Quantitative Methods in Chemistry
Prof. Dr. Aasheesh Srivastava, Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research-Bhopal

Lecture-15
Hypothesis testing and Finding Outliers Part 01

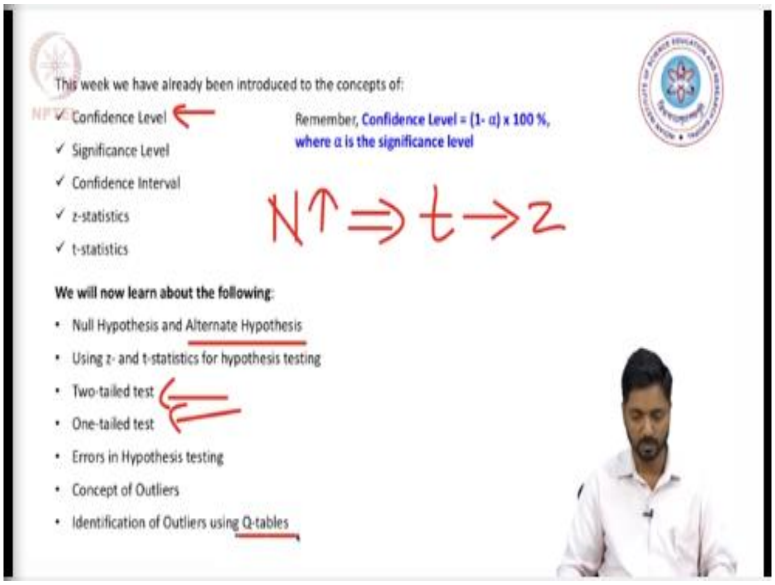
(Refer Slide Time: 00:28)

Quantitative Methods in Chemistry

Week 4, Lecture 2

Hello and welcome back to this NPTEL course titled quantitative methods in chemistry. This is week 4 and now we will be starting lecture 2 of this week.

(Refer Slide Time: 00:41)



This week we have already been introduced to the concepts of:

- ✓ Confidence Level
- ✓ Significance Level
- ✓ Confidence Interval
- ✓ z-statistics
- ✓ t-statistics

Remember, Confidence Level = (1- α) x 100 %, where α is the significance level

$N\uparrow \Rightarrow t \rightarrow z$

We will now learn about the following:

- Null Hypothesis and Alternate Hypothesis
- Using z- and t-statistics for hypothesis testing
- Two-tailed test
- One-tailed test
- Errors in Hypothesis testing
- Concept of Outliers
- Identification of Outliers using Q-tables

So for in this week we have already been introduce the concepts of confidence level, significance. And what we understand is that the confidence level is essentially a probability of finding the population mean around the sample mean. Similarly we have the concept of significance level which is a concept that tells us that how probable is a result being outside the confidence interval.

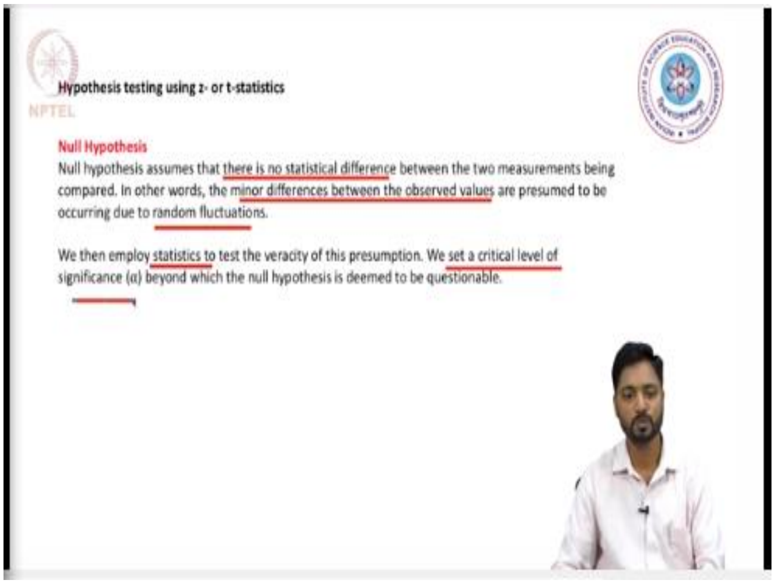
Now how do we define confidence interval, confidence interval is the interval around which the sample mean is expected to lie with a certain probability. So we also got introduced to the concepts of z and t statistics. And we got to know that when we are having a very good estimate of the population standard deviation. Then we can use the z-statistics however for small samples for which the standard deviation, population standard deviation is not accurately known.

In those cases we can use the t statistics and we also understood that when our number of readings for a sample is large than the t-statistics tends to the z-statistics. In other words when N value increases then the t statistics the z statistics. So with this basis now let us start discussing what we will be covering in this lecture 2, we will be talking about the null hypothesis and the alternate hypothesis.

And we will see how z and t statistics can be used to test the hypothesis in our hand. And we will also understand how the alternate hypothesis dictates which of the 2 test which is two tail test or the 1 tail test is to be applied. And as we progress we will also understand how errors in hypothesis testing come about and what is the importance and significance of these errors in terms of the inferences that we make.

Finally we will get introduced to the concept of outliers and we will utilize the Q tables for identifying the outliers in our dataset.

(Refer Slide Time: 03:50)



Hypothesis testing using z- or t-statistics

NPTEL

Null Hypothesis

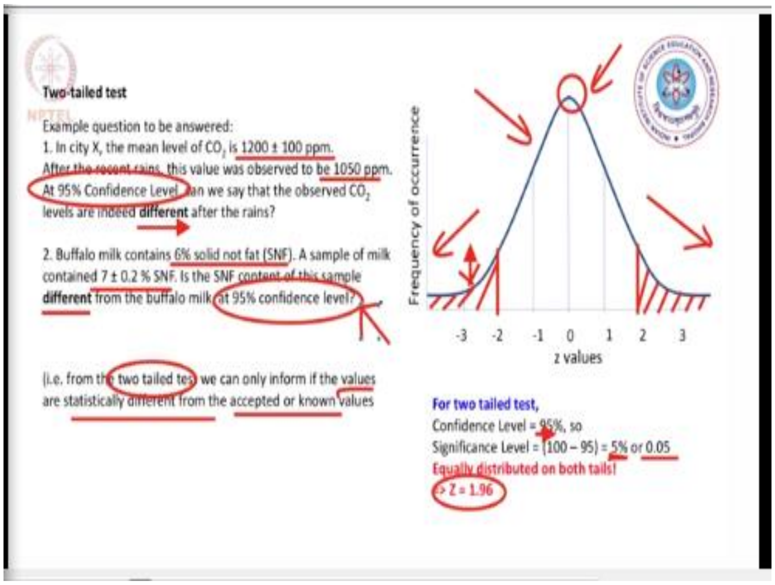
Null hypothesis assumes that there is no statistical difference between the two measurements being compared. In other words, the minor differences between the observed values are presumed to be occurring due to random fluctuations.

We then employ statistics to test the veracity of this presumption. We set a critical level of significance (α) beyond which the null hypothesis is deemed to be questionable.

So moving ahead let us first describe what we understand by null hypothesis, null hypothesis is a presumption which says that there is no statistical significance between 2 measurements being compared. In other words any minor differences between the observed values is presume to come due to random fluctuations and is not in real. And this presumption of null hypothesis is then tested using statistical tools such as the z test and the t test.

And we set a critical level of significance alpha beyond which the null hypothesis is presumed to be questionable and cannot be rely it upon.

(Refer Slide Time: 04:51)



Two-tailed test

NPTEL

Example question to be answered:

1. In city X, the mean level of CO_2 is 1200 ± 100 ppm. After the recent rains, this value was observed to be 1050 ppm. At 95% Confidence Level, can we say that the observed CO_2 levels are indeed **different** after the rains?

2. Buffalo milk contains 6% solid not fat (SNF). A sample of milk contained 7 ± 0.2 % SNF. Is the SNF content of this sample **different** from the buffalo milk at 95% confidence level?

(i.e. from the two tailed test we can only inform if the values are statistically different from the accepted or known values

For two tailed test,

Confidence Level = 95%, so

Significance Level = (100 – 95) = 5% or 0.05

Equally distributed on both tails!

=> Z = 1.96

And now let us move to the concept of the two tail test which is a test which is most routinely utilized for comparing 2 datasets. So let us understand this concept by taking 2 examples, the first example is that of a certain city in which the mean level of carbon dioxide is listed as 1200 + - 100 ppm. While after say a recent spell of rains this value was fond to be 1050 ppm. Now the question that we have at our hand is at 95% confidence level can be say that the observed value of carbon dioxide levels.

After the rains is indeed different from the commonly observed value of 1200 + - 100 ppm. Another question in the same line could be that of a buffalo milk sample, where we know that, typically buffalo milk should contain about 6% of solid not fat. But this sample that we are dealing with contains 7 + - 0.2% solid not fat. Now the question that we want to answer is that is the SNF content of the sample at our hand is different from the actual buffalo milk at 95% confidence level.

So, I want you to take note of 2 points, one is that we test the hypothesis at a particular confidence level. In this case as well as in this case, this number is 95% confidence level. And another thing is that we are being asked only to tell whether the reading at hand is indeed different from an accurate value or a population mean value. So, in those cases our question is to only address whether these numbers are different or not.

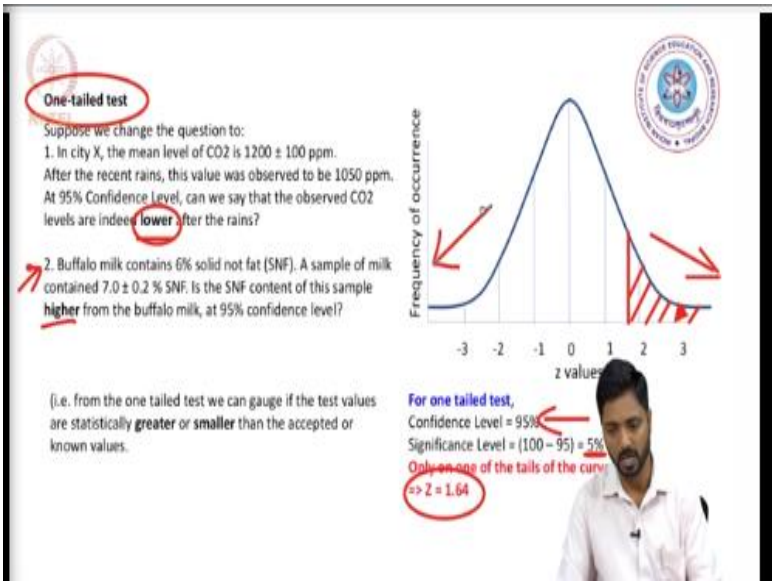
So, in such cases we will apply the two tail test, because reading can be either low or it can be higher than the mean value of a certain population. And as we have already discussed that are randomly selected sample from a large population will follow Gaussian profile and the mean value will be the most probable value as well. So, as we move away from this mean value the probability of occurrence of a certain reading reduces significant.

So, in nutshell if we have to tell whether the values being compared are statistically different from accepted or known values then we apply the two tail test. And for the two tail test at 95% confidence level what we imply is obviously the significance level is 5% which is nothing but 100 - 95 or if we write it in terms of decimal 6.05. But this 5% significance region is distributed evenly on the two tails of the Gaussian profile shown here.

Now, when that happens essentially we are talking about a z value of + - 1.96. So the region that I am going to highlight now, this whole region which has been highlighted with the red lines is to be considered as the region of significant. So, any reading that lies in this region will be considered statistically significantly different from the mean value which is shown at the peak of the Gaussian profile.

And we state our results at a certain confidence level. So 2 points to remember from this slide are that we are reporting our hypothesis testing at a certain confidence level. Typically 95% but it can be 99% or 99.9% depending upon how accurately we wish to report our findings. And another important point to be noted is that a two tail test can be applied only when the question at hand is whether the readings being compared are different or not.

(Refer Slide Time: 10:28)



One-tailed test

Suppose we change the question to:

1. In city X, the mean level of CO2 is 1200 ± 100 ppm.
After the recent rains, this value was observed to be 1050 ppm.
At 95% Confidence Level, can we say that the observed CO2 levels are indeed **lower** after the rains?

2. Buffalo milk contains 6% solid not fat (SNF). A sample of milk contained 7.0 ± 0.2 % SNF. Is the SNF content of this sample **higher** from the buffalo milk, at 95% confidence level?

(i.e. from the one tailed test we can gauge if the test values are statistically **greater** or **smaller** than the accepted or known values.

For one tailed test,
Confidence Level = 95%
Significance Level = (100 – 95) = 5%
Only on one of the tails of the curv
=> Z = 1.64

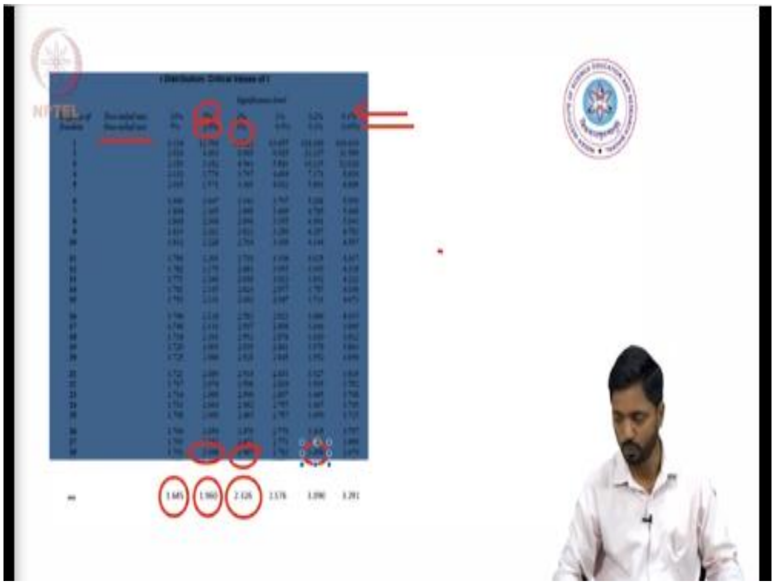
Now, if we rephrase this question as to whether the readings are lower or higher, then in those cases we are concerned only with one of the two tails of the Gaussian profile. So, in those cases, if we want to tell whether a measurement is lower than the population standard deviation, population mean, in those cases we will be using the left tail of the Gaussian profile. However, if the question at hand for example in the question 2 highlighted here, the question at hand is whether the reading is higher.

And when we are dealing with that we are going to be concerned with only the right hand side of the Gaussian profile or the right tail of the Gaussian profile. So, all the significant region will be lying at one of the tails and that is why this test is known as the one tailed test. And when we talk about one tailed test at 95% confidence level, we imply again the significance level of 5% being lying on only one side of the tail.

So, for example for the question number 2 where the buffalo male sample is to be compared with a standard sample with 6% solid not fat. In this case, we need to only tell whether at say 5% significance level the reading is different from the standard or not. So, our significance region is restricted to only one side and that implies that we use a different z value which is 1.64 in this case.

Now if the value to be compared is supposed to be higher than a population mean then we use the z value of + 1.64 and that becomes our critical z value. Similarly, if we are concerned with whether the reading is lower than the population mean, then we use a z critical value of - 1.64 being concerned with only the left hand side of the tail. So, I hope with these examples, it is clear which of the 2 tests are to be employed and how is what we will see now.

(Refer Slide Time: 13:45)

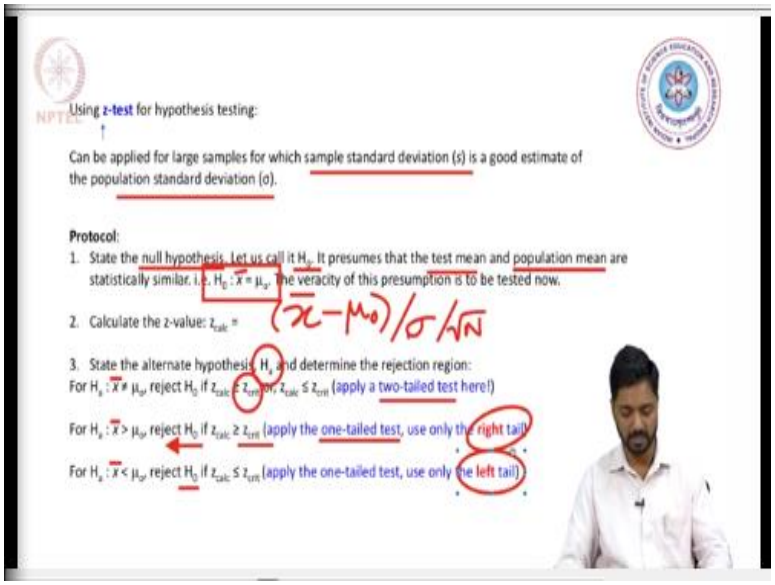


t Distribution: Critical Values of t

∞ 1.645 1.960 2.326 2.576 3.090 3.291

So, before that I want to reemphasize that our t tables already incorporate this point by giving us the significant levels for a two tail test and for a one tailed test. And what you will see is that for a one tail test, the significance value is half that for a two tail test ok. So, coming to 95% confidence level if we are applying the two tail test, then the z value was 1.96 while if the test being applied was only a one tail test in the z value was different and it was 1.645.

(Refer Slide Time: 14:49)



Using **z-test** for hypothesis testing:

Can be applied for large samples for which sample standard deviation (*s*) is a good estimate of the population standard deviation (σ).

Protocol:

1. State the null hypothesis. Let us call it H_0. It presumes that the test mean and population mean are statistically similar. i.e. $H_0 : \bar{x} = \mu_0$. The veracity of this presumption is to be tested now.
2. Calculate the z-value: $z_{calc} =$ $(\bar{x}-\mu_0)/\sigma/\sqrt{N}$
3. State the alternate hypothesis, H_a and determine the rejection region:

For $H_a : \bar{x} \neq \mu_0$, reject H_0 if $z_{calc} \geq z_{crit}$ or, $z_{calc} \leq z_{crit}$ (apply a two-tailed test here!)

For $H_a : \bar{x} > \mu_0$, reject H_0 if $z_{calc} \geq z_{crit}$ (apply the one-tailed test, use only the right tail)

For $H_a : \bar{x} < \mu_0$, reject H_0 if $z_{calc} \leq z_{crit}$ (apply the one-tailed test, use only the left tail)

Now, let us see how z test can be applied for hypothesis testing. So, it has to be clearly understood that when the sample standard deviation is a good estimate of the population standard deviation. That is when the z test can be applied and protocol for applying the z test is, that we first state the null hypothesis. And null hypothesis is represented by H 0 and of course it presumes that the test mean and the population mean are statistically similar.

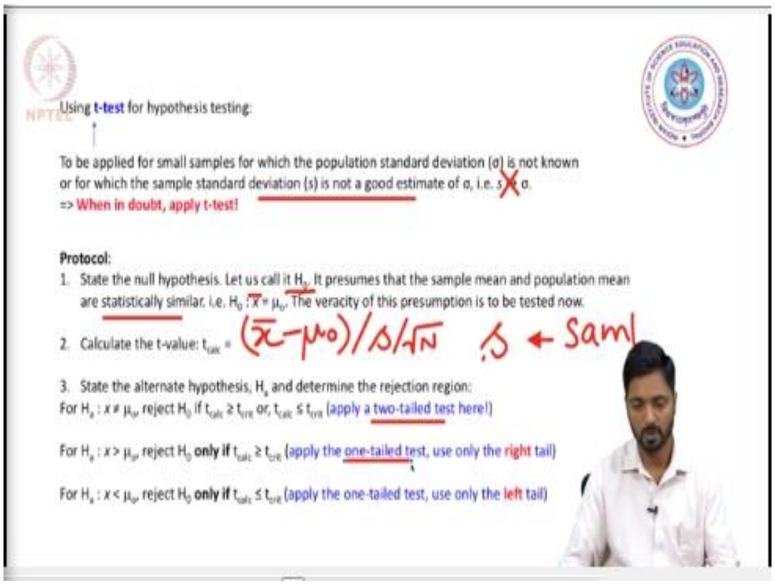
And this is mathematically written as what I have indicated in the box H 0 is x bar to indicate that it is a main is equal to mu 0, x bar is the sample mean and mu 0 is the population mean. Now we calculate z value based on the data with available to us. So, we do x bar - mu 0 and divide it by sigma by root N, where N is the number of ratings in a dataset. So, when we do our z calculation this way, we can compare this calculated z value with a critical z value which we have denoted as z critical.

And now comes the question that we need to address whether it is only about x bar being not equal mu 0, in those cases we will apply the two tail test. And we will compare whether the z calculated value is either greater than the z critical or less than the z critical. However, if our alternate hypothesis which is being denoted by H a is whether the sample mean value is greater than the population mean.

In those cases we can reject the null hypothesis only if the z calculated is greater than the z critical. And in that case we will be applying the one tail test, so the z value z critical value has to be chosen appropriately and we will be concerned only with the right tail of the Gaussian profile. If the alternate hypothesis is that the sample mean is less than the population mean, in those cases the null hypothesis is to be rejected.

If the z calculated is less than the z critical value at a certain confidence level. And again we will be using only one of the tails which is the left tail in this test ok.

(Refer Slide Time: 18:22)



Using **t-test** for hypothesis testing:

To be applied for small samples for which the population standard deviation (σ) is not known or for which the sample standard deviation (s) is not a good estimate of σ, i.e. $s \cancel{\approx} \sigma$.

=> **When in doubt, apply t-test!**

Protocol:

1. State the null hypothesis. Let us call it H_0. It presumes that the sample mean and population mean are statistically similar. i.e. $H_0 : \bar{x} = \mu_0$. The veracity of this presumption is to be tested now.
2. Calculate the t-value: $t_{calc} = (\bar{x}-\mu_0)/s/\sqrt{N}$ $s \leftarrow$ Saml
3. State the alternate hypothesis, H_a and determine the rejection region:

For $H_a : x \neq \mu_0$, reject H_0 if $t_{calc} \geq t_{cri}$ or, $t_{calc} \leq t_{cri}$ (apply a two-tailed test here!)

For $H_a : x > \mu_0$, reject H_0 **only if** $t_{calc} \geq t_{cri}$ (apply the one-tailed test, use only the right tail)

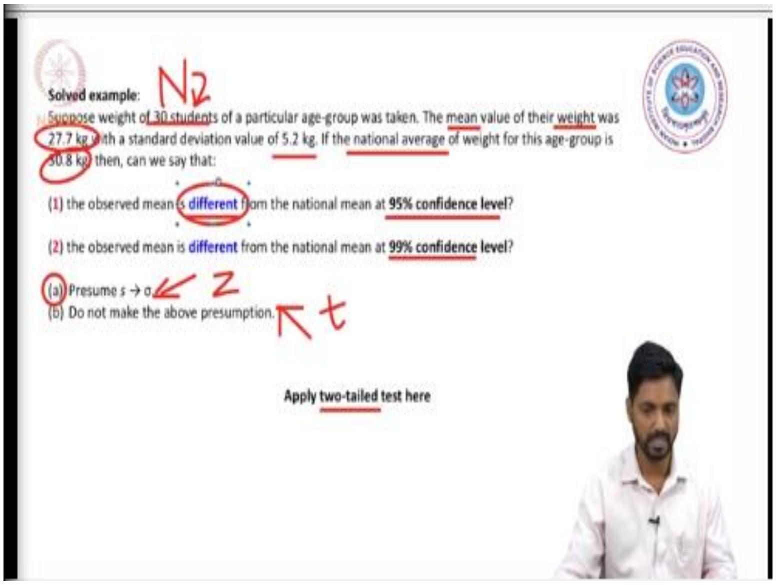
For $H_a : x < \mu_0$, reject H_0 **only if** $t_{calc} \leq t_{cri}$ (apply the one-tailed test, use only the left tail)

Now under which conditions do we apply the t test for hypothesis testing the t test is to be applied, when we do not have a good estimate of the population standard deviation or in other words, the presumption that the sample standard deviation is similar to the population standard deviation cannot be considered true. In those cases or whenever you are in doubt you should apply the t test, so the protocol is very similar for the t test as well.

We again come up with a null hypothesis, which presumes that the compared values are statistically similar. Now these compared values are the sample mean and the population mean for us. So we state the H 0 accordingly, and we test the veracity of this presumption by applying the t test, the formula also remains more or less similar. However, the only difference that comes now is that we are talking about the sample standard deviation s instead of the population standard deviation.

Now as previously we will apply either the two tail test or the one tail test depending upon the hypothesis that we want to test.

(Refer Slide Time: 20:26)



Solved example: N2

Suppose weight of 30 students of a particular age-group was taken. The mean value of their weight was 27.7 kg with a standard deviation value of 5.2 kg. If the national average of weight for this age-group is 30.8 kg then, can we say that:

(1) the observed mean is different from the national mean at **95% confidence level**?

(2) the observed mean is different from the national mean at **99% confidence level**?

(a) Presume $s \rightarrow \sigma$. z

(b) Do not make the above presumption. t

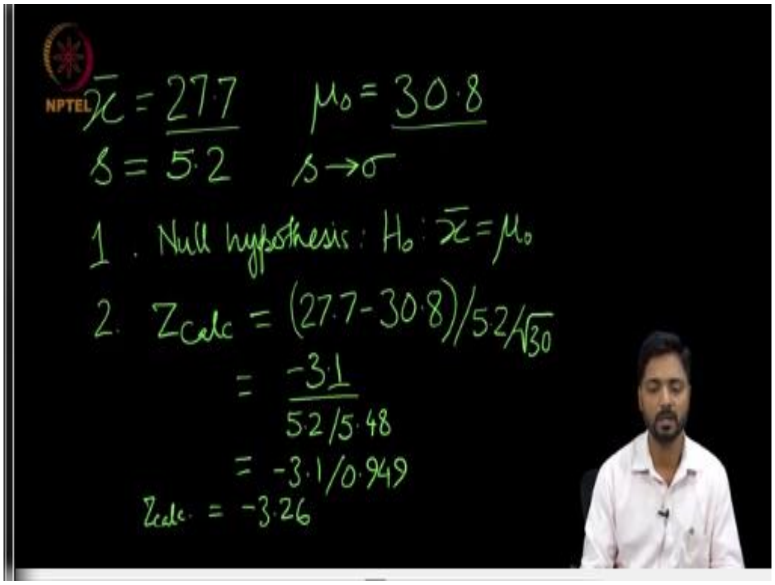
Apply two-tailed test here

Now let us take a real example to understand this concept in a greater detail. Suppose, we are talking about 30 students over N value is 30. And these students have mean weight of 27.7 kilograms, and the standard deviation for this sample is 5.2 kilograms. And there is an average weight which is a national average, which is available to us from the previous studies. And this tells us that for this age group the weight of the student should have been 13.8 kgs.

Now under these circumstances can we say that the observed mean is different from the national mean at 95% confidence level or in part 2 at 99% confidence level and we make a presumption in part A that the sample standard deviation s is same as the population standard deviation. In part B we will see if we do not presume this, then what happens. So in other words for a, we will be applying the z test for b we will be applying the t test.

And now let us apply that two tail test here, because we are only to address whether the readings are different or not, so let us do this exercise.

(Refer Slide Time: 22:44)



$\bar{x} = \underline{27.7}$ $\quad \mu_0 = \underline{30.8}$

$s = 5.2$ $\quad s \to \sigma$

1. Null hypothesis : $H_0 : \bar{x} = \mu_0$

2. $Z_{calc} = (27.7 - 30.8)/5.2/\sqrt{30}$

$= \dfrac{-3.1}{5.2/5.48}$

$= -3.1/0.949$

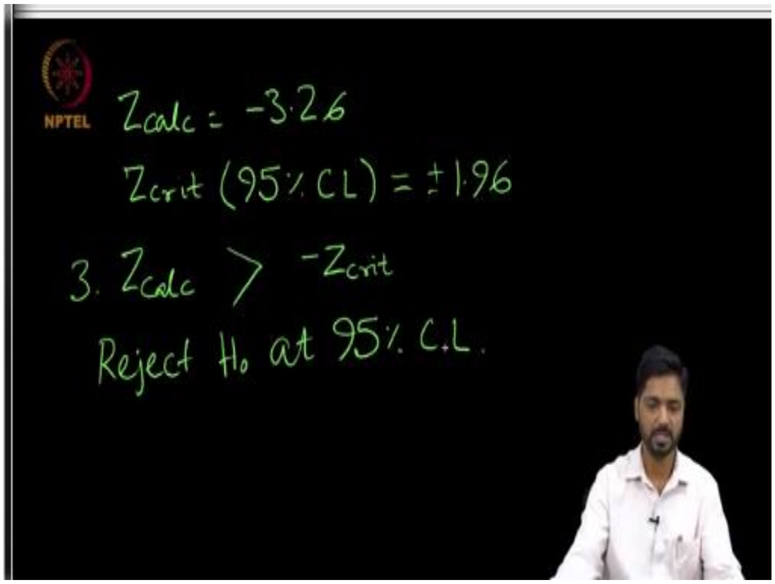
$Z_{calc.} = -3.26$

So, what we know from the obtain data is that the x bar value is 27.7 kilograms while the mu 0 value was supposed to be 30.8 kilograms, what we also know is that s = 5.2 kilograms. So, with this data available, and the presumption that s tends to sigma it would be easy to figure out whether the x bar value of 27.7 is statistically different from the national average of 30.8 kilograms.

So, let us do first is the null hypothesis and this states that or this presumes that the x bar value is same as the mu 0 value and the veracity of this hypothesis is to be tested now. So, we will do in the second step to the calculation of the z value for this dataset that will come out to be 27.7 - 30.8 divided by root 30. Now, this comes out to be - 3.1 divided by 5.2 by 5.48 or finally this turns out to be - 3.1 by 0.949.

In other words, the final z calculated value turns out to be - 3.26. Now, what we need to do is to calculate this z value. In the third step we will compare the z value with a z critical.

(Refer Slide Time: 25:32)



$Z_{calc} = -3.26$

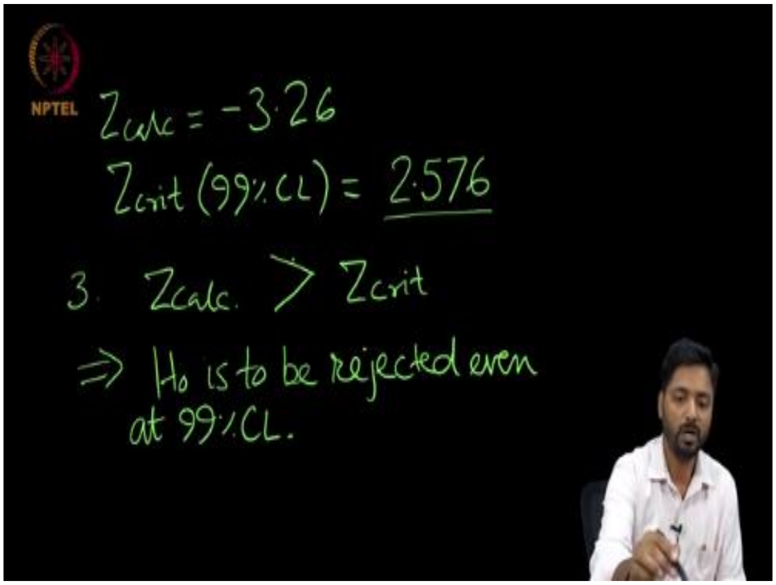
$Z_{crit} (95\% \text{ CL}) = \pm 1.96$

3. $Z_{calc} > -Z_{crit}$

Reject H_0 at 95% C.L.

So, z calculated was - 3.26 and z critical at 95% confidence level for a two tail test is to be checked and this should be equal to + - 1.96. So z calculated is greater than the negative value of the z critical. Then we say that we need to reject H 0 the null hypothesis at 95% confidence level ok. Now this was the part one of the question in the second part we wanted to ask if at 99% confidence level also we can say that.

(Refer Slide Time: 26:46)



$Z_{calc} = -3.26$

$Z_{crit} (99\% \text{ CL}) = \underline{2.576}$

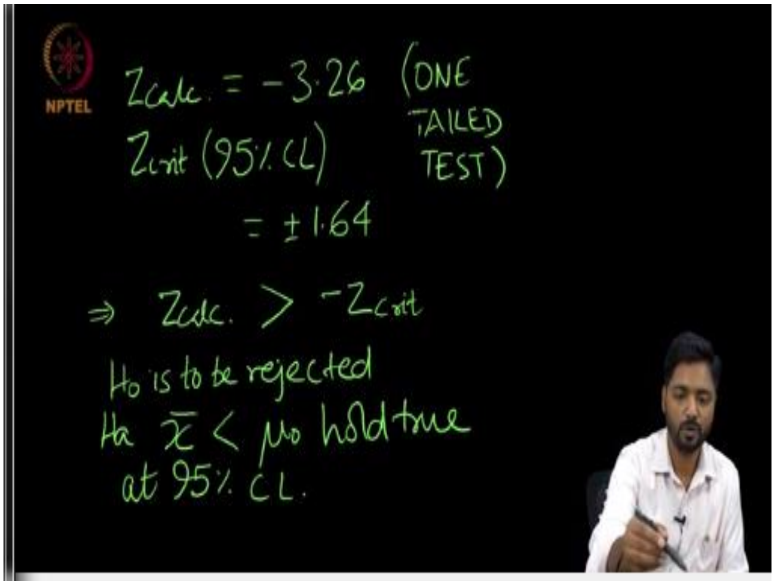
3. $Z_{calc} > Z_{crit}$

$\Rightarrow$ H_0 is to be rejected even at 99% CL.

So, we just need to compare the calculated z value which is not going to change with a z critical value 99% confidence level which now will be different and z critical value now is 2.576. You can look at the t table to check that this number being reported for the z critical value is correct or not. Because you need to report this number and then make a comparison and when we compare, we still find the third step still is that the z calculated value is greater than the z critical value.

And which implies that the null hypothesis is to be rejected even at 99% confidence level. So we can report our conclusions at 99% confidence level. That the mean value of the weight of these 30 students is indeed different from that of the national average, because that was the question that we were dealing with. However if the mean weight of the students 30 students that we are dealing with is different not just different but is lower than the population mean which is 30.8 kilograms.

(Refer Slide Time: 28:57)



$Z_{calc.} = -3.26$ (ONE TAILED TEST)

Z_{crit} (95% CL) $= \pm 1.64$

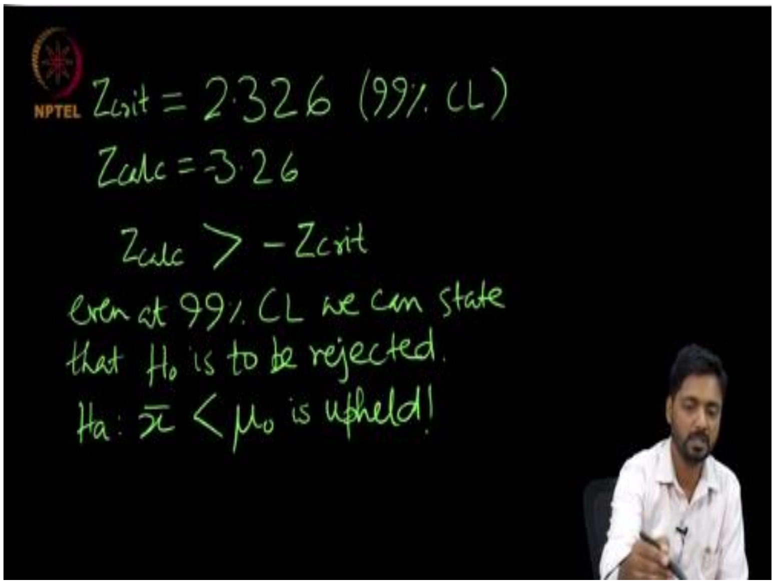
$\Rightarrow Z_{calc.} > -Z_{crit}$

H_0 is to be rejected

H_a $\bar{x} < \mu_0$ hold true at 95% CL.

In this case, we will be applying the one tail test and the z calculated value however will not change the z critical values nonetheless will be changed here. So if we are talking about the 95% confidence level. Now we need to apply the one tailed test and hence our z critical value will be + - 1.64. This implies the z calculated value is greater than the negative z critical value of **1.** -1.64 in other words even here the null hypothesis is to be rejected. And the alternate hypothesis which is x bar is less than mu 0 holds true at 95% confidence level.

(Refer Slide Time: 30:48)



$Z_{crit} = 2.326$ (99% CL)

$Z_{calc} = -3.26$

$Z_{calc} > -Z_{crit}$

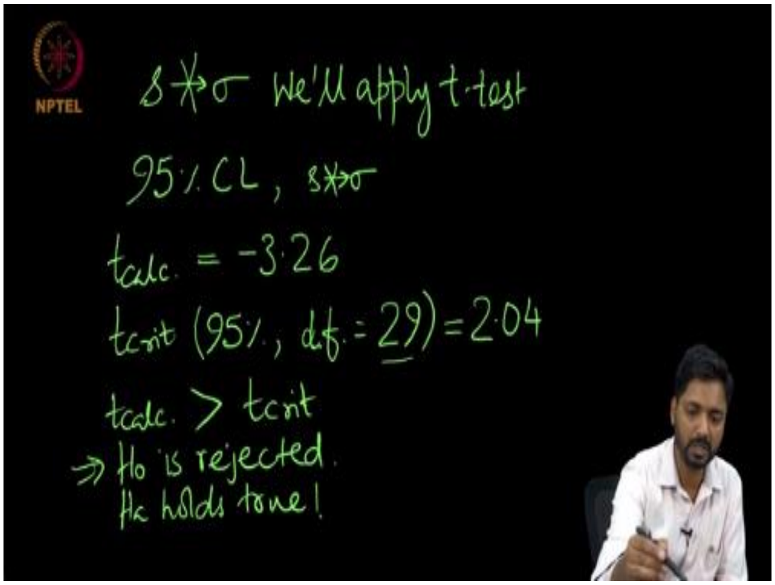
even at 99% CL we can state that H_0 is to be rejected.

$H_a: \bar{x} < \mu_0$ is upheld!

Now, repeating the same thing for 99% confidence level, we need to let us quickly go to the t table and see what the z value should be used here. We are using the one tail test and we want to see the 99% significance level or 1% significance level 99% confidence level, so our z value is 2.326. When we use this critical value also, we see that the z calculated value of 3.26 is still greater than the negative z critical value.

So even at 99% confidence level we can state that the null hypothesis is to be rejected or the alternate hypothesis that x bar value is less than mu 0 is upheld ok. Now the situation will change if our presumption that s is not.

(Refer Slide Time: 33:10)



$s \not\to \sigma$ We'll apply t-test

95% CL, $s \not\to \sigma$

$t_{calc.} = -3.26$

$t_{crit}(95\%, d.f. = \underline{29}) = 2.04$

$t_{calc.} > t_{crit}$

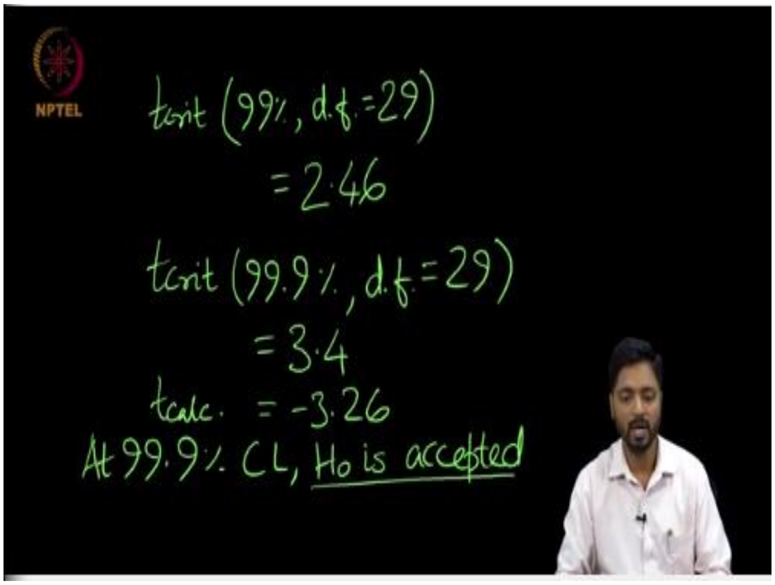
$\Rightarrow H_0$ is rejected.

H_a holds true!

Our presumption that s is tending to sigma does not hold true. In that case we will be applying the t test and use the appropriate t values. Let us consider the first case which was at 95% confidence level s not tending to sigma. In that case, our t calculated would be still – 3.26, but the t critical value would take into account the confidence level as well as the number of readings in terms of the degrees of freedom, and the degrees of freedom now is 29.

So, let us look at the t critical value from the table at 29 degrees of freedom, while we do not have it, what we definitely have it the t value at 95% confidence level and at 28 degrees of freedom. So, we can extrapolate this number to 2.04 and what we observe is that the t calculated is again greater than the t critical. So at 95% confidence level our H 0 is rejected and the alternate hypothesis holds true ok. Even if we do the same exercise at 99% confidence level, I will just give you the t critical value in that case.

(Refer Slide Time: 35:56)



t_{crit} (99%, d.f. = 29)

= 2.46

t_{crit} (99.9%, d.f = 29)

= 3.4

$t_{calc.}$ = −3.26

At 99.9% CL, <u>H_0 is accepted</u>

So, how you write t critical for 99% confidence level and degrees of freedom is equal to 29. We will go back to the table and we are looking at the 99% confidence level and around this will be the reading. Let us presume that this reading is 2.46 ok, even if we extend it to 99.9% confidence level and degrees of freedom is equal to 29 the t critical value will be 3.4 now. While our t calculated to recollect was - 3.26.

So, at 99.9% confidence level we cannot reject the null hypothesis or in other words the null hypothesis is to be accepted. So I hope with this example you understood how t critical and z critical values are to be found and how you can utilize the t test or the z test to tell whether a reading is different from a population mean. And not just different whether it is greater than or lower than the population mean at a certain confidence level.