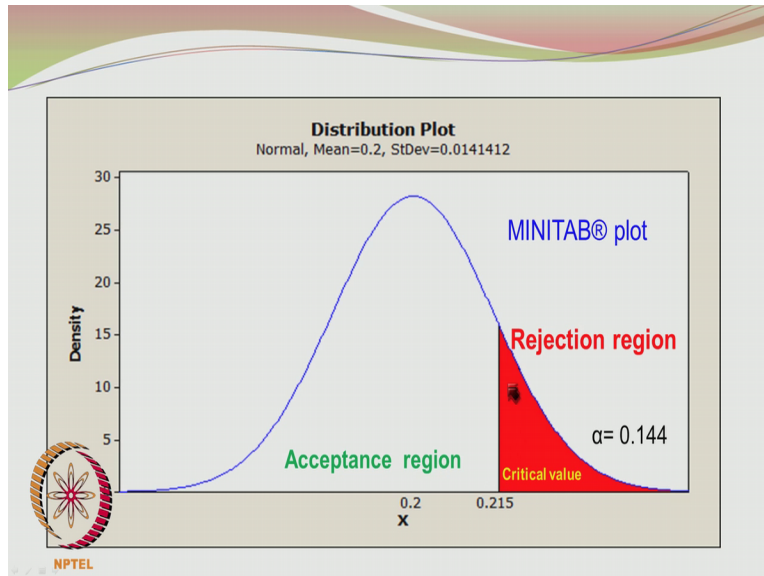
So now this is very nicely brought out in this MINITAB plot, here you are having the normal distribution for the sample means okay, so this is the normal distribution for the sample means. Well you have speculating on the mean value, and you are also knowing the population variance, so you can also construct the population normal distribution with this data, but that is not really necessary because we are going to use the sample to make the conclusions.
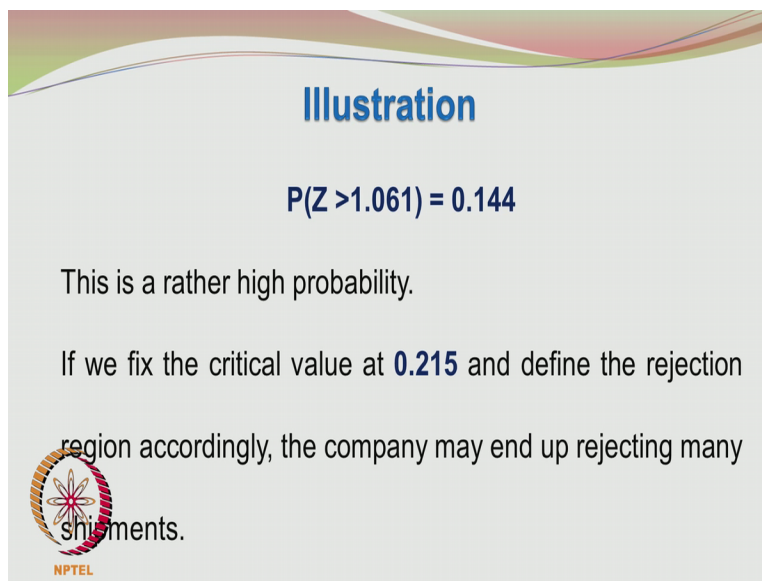
Since we are using the sample, we construct the sampling distribution of the means using the given data, so please check this out to see whether sigma squared/n and then taking square root of that to get sigma/root n, whether that value comes to 0.01414. So anyway we have the sampling distribution of the means with mean value of 0.2 ppm, and sigma/root n value of 0.01414 ppm, so this is the distribution and our actual sample indicated a mean value of 0.215 ppm.

So if you recollect that number, so you are having a value of 0.215 ppm, so this 0.215 ppm lies here, and the probability of taking random samples with mean impurity value of 0.215 ppm or higher is in this red portion, and that is coming as 0.144. So this 0.144 is a pretty high probability okay, so it would be very difficult to defend the rejection of the null hypothesis, anyway so 0.215 ppm is the critical value, and the region where the null hypothesis is accepted is the acceptance region.

So any impurity level below 0.215 ppm would be accepted if 0.215 ppm was kept as the critical value okay, and the regions beyond this critical value of 0.215 ppm is called as the rejection region. Now if you say 0.215 ppm is pretty harsh, then you can of course keep your critical value at let us say 0.22 ppm or 0.23 ppm, there is a small typo in this graph, let me see if I can correct it. So going back if 0.215 was set as the critical value, the area under the curve beyond the critical value is 0.144, and that becomes the rejection region.

And the one below the rejection region is called as the acceptance region, the critical value increases to let us say 0.24 ppm it may come somewhere here, and when you increase the value of the critical value or if you increase the critical value, then you can notice that the rejection region will actually shrink and the acceptance region will expand okay. So if you had set a critical value as 0.24 ppm, then the mean impurity of 0.215 ppm would have lied in the acceptance region, and you would have accepted the shipment.

(Refer Slide Time: 34:53)

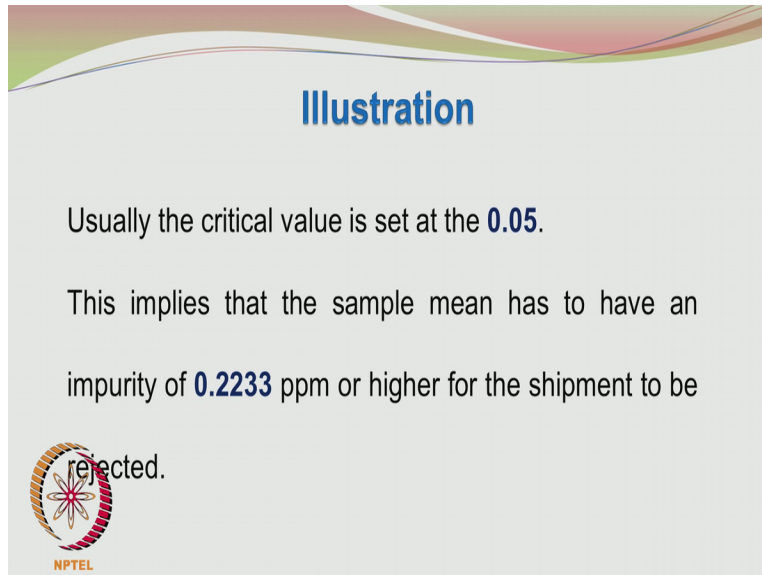So this is rather a high probability 0.144, and if you fix the critical value at 0.215 ppm, and define the rejection accordingly, the company may end up rejecting shipments okay that is probably not fair to the supplier and detrimental to the company in the long run okay. So if the probability value had been <=0.05, then the company would have been more justified in rejecting the shipment, rejecting the shipment at 0.144 probability is pretty harsh.

So this is the kind of quantitative arguments, you may present before the company to justify your decision making process.

(Refer Slide Time: 35:56)

So the critical value is set such that the probability of z exceeding this critical value is 0.05, so that is the real meaning of this sentence, this is the actual meaning of this sentence. Usually, the critical value is set such that the probability of the standard normal variable equaling or exceeding this value is only 0.05. So if you had set an alpha value of 0.05, then the corresponding sample impurity would have been 0.2233 ppm.

So only if the mean impurity in this sample was 0.2233 ppm, you could have been justified in rejecting the shipment okay. Well it may be difficult for you to measure 0.2233 ppm so accurately, this is a number just thrown up by the spreadsheet, we will continue in the next lecture.