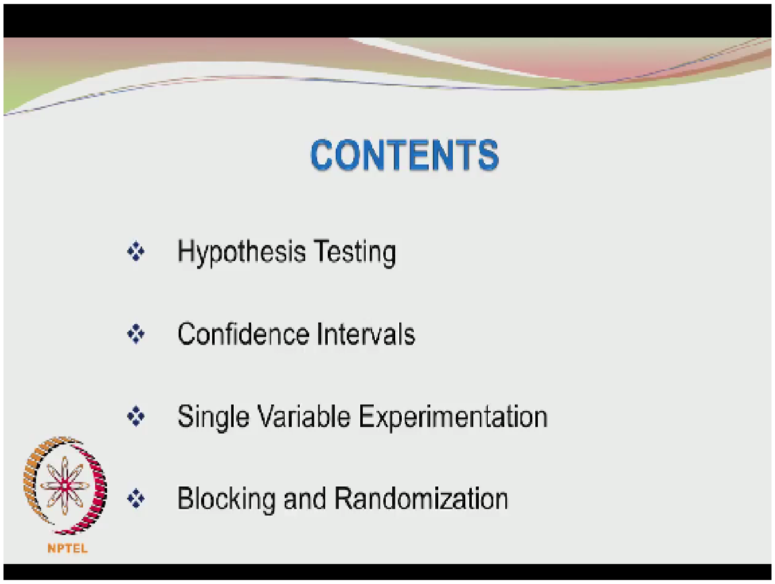
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture – 28
Example Set -6 - Part A

Hello, good to have you back. In today's classes, we will be solving a few problems, typical problems illustrating some of the concepts we have learnt, especially pertaining to confidence intervals on the standard deviation, blocking and randomisation in single variable experimentation. As I have told you previously, I request you to solve the problems independently before looking at the solutions.

By this way, not only you can catch your mistakes, you may possibly catch some of mine. Hopefully, I have not made any mistake. I verified the solutions independently but who knows. There may be possibility of error and if you pick up any error, please let me know, okay. Let us get on.

(Refer Slide Time: 01:27)

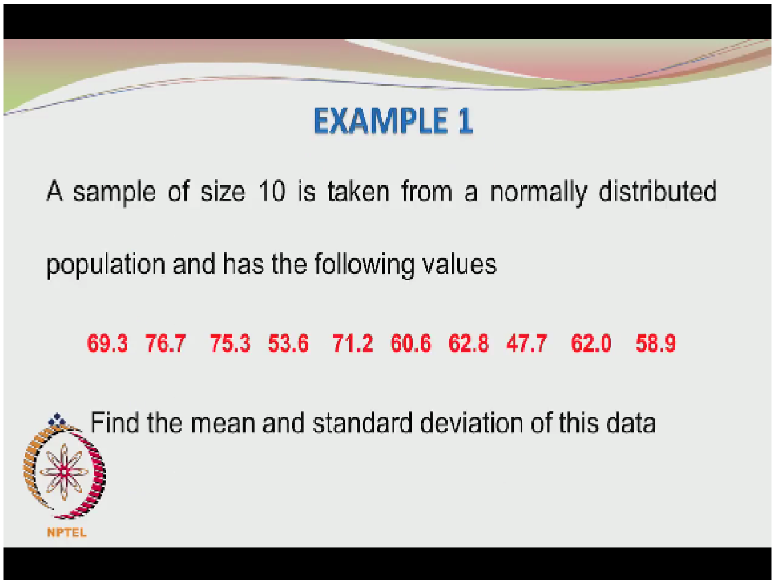


CONTENTS

- Hypothesis Testing
- Confidence Intervals
- Single Variable Experimentation
- Blocking and Randomization

The contents of this examples set are problems solved in hypothesis testing, confidence intervals, single variable experimentation, blocking and randomisation.

(Refer Slide Time: 01:38)



EXAMPLE 1

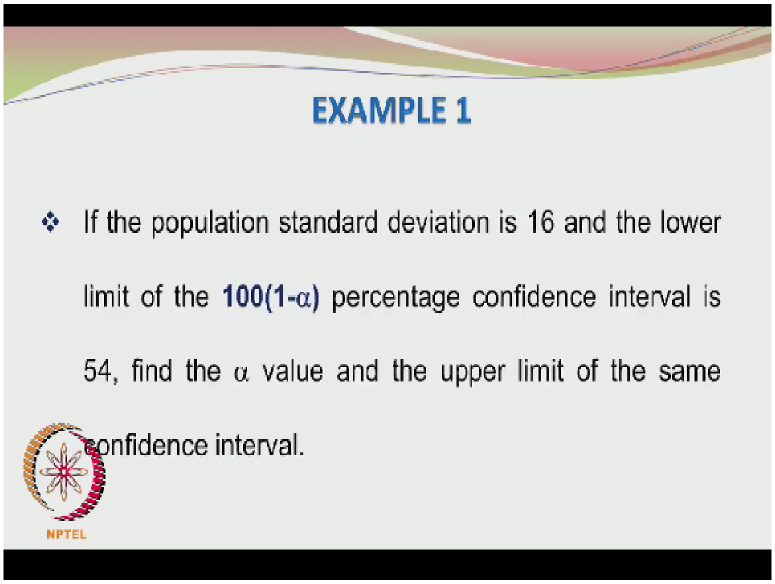
A sample of size 10 is taken from a normally distributed population and has the following values

69.3 76.7 75.3 53.6 71.2 60.6 62.8 47.7 62.0 58.9

- Find the mean and standard deviation of this data

Without any further ado, let us get on to the first example. We have a sample of size 10 taken from a normally distributed population with the following values. Well I do not know what these numbers represent. They do not look like marks. They look more like weights of people; nobody is grossly overweight. Anyway, so the numbers are 69.3 76.7 75.3 53.6 71.2 60.6 62.8 47.7 62 58.9. So you are asked to first find the mean and the standard deviation of this data. This should be pretty simple for all of you.

(Refer Slide Time: 02:48)



EXAMPLE 1

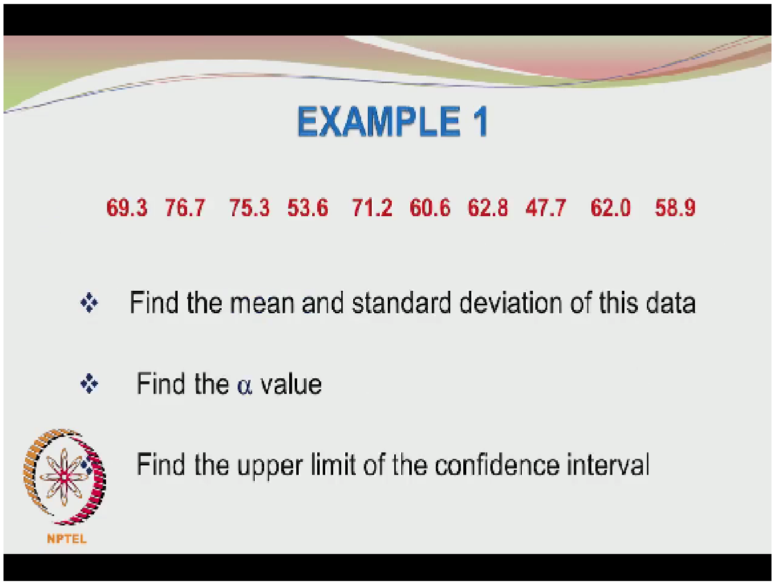
- If the population standard deviation is 16 and the lower limit of the 100(1-α) percentage confidence interval is 54, find the α value and the upper limit of the same confidence interval.

And the next part of the example is to find the level of significance. You have to find the alpha value given the lower limit of the 100*1-alpha percentage confidence interval is 54 and the population standard deviation is 16. So the population is said to be normal and the standard deviation is provided to you. You are also given the lower limit to be 54. Well you can argue what is the lower limit for.

Well that is not given. It is assumed to be the mean because you cannot have this data a standard deviation of 54. So you will not have a lower limit as 54 and hence it has to be the mean or the average mu of the parameter, mu, the parameter of the population. So you are given the confidence interval lower limit. You are also given sigma.

So you have to find out what is the alpha value using this information and after you have found the alpha value, it should be a piece of cake to find the upper limit of the same confidence interval. Well instead of asking the question directly, you are given one part of the problem and you are asked to find the parameter and then hence find the remaining missing part.

(Refer Slide Time: 04:49)



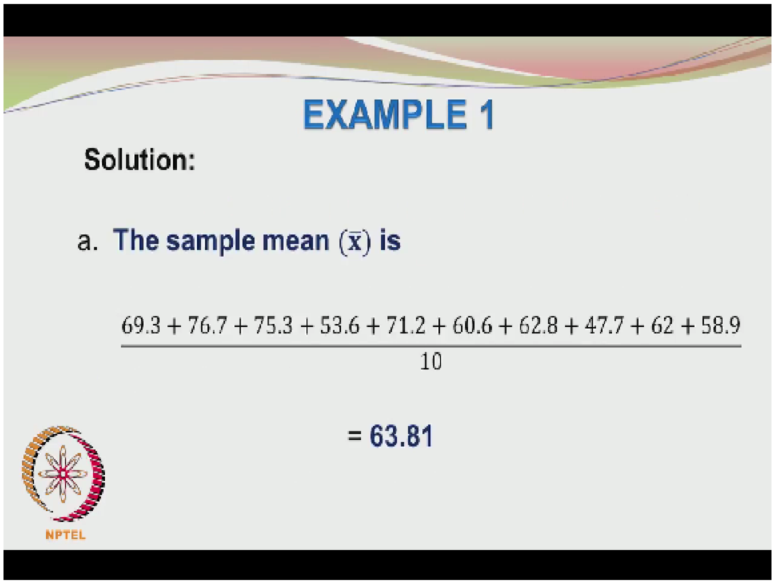
EXAMPLE 1

69.3 76.7 75.3 53.6 71.2 60.6 62.8 47.7 62.0 58.9

- Find the mean and standard deviation of this data
- Find the α value
- Find the upper limit of the confidence interval

So to summarize, you have to find the mean and standard deviation of the data given above, the alpha value and the upper limit of the confidence interval. Wish I had given the data in an ascending order in which case we could have found the median and seen how close the mean was to the median that I think you can do. Let us see how many data points are there 1 2 3 4 5 6 7 8 9 10, okay. So you are having 9+1 10 data points.

(Refer Slide Time: 05:31)



EXAMPLE 1

Solution:

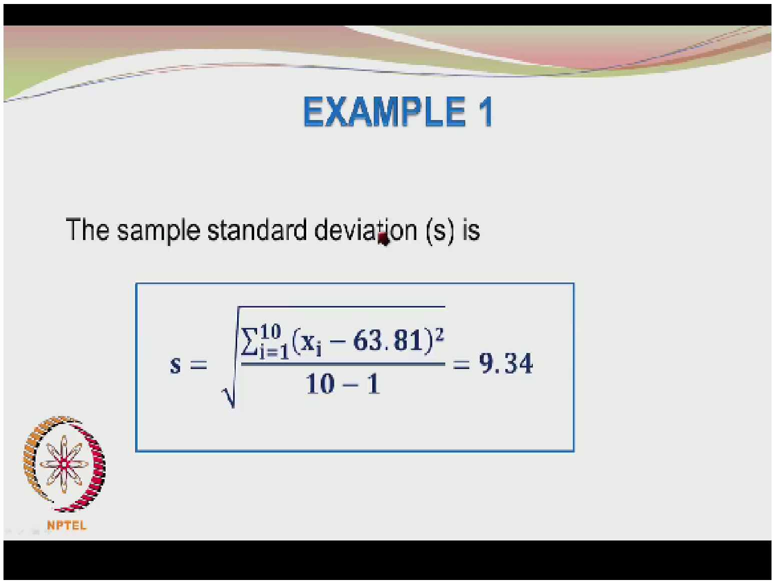
a. The sample mean ($\bar{x}$) is

$$\frac{69.3 + 76.7 + 75.3 + 53.6 + 71.2 + 60.6 + 62.8 + 47.7 + 62 + 58.9}{10}$$

$$= 63.81$$

And the sample mean is 63.81. Well you could have reported the answer as 63.8 but no harm in adding the second decimal. So 63.81 is the sample mean.

(Refer Slide Time: 05:50)



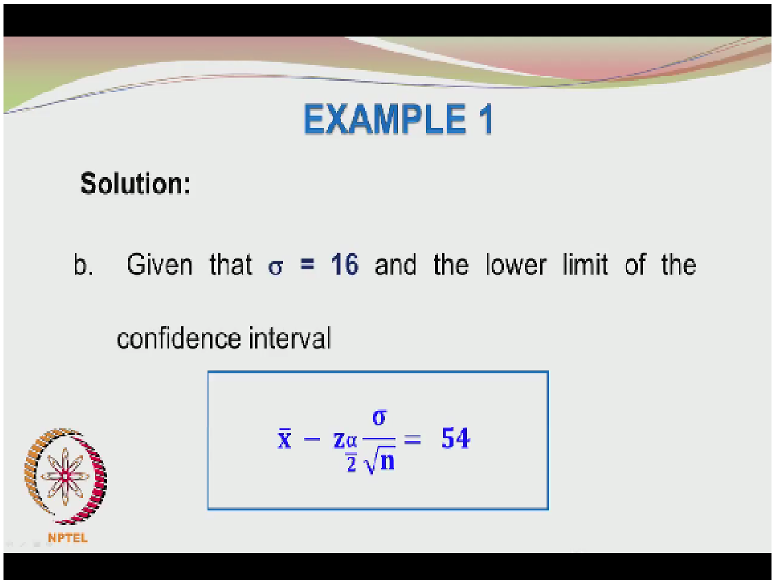
EXAMPLE 1

The sample standard deviation (s) is

$$\boxed{s = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 63.81)^2}{10 - 1}} = 9.34}$$

The sample standard deviation is obtained by taking the individual xi values, subtracting the sample mean from it, squaring it, so we are getting a deviation from the sample mean. Then we square it and add up all the sum of the square of the deviations, divide it by number of sample elements, which is 10 in this case, -1 and so we get the one under the square root. Since you are being told to find the standard deviation, yes we had to take the square root of the sum of squares of deviations/n-1 and that answer comes to 9.34, okay.

(Refer Slide Time: 06:39)



EXAMPLE 1

Solution:

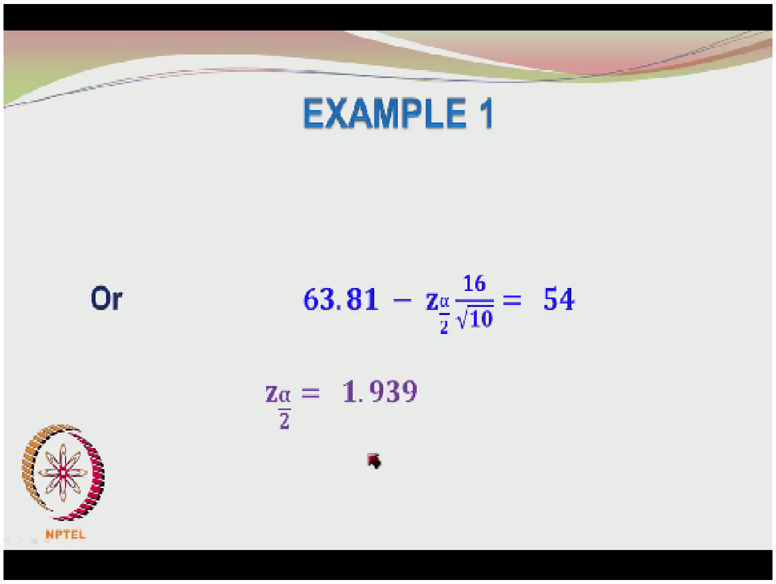
b. Given that σ = 16 and the lower limit of the confidence interval

$$\boxed{\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 54}$$

So you are given the population standard deviation as 16, it is quite different from the sample standard deviation of 9.34 but that is expected. It just shows that the sample value in this particular case is quite far from the actual population standard deviation and the lower limit of the confidence interval is given as 54. So since you are given normal distribution and you are given the value of sigma and then you are also told it is the lower limit which means that it is a 2-sided confidence interval.

So using this information, we know that it should be alpha/2 which we should use. So we get x bar-z alpha/2 sigma/root n. We use alpha/2 because it is a 2-sided confidence interval. X bar is a sample mean, z is the value of the standard normal variable corresponding to the upper tail alpha/2 percentage points. Sigma/root n, sigma is the given standard deviation and n is the number of data points in the sample or simply the sample size. So you are given x bar, you are not actually given, you found out x bar. You are given sigma. You know n and you have to find z alpha/2, right.

(Refer Slide Time: 08:27)



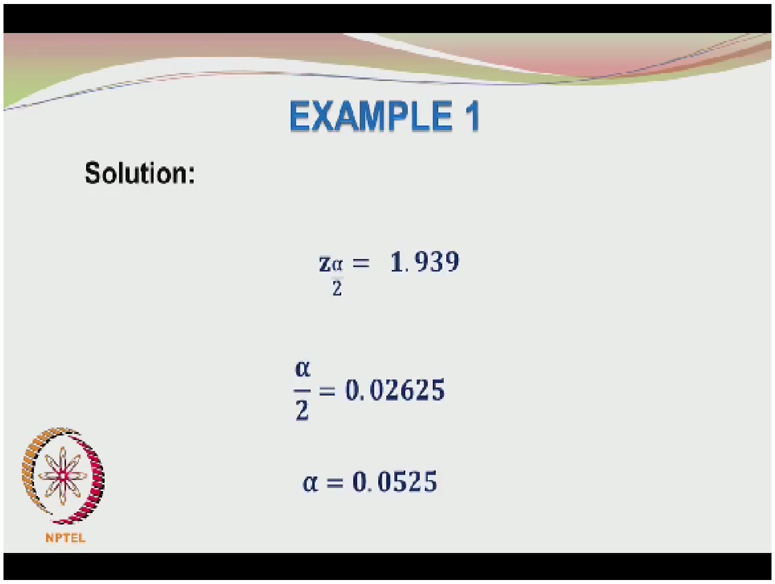
EXAMPLE 1

Or $$63.81 - z_{\frac{\alpha}{2}} \frac{16}{\sqrt{10}} = 54$$

$$z_{\frac{\alpha}{2}} = 1.939$$

So 63.81-z alpha/2*16/ root 10=5.4, sorry =54 and so 63.81-54, 9.81, 9.81, let us say it is 10, 10*3, 30, 30/16 is approximately 2 and so z alpha/2 is coming to about 1.939 exactly.

(Refer Slide Time: 09:04)



EXAMPLE 1

Solution:

$$z_{\frac{\alpha}{2}} = 1.939$$

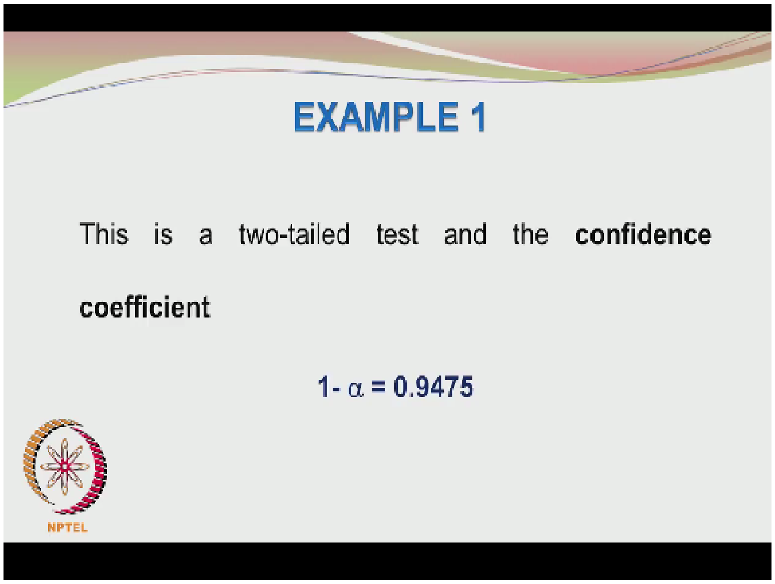
$$\frac{\alpha}{2} = 0.02625$$

$$\alpha = 0.0525$$

So z alpha/2=1.939. So we have to look at the standard normal distribution curve and we have to locate the 1.939 on the x-axis and see the probability beyond the value. What I am trying to say here is we have to find the area under the normal probability curve, the standard normal probability curve beyond the value of 1.939. So alpha/2 comes to around 0.02625 which means alpha=0.0525, slightly different from the usual 0.05 we use.

So alpha is coming to 0.0525. It is important that you use the normal probability curve properly, locate 1.939 and then find the cumulative probability up to 1.939 and hence the probability or the area under the curve above 1.939 would be 0.02625. So alpha then comes to twice this 0.02625 which is 0.0525.

(Refer Slide Time: 10:34)



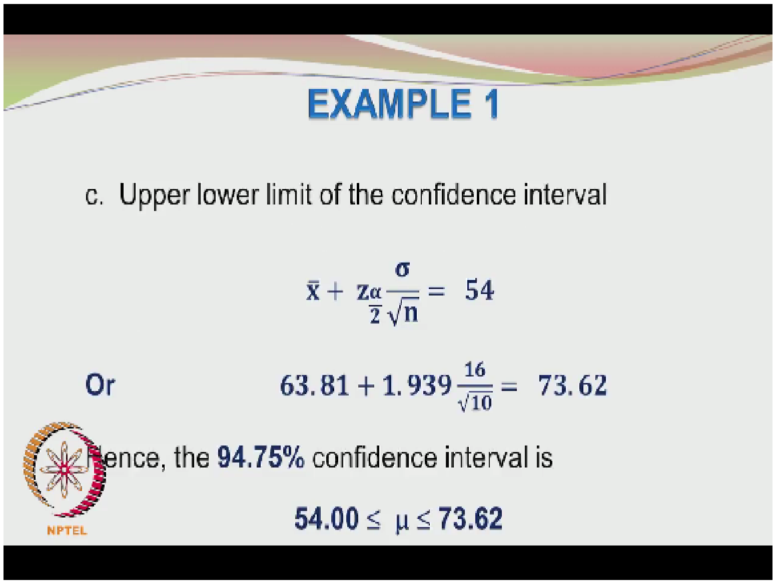
EXAMPLE 1

This is a two-tailed test and the **confidence coefficient**

$$1 - \alpha = 0.9475$$

So 1-alpha would be = 0.9475. Here you are not using alpha/2, we are using alpha. It is a 2-tailed test. We have to look at the total area including the area on the rights side of the tail and the left side of the tail. So it will be 2*alpha/2 and so the alpha is 0.0525. So 1-0.0525 is 0.9475.

(Refer Slide Time: 11:06)



EXAMPLE 1

c. Upper lower limit of the confidence interval

$$\bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 54$$

Or $$63.81 + 1.939 \frac{16}{\sqrt{10}} = 73.62$$

Hence, the **94.75%** confidence interval is

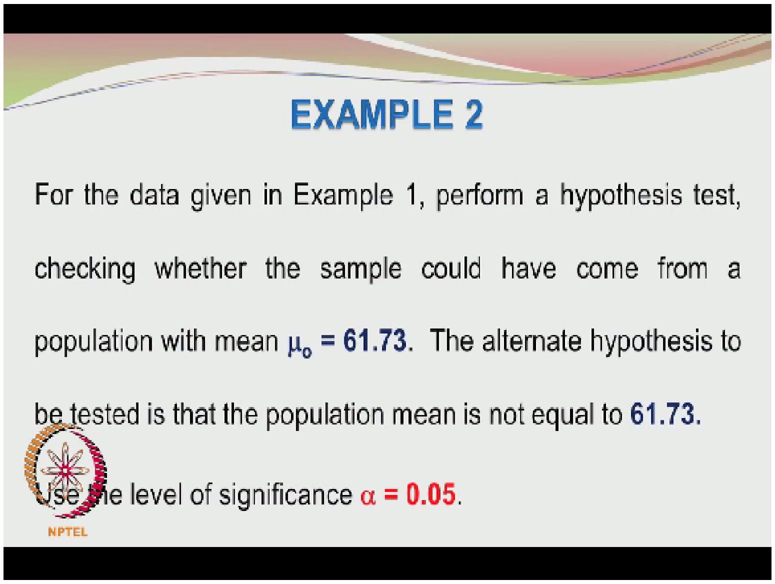
$$54.00 \leq \mu \leq 73.62$$

That is the confidence coefficient. Now all the information is known to us. We know z alpha/2. We know all other information. So 63.81+1.939*16/root 10 that comes to 73.62. So we report the

94.75% confidence interval as 54 <= mu <= 73.62. So if you generate a large number of confidence intervals based on a large number of random samples drawn from the normal population.

It is expected that 94.75 percentage of the constructed confidence intervals would encompass or surround or have the population parameter mu within their bounds. So out of let us say 100 confidence interval is constructed based on 100 random samples, about the 95% of them or 95 of them, would surround the population parameter mu. So it is important that we understand the meaning of the confidence interval.

(Refer Slide Time: 12:44)



EXAMPLE 2

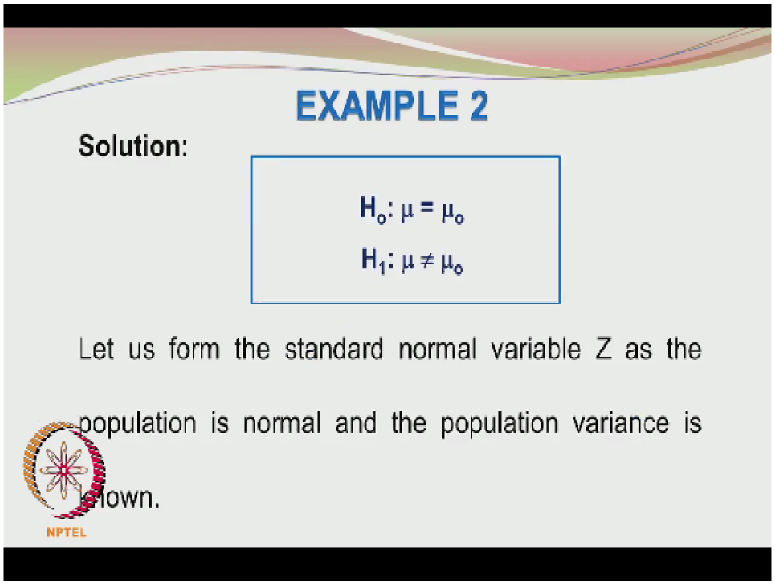
For the data given in Example 1, perform a hypothesis test, checking whether the sample could have come from a population with mean $\mu_o = 61.73$. The alternate hypothesis to be tested is that the population mean is not equal to 61.73.

Use the level of significance $\alpha = 0.05$.

Quickly we move on to the second problem which is based on the first problem. For the data given in example 1, perform a hypothesis test checking whether the sample could have come from a population with mean mu0=61.73. The alternative hypothesis to be tested is the population mean != 61.73. Sigma is known, we are speculating on the value of mu of the population.

According to null hypothesis, it is 61.73. We have drawn a sample which is having a slightly higher mean. So we have to see whether it has indeed come from this population of mean, 61.73, whether the elements of the sample were actually taken from a population of mean 61.73. The level of significance to be used is 0.05, okay.

(Refer Slide Time: 13:53)



EXAMPLE 2

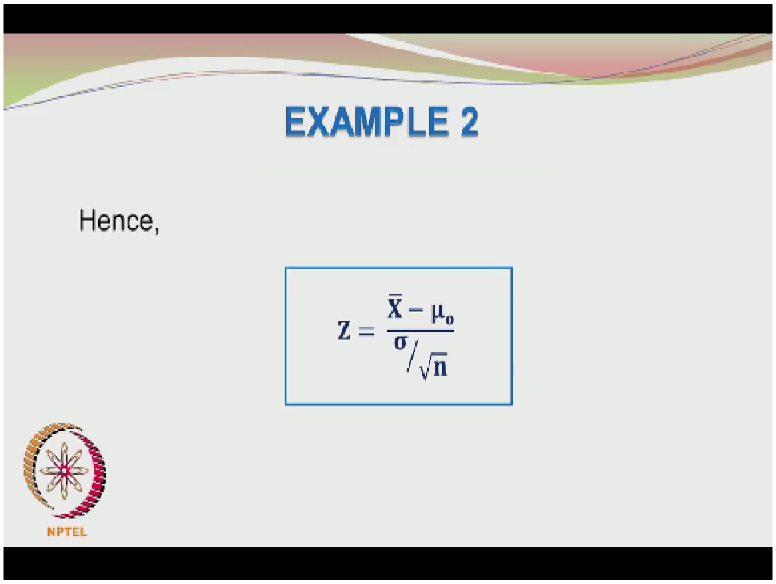
Solution:

$$\boxed{\begin{array}{c} H_o: \mu = \mu_o \\ H_1: \mu \neq \mu_o \end{array}}$$

Let us form the standard normal variable Z as the population is normal and the population variance is known.

H0 mu=mu0 and H1 mu != mu0. You please enter the statement; otherwise, your class teacher may cut half a mark or 1 mark for not explicitly stating this upfront. Let us form the standard normal variable as the population is normal and the population variance is known.

(Refer Slide Time: 14:23)



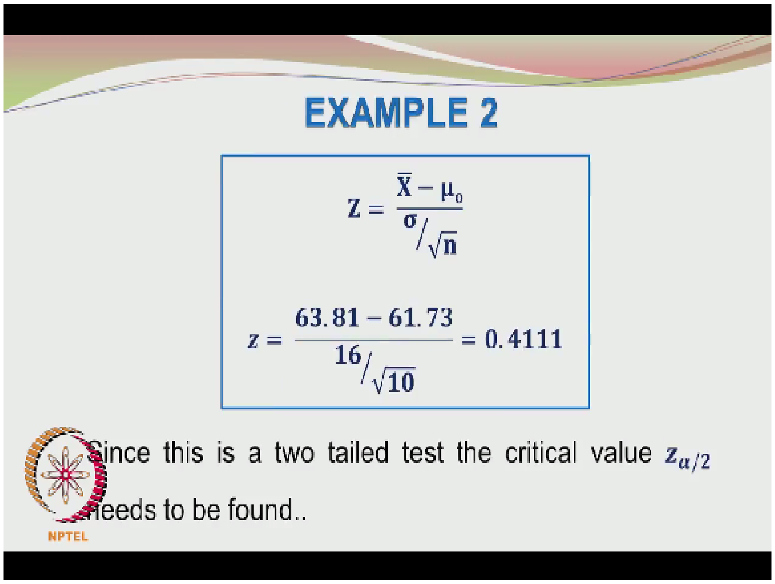
EXAMPLE 2

Hence,

$$\boxed{Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}}}$$

So z=the standard normal variable, standard normal has a mean of 0 and standard deviation of 1. So we get z=X bar-mu0/sigma/root n.

(Refer Slide Time: 14:43)



EXAMPLE 2

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

$$z = \frac{63.81 - 61.73}{16/\sqrt{10}} = 0.4111$$

Since this is a two tailed test the critical value $z_{\alpha/2}$ needs to be found..

So we are having X bar, the sample has been taken and we know the mean to be 63.81 and mu0 is 61.73, sigma is 16 and n is 10. I think 0.27, 1.81, so 2.08, 6.34, okay, about 6.4/16 would be about 0.4, right. So we are having this z value as 0.411. The 2-tailed test has to be performed because the alternative hypothesis is mu != mu0. So we do a 2-tailed test and also try to find out what is the critical value z alpha/2.

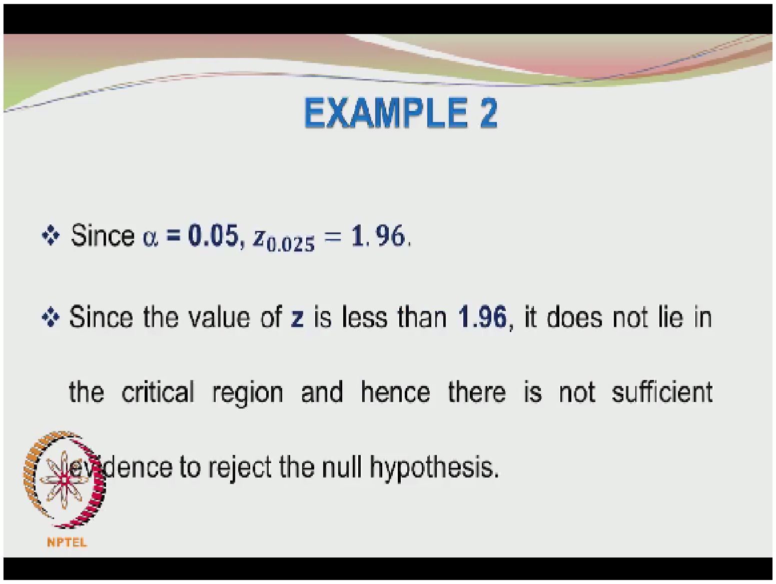
Please remember that we are having the standard normal and the standard normal is created by normalizing the sample mean by first subtracting it with mu0, the speculated population mean and most importantly, we are dividing it by sigma by root n. We are not dividing it by sigma. In both the cases, whether you divide by sigma by root n or sigma, it would not have really mattered because it would have been a standard normal variable but we have to see what is the probability distribution we are looking at now.

We are having a normally distributed sampling distributions of the mean and the sampling distribution of the mean would be centered at mu0 and have a standard deviation of sigma/root n. So this sampling distribution of the mean which is the normal distribution is being converted into a standard normal distribution and hence we have to do the transformation z=X bar-mu0/sigma/root n to show that we are normalizing the sampling distributions of the mean and once we have done that, we get the value of z.

Now in this sampling distribution of the mean, we have 2 regions, one is the acceptance region and another is the rejection region. If our statistic z value is such that the value here exceeds the critical value, then the statistic is lying in the rejection region and you have to reject the null hypothesis. If the value of z is lying in the acceptance region, then now you have to except the null hypothesis.

So how do you know whether this z value of 0.411 is lying in the rejection region or in the acceptance region. For that, we need to find the value of z alpha/2. We have to find the critical value which divides or serves as the boundary between the acceptance region and the rejection region. Since it is 2-tailed test, we have to find what is the value of Z alpha/2.

(Refer Slide Time: 18:37)



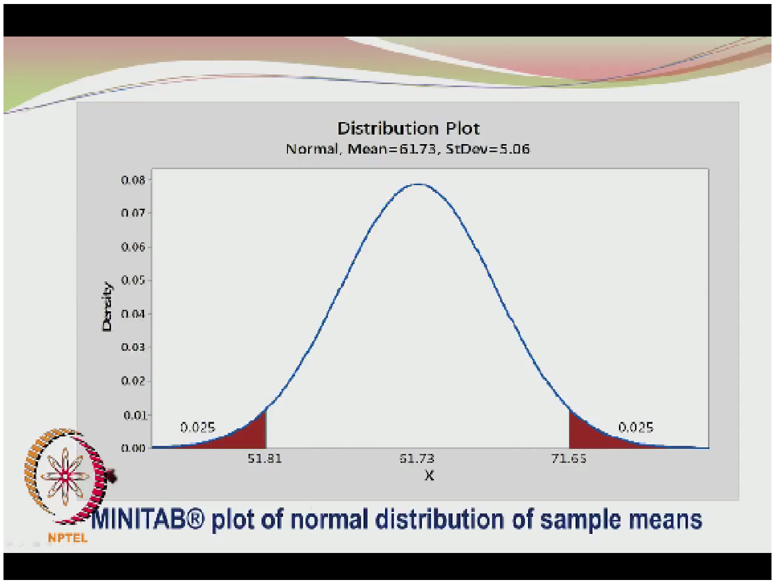
EXAMPLE 2

- Since α = **0.05**, $z_{0.025} = 1.96$.
- Since the value of **z** is less than **1.96**, it does not lie in the critical region and hence there is not sufficient evidence to reject the null hypothesis.

So z 0.05/2 which is z0.025 is 1.96 and this is a very famous and popular number. It crops up very frequently and so I think by the end of the course if not by now, you will know this number by heart. Similarly, you should also be knowing what is the value of z 0.05. So anyway, no need to memorize, I am just telling that since a number becomes very familiar, we tend to remember it like mobile numbers, our ID's and so on, okay.

So the value of 0.4111 is lesser than 1.96. It does not lie in the critical region and hence we can claim that there is not sufficient evidence to reject the null hypothesis.

(Refer Slide Time: 19:39)

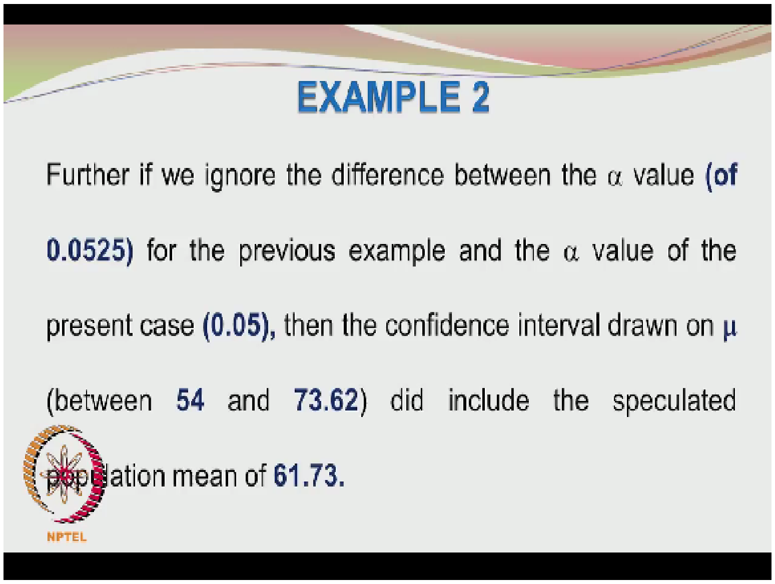


MINITAB® plot of normal distribution of sample means

So I have plotted this in the Minitab. This is the sampling distribution of the mean and normal distribution centered around the speculated value of 61.73 and so the critical value corresponds to a mean of 71.65 and the mean of 51.81. In our present case, the sample mean is higher than the speculated mean. We have to see whether that sample mean is lying between 61.73 and 71.65 or it is lying between, sorry, it is lying beyond 71.65.

In our present case, the sample mean came to 63.81. So 63.81 is going to be present somewhere here and that is lying well before the rejection region. This is the acceptance region. So we accept the null hypothesis. So only if the sample mean had taken a value of 71.65 or 51.81 or lower, what I am trying to say here is if the sample mean had a value beyond 71.65 or a value below 51.81, we would have rejected the null hypothesis; otherwise, we are safe in accepting it. All right.

(Refer Slide Time: 21:19)



EXAMPLE 2

Further if we ignore the difference between the α value **(of 0.0525)** for the previous example and the α value of the present case **(0.05),** then the confidence interval drawn on μ (between **54** and **73.62**) did include the speculated population mean of **61.73.**

We could have done this problem even without doing the hypothesis test in an approximate manner. These approximate solutions are pretty useful. It gives in the order of magnitude estimate of the correct answer and it also tells us okay you are on the right track. So if you look at the previous example, we were our finding the alpha value as 0.0525, whereas the present value of alpha is 0.05.

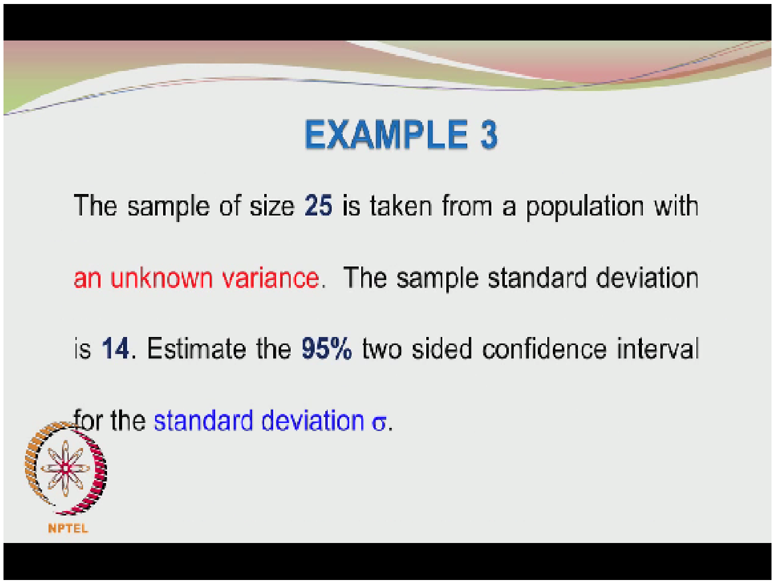
So we could have used information given in the first example in an approximate manner to check whether our null hypothesis is correct even without doing any further calculations. So the information in the first example could have been used to at least approximately solve the second example without going through the procedure of standard normalization and the finding the critical value and all that.

The important thing to notice, we are having only 1 random sample. From the random sample, we can do hypothesis testing or we can construct the confidence interval and even though both of them have slightly different uses, both of them pretty much give the same conclusion regarding the sample, regarding, I made a mistake here, regarding the population parameter mu. So when we looked at the confidence interval, it bounded the values 54 and 73.62.

So this particular 94.75% confidence interval said mu is located between 54 and 73.62. Now the speculated population mean is 61.73. So since 61.73 is lying between these 2 numbers, then we can say that yes the parameter mu is falling between the upper and lower limits of the confidence interval. So we can accept the null hypothesis. For this claim, we are neglecting the difference between 94.75% confidence interval and 95% confidence interval.

Well the approximation is not a bad one because our mean value of 63., sorry 61.73 is well within these 2 bounds. If it had been something like 55 or 70, then we would have not been sure whether the 94.75% bounds are really including the population parameter but it is well within the bounds, so we can reasonably assume even without doing any further calculations that S the speculated mean value of 61.73 is indeed lying between the lower and upper bounds of the 95% confidence interval and hence the null hypothesis may be accepted, okay.

(Refer Slide Time: 24:38)



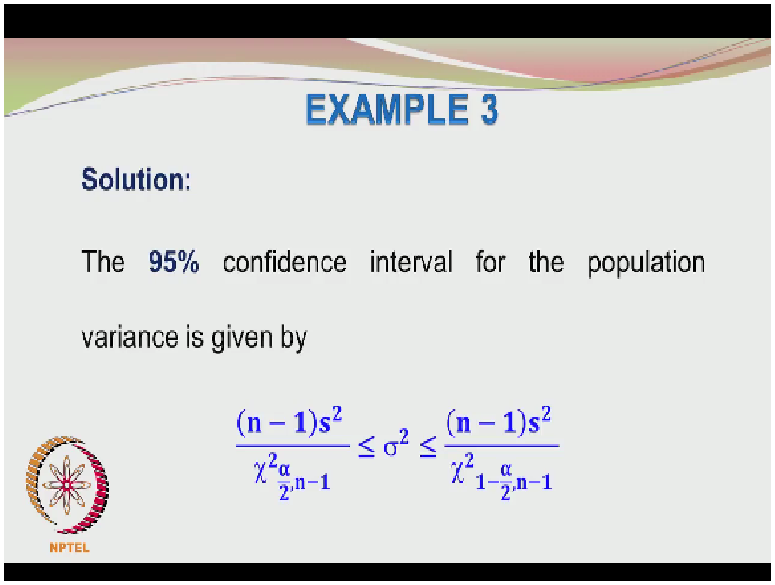
EXAMPLE 3

The sample of size **25** is taken from a population with an unknown variance. The sample standard deviation is **14**. Estimate the **95%** two sided confidence interval for the standard deviation σ.

Let us move on to the third example. Here we are having a sample of size 25 which is taken from a population with an unknown variance, right. The sample standard deviation is 14 and estimate a 95% 2-sided confidence interval for the standard deviation sigma. So please read the problem statement, write down the information given and then think of how to solve it. Here we are noticing that X bar is not given.

Only the sample standard deviation S is given to us and then you may think that enough information is not provided if we had read the question hastily. We are not asked to find the confidence interval of the population mean mu. We are asked to find the confidence interval for the standard deviation sigma. We want to find the lower bound and upper bound for a Sigma. So just think about how you would go about solving this problem, right.

(Refer Slide Time: 25:59)



EXAMPLE 3

Solution:

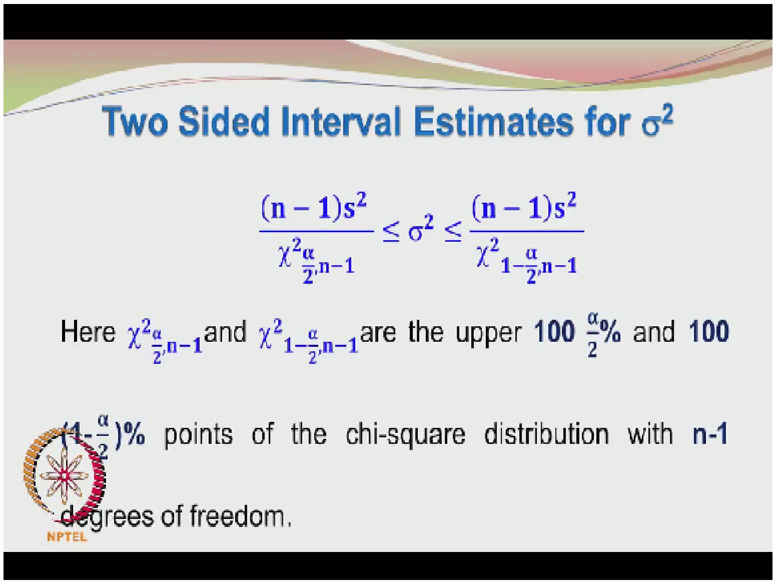
The **95%** confidence interval for the population variance is given by

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

I hope you solve the problem by now. If not, let us do it together, 95% confidence interval for a population standard deviation is given by this. Actually there is a typo, let me correct it immediately. It is for the population variance, okay. Here we go. So the 95% confidence interval for the population variance is given by n-1s squared/chi-squared alpha/2n-1 <= sigma squared <= n-1 s squared/chi-squared 1-alpha/2 n-1.

We are using the chi-square distribution because it represents the distributions of the variances. So we have to find the chi-squared percentage points corresponding to alpha value of 0.05. So alpha value would be 0.05, alpha/2 will be 0.025, n-1 would be the degrees of freedom associated with the sample standard deviation, the sample size is 25, the degrees of freedom would be then 24. So we have to use 25-1 and the alpha/2 would be 0.05/2, n-1 would be 25-1 which is 24.

(Refer Slide Time: 27:49)



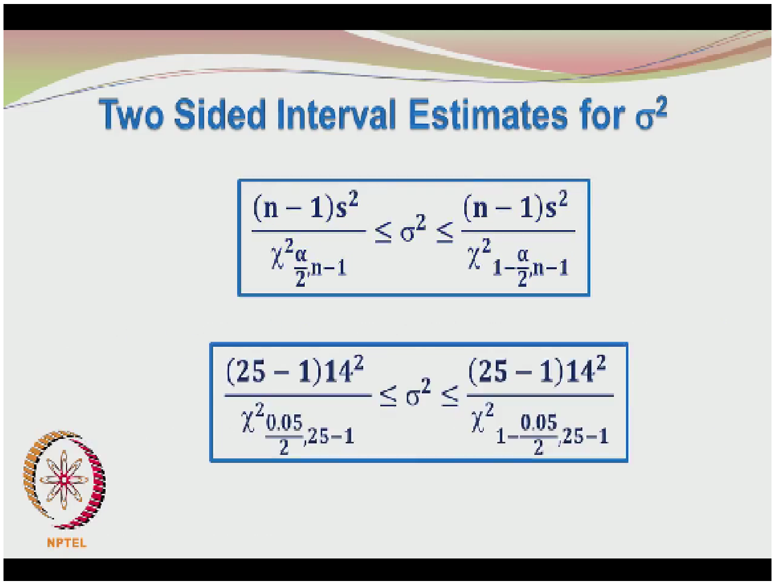
Two Sided Interval Estimates for σ^2

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

Here $\chi^2_{\frac{\alpha}{2},n-1}$ and $\chi^2_{1-\frac{\alpha}{2},n-1}$ are the upper 100 $\frac{\alpha}{2}$% and 100 $(1-\frac{\alpha}{2})$% points of the chi-square distribution with n-1 degrees of freedom.

Sigma squared is not known. So the confidence interval for sigma squared requires only the sample information, especially the sample variance. So we can find the upper and lower bounds for sigma squared. Chi-squared alpha/2 n-1 and chi-squared 1-alpha/2 n-1 are the upper 100 alpha/2% and upper 100*1-alpha/2% point of the chi-square distribution with n-1 degrees of freedom.

(Refer Slide Time: 28:27)



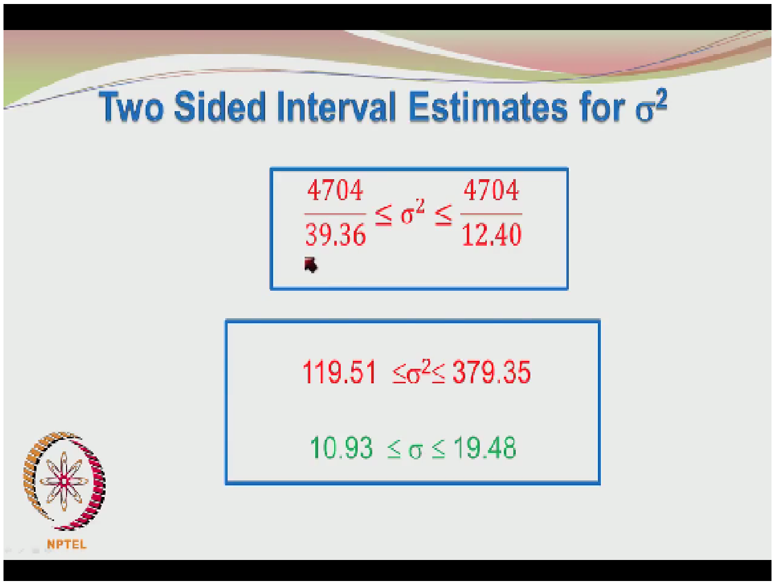
Two Sided Interval Estimates for σ^2

$$\boxed{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}$$

$$\boxed{\frac{(25-1)14^2}{\chi^2_{\frac{0.05}{2},25-1}} \le \sigma^2 \le \frac{(25-1)14^2}{\chi^2_{1-\frac{0.05}{2},25-1}}}$$

So plugging the numbers, n is 25, 25-1 s squared is 14 squared, yes. Sample standard deviation is 14. So s squared will be 14 squared. So how much is 14 squared, let us quickly do it. So 14 squared would be 196 and then we have to find out these values.

(Refer Slide Time: 29:13)



Two Sided Interval Estimates for σ^2

$$\frac{4704}{39.36} \le \sigma^2 \le \frac{4704}{12.40}$$

$$119.51 \le \sigma^2 \le 379.35$$

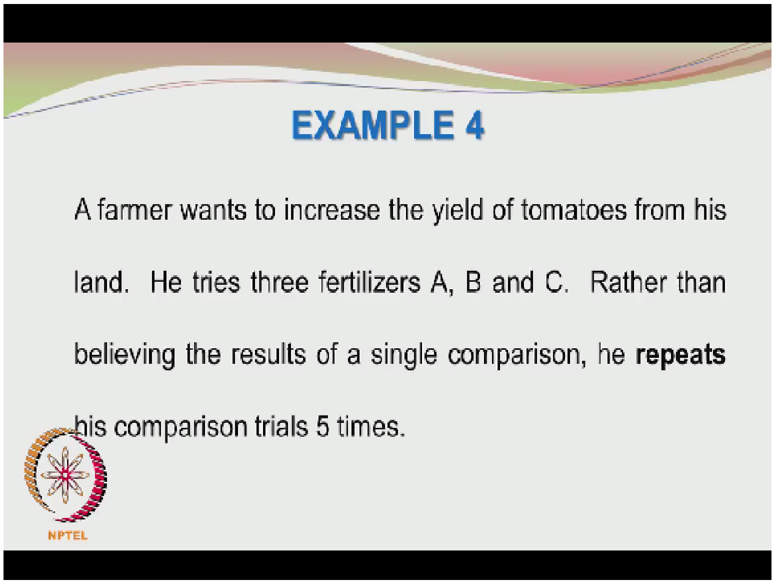
$$10.93 \le \sigma \le 19.48$$

And before we find the confidence interval for sigma, we have to find the confidence interval for sigma squared and 196*24. If you think 196 as 224*200 would be 4800. We are getting close to that 4704 and unfortunately we cannot do any quick calculations for the chi-square. We have to use the tables only. Remember the chi-square values are given in the tables. So please do not take the values given in the table and then square it, okay.

It is the chi-square distribution values, the probability values for the chi-square distribution which are given in the tables. So 39.36 for chi-squared alpha/2 n-1, chi-square alpha/2 n-1 was 39.36 and for chi-squared 1-alpha/2 n-1, we have 12.40 and so the confidence interval, the 95% confidence interval for the population variance sigma squared is 119.51 <= sigma square <= 379.35.

If I take the square root, if I take this as 121 square root of 121, would be 11, 10.93 <= sigma, <=, if I take this as 400, this would be approximately 20. So <= 19.48. It is also important for us to have some approximate calculations with us just to make sure that we did not punch the wrong numbers in the calculator, right.

(Refer Slide Time: 31:32)



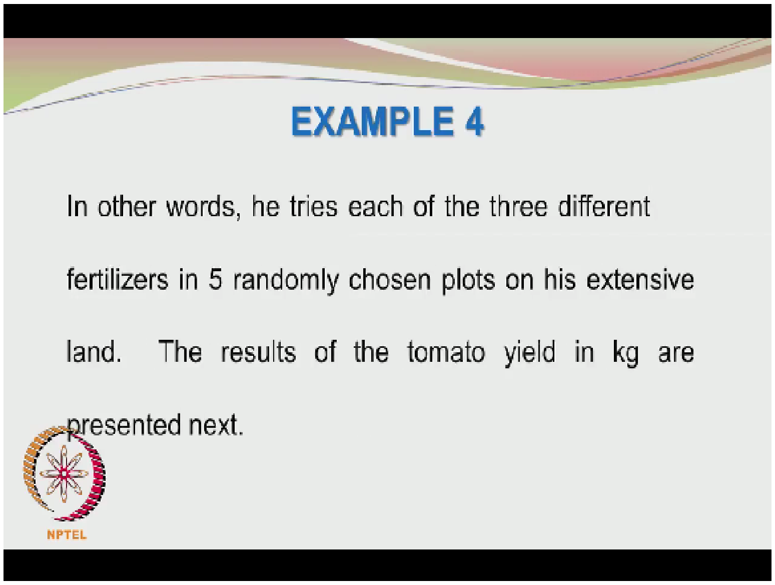
EXAMPLE 4

A farmer wants to increase the yield of tomatoes from his land. He tries three fertilizers A, B and C. Rather than believing the results of a single comparison, he **repeats** his comparison trials 5 times.

Let us now move on to the next example. Here we have a farmer trying to increase the yield of tomatoes from his land. He tries out 3 different fertilizers A, B and C. Rather than believing the results of a single comparison, he repeats his comparison trials 5 times. Well why tomatoes, why not anything else? Well I guess tomatoes now are very expensive, about Rs.50 a kilo.

So I guess if he can increase the yield of tomatoes by using the correct fertilizer, he may definitely make a killing. Anyway we have to do the problem given to us. So let us see whether the 3 fertilizers A, B and C are pretty much the same in influencing the yield or one fertilizer is better than the other and being skeptical, he is repeating the experiments 5 times, okay.

(Refer Slide Time: 32:46)

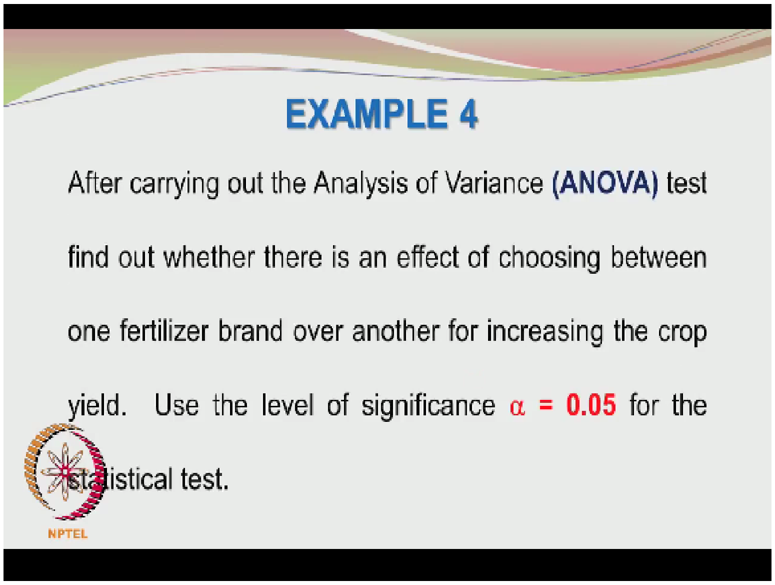


EXAMPLE 4

In other words, he tries each of the three different fertilizers in 5 randomly chosen plots on his extensive land. The results of the tomato yield in kg are presented next.

He seems to be having quite a bit of land. So what he does is, takes 5 plots or 5 fields and within each field, he tries out 3 different fertilizers, okay and he is noting down the yield in kilograms.

(Refer Slide Time: 33:18)

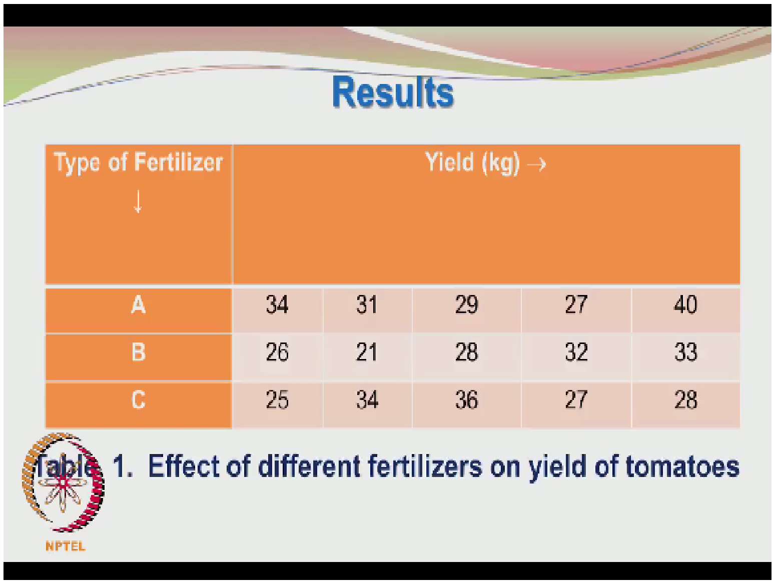


EXAMPLE 4

After carrying out the Analysis of Variance **(ANOVA)** test find out whether there is an effect of choosing between one fertilizer brand over another for increasing the crop yield. Use the level of significance $\alpha = 0.05$ for the statistical test.

What do we have to do with this data? We have to carry out an analysis of variance test and conclude whether there is an effect of choosing between one fertilizer brand over another for increasing the crop yield. We use the alpha value of 0.05 for the statistical test.

(Refer Slide Time: 33:43)



Results

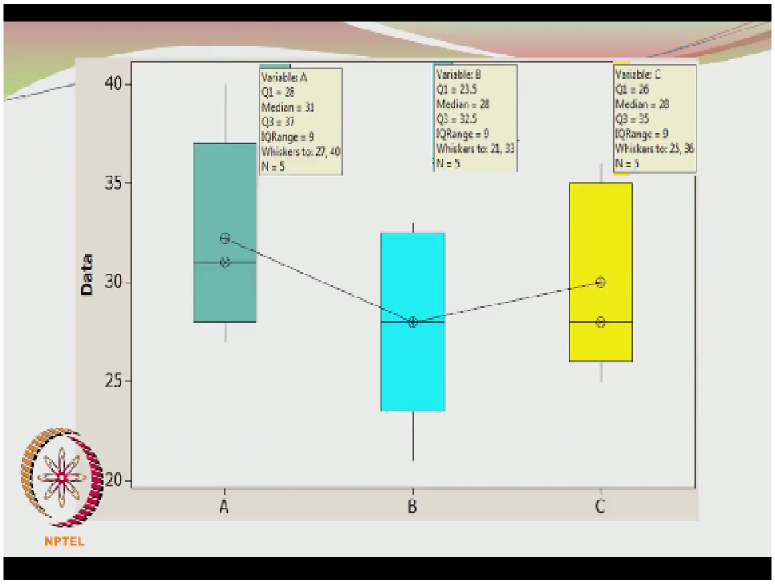
Type of Fertilizer ↓	Yield (kg) →				
A	34	31	29	27	40
B	26	21	28	32	33
C	25	34	36	27	28

Table 1. Effect of different fertilizers on yield of tomatoes

So this is the data and you can see that there is only 1 variable which is the type of fertilizer. We are having 3 levels of that variable or factor and these may be called as different treatments, treatment A, treatment B, treatment C and the yields are plotted in the table 34 31 29 27 40, 26

21 28 32 33, 25 34 36 27 28. So it looks on the face of it that well not a great difference between the yields, okay. So it is difficult to look at the data from an overall perspective and then make some conclusions. Let us see whether the box plots will help us.

(Refer Slide Time: 34:49)



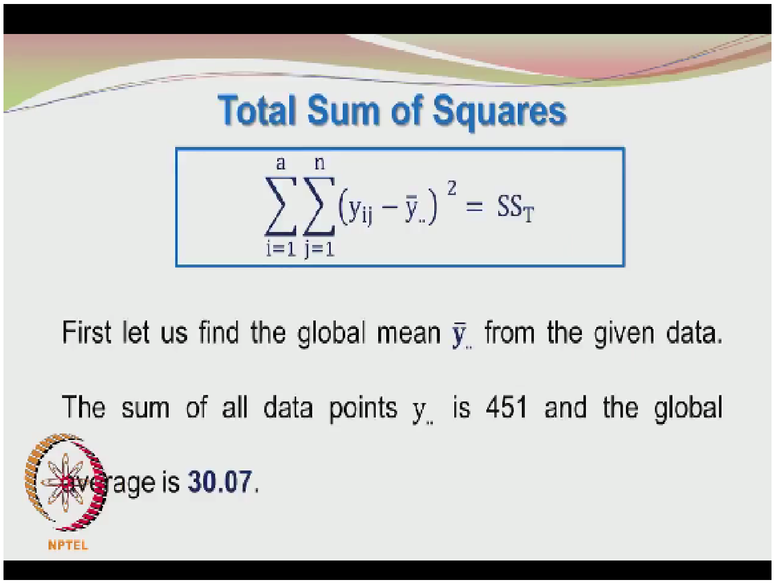
So these are the box plots and this is for fertilizer A. These are the values for the 5 repeats. These are the values for the 5 repeats and these are the values for the 5 repeats. So if you look closely, the first quartile is 28. The median is at 31. Mean is maybe 32, slightly above that. So mean and median are not coinciding for this particular case and the third quartile is 37. Interquartile range is 37-20 which is 9 and these are the whiskers.

Whiskers are going to 27 on the lower side and they are going to 40 on the other side because for fertilizer A, you are even getting 40 and even a low yield or lowest yield of 27 for fertilizer A. Similarly, for fertilizer B, the noticeable thing is the mean and median are coinciding and the whiskers are running from 21 to 33 and then you have the fertilizer C where the median and mean are quite different and the whiskers are sort of equally or equidistant from the first and third quartiles.

What I am trying to say is the whiskers from the first quartile is having a length almost equal to the whiskers length from the third quartile. Anyway so by looking at the data here that does not seem to be any great difference between the 3 fertilizers but that is a very subjective opinion.

Somebody else may claim that fertilizer A is better when compared to fertilizer B and fertilizer B is better when compared to fertilizer C. So let us see what conclusion we get after carrying out a proper statistical test.

(Refer Slide Time: 37:06)



Total Sum of Squares

$$\boxed{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^{2} = SS_T}$$

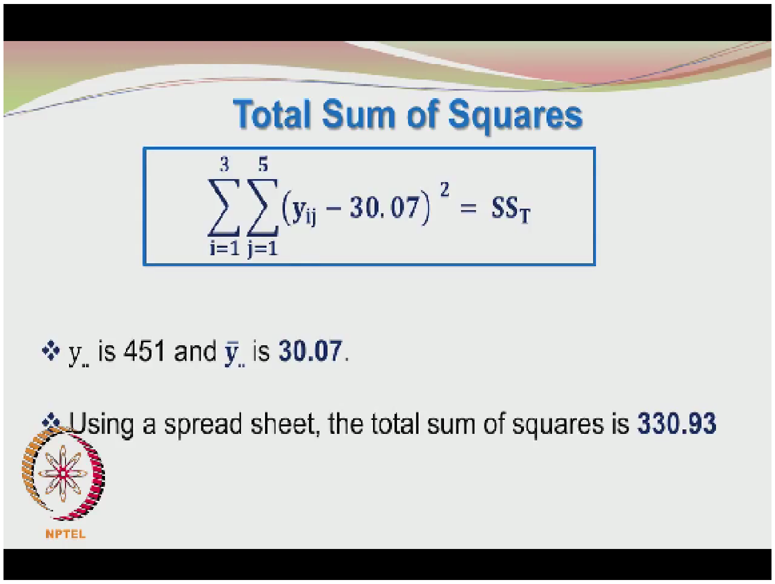
First let us find the global mean $\bar{y}_{..}$ from the given data.

The sum of all data points $y_{..}$ is 451 and the global average is **30.07**.

So we have to find the total sum of squares sigma=1 to a where a is the number of treatments, j=1 to n where n is the number of repeats for a given treatment and yij is the observation of the crop yield-y bar .. whole squared=total sum of squares. If you add up all these points, you should get 451. I am not going to do that. I do not have a calculator with me but I guess if you add up all these numbers 5*3 15 numbers you should get 451, 451/15 and that is approximately 30.

So 451/15 you will get approximately 30 or exactly 30.07. That is the global average. So I have found out y bar .. as 30.07.

(Refer Slide Time: 38:27)



Total Sum of Squares

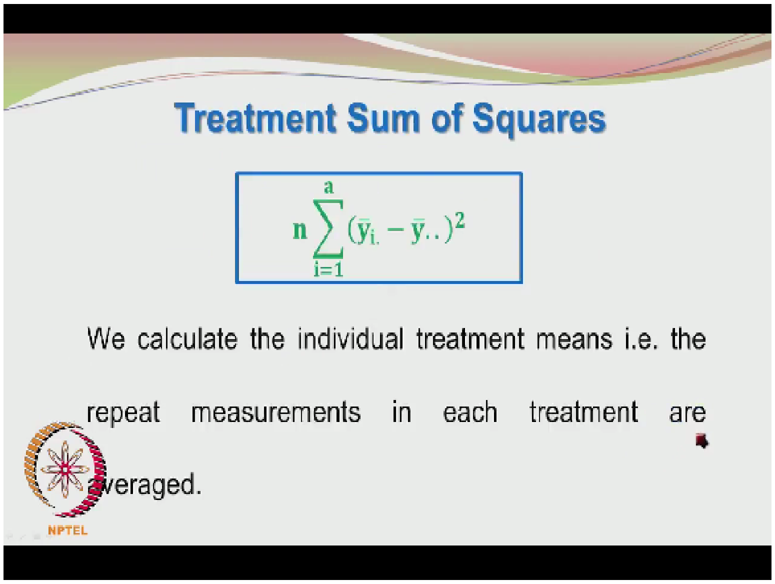
$$\sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij}-30.07)^2 = SS_T$$

❖ $y_{..}$ is 451 and $\bar{y}_{..}$ is 30.07.

❖ Using a spread sheet, the total sum of squares is 330.93

And we can use a spreadsheet to calculate all these values quickly. Montgomery and Runger have proposed shortcut formula which you may also refer to it but if you have a spreadsheet, I think, directly you may use the formula. So if I do that, the total sum of squares is coming to 330.93.

(Refer Slide Time: 38:54)



Treatment Sum of Squares

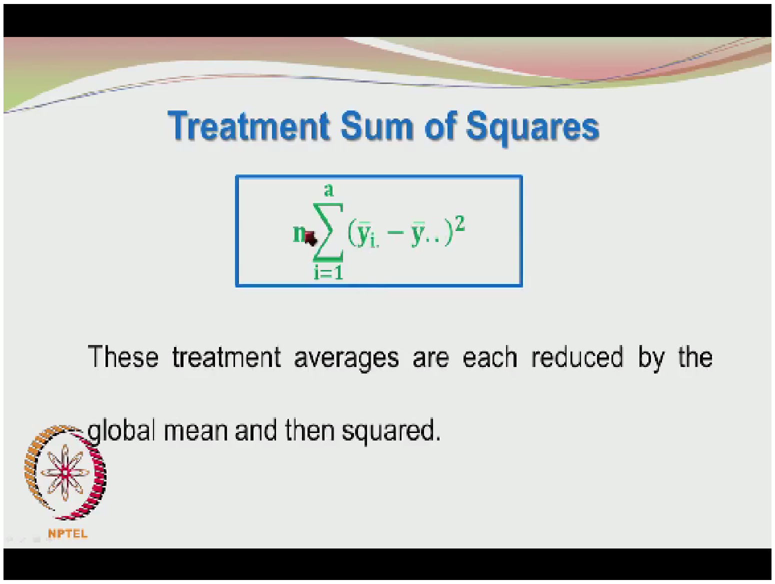
$$n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$$

We calculate the individual treatment means i.e. the repeat measurements in each treatment are averaged.

The treatment sum of squares is given by n*i=1 to 1 y bar i.-y bar .. whole squared. So we can calculate the individual treatment means the repeat measurements in each treatment are averaged, okay. So what I am trying to say is first we have to calculate y bar i.. How do I calculate y bar i.? For a given I, let us say for a given A fertilizer, I add up all these numbers and then divide by 5.

So I get 161, 161/5, 32.2. So that would be the average for the first treatment. For the second treatment, I take the average that would be 47 75 107 140 140/5 would be 28 and is it 28, let us check it out. Let us look at the box plot. Median is 28, median and mean are coinciding. So it is indeed 28, so no problem with that. This is quite simple, 59 95 122+28, it is 150, 150/5 is 30. So this average would be 30; 32.2, 28 and 30, right. So we have found the treatment means.

(Refer Slide Time: 40:54)



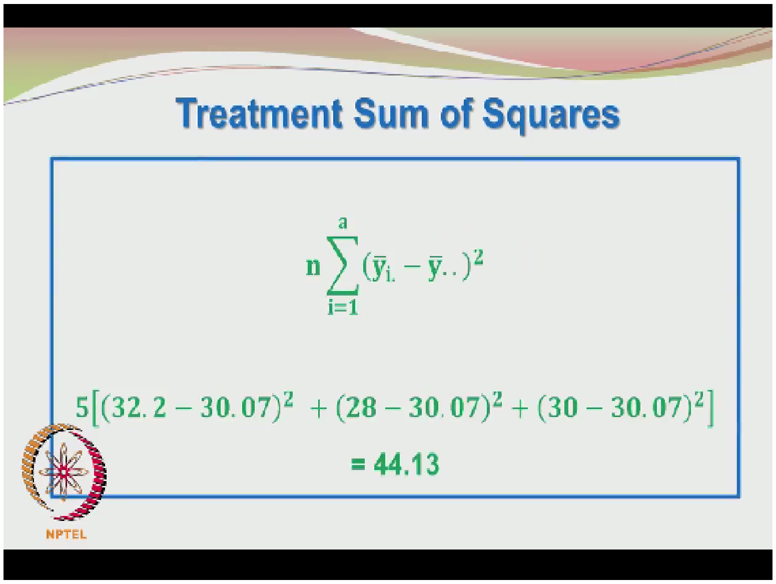
Treatment Sum of Squares

$$n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$$

These treatment averages are each reduced by the global mean and then squared.

And then we have to subtract the global average from each of the treatment means and add it up and then multiply it by the number of repeats. This will give us the treatment sum of squares.

(Refer Slide Time: 41:08)



Treatment Sum of Squares

$$n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2$$

$$5[(32.2-30.07)^2+(28-30.07)^2+(30-30.07)^2]$$

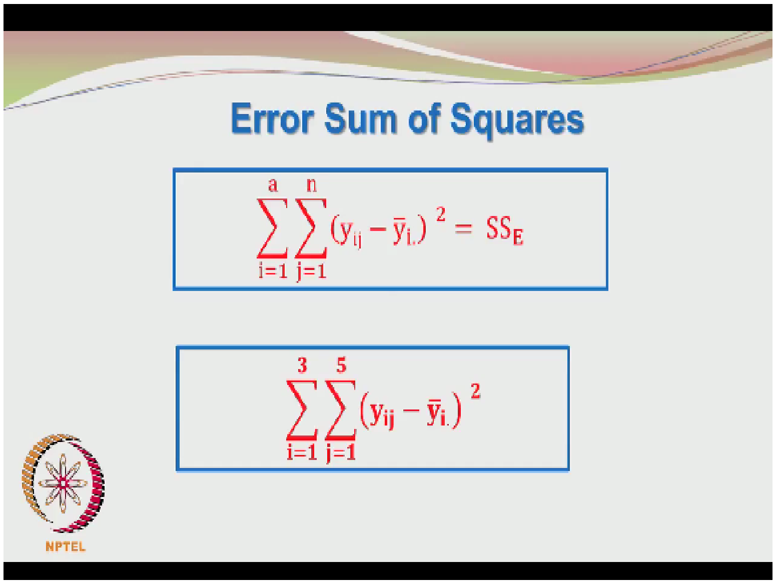
$$= 44.13$$

So n is 5, 32.2-30.07, the global mean is 30.07, 28-30.07 for the second treatment. For the third treatment is 30-30.07 and squaring all these deviations from the global mean and then multiplying by 5, I will get 44.13. So we have found out the total sum of squares. We have found out the treatment sum of squares. Next we have to find the error sum of squares. Since we know the treatment sum of squares and the total sum of squares, we can subtract the treatment sum of squares from the total sum of squares to get the error sum of squares.

This is correct provided you have accurately done the calculations. It is always good to have an independent check when you are doing these kind of problems. So what I would rather recommend is to carry out the calculations for the error sum of squares independently and then find out the number, add this error sum of squares with the treatment sum of squares to get the total sum of squares, compare the total sum of squares with the value you have already obtained.

So this will tell you that your calculation are in order. So I would rather go for the error sum of squares in an independent manner.

(Refer Slide Time: 42:39)



Error Sum of Squares

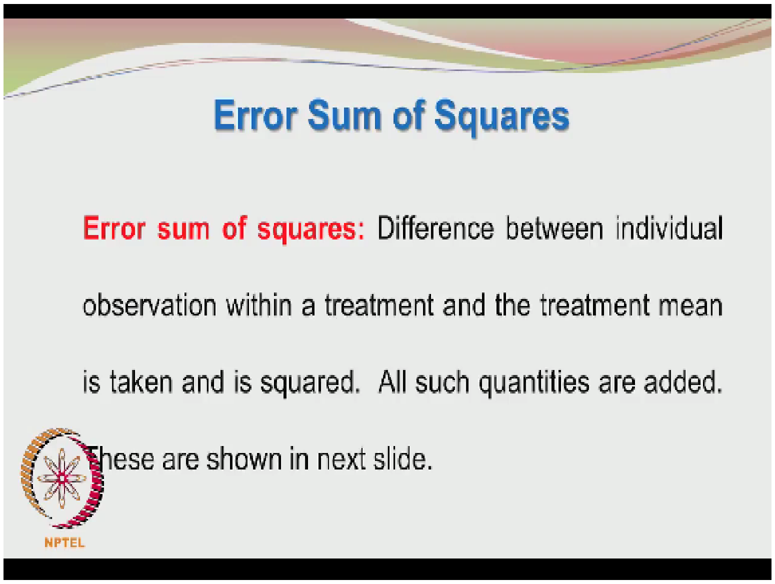
$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^{2} = SS_{E}$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{i.})^{2}$$

The error sum of squares is yij-y bar i. whole squared summed over i and j running from a and n respectively. So we are having the error sum of squares. What I am doing is from each observation, I am subtracting out the treatment mean corresponding to the appropriate treatment, okay. So fast I will fix the value of a treatment for the n repeats in that particular treatment.

I will find deviation of the observation from the treatment mean, then square it, then add it, then go to the next treatment i=2. Then I will again subtract the second treatment mean from each of the repeat values that will give me the deviations. I will square the deviations. Similarly, I will go and do it for the third treatment. This will get me the error sum of squares.

(Refer Slide Time: 43:36)

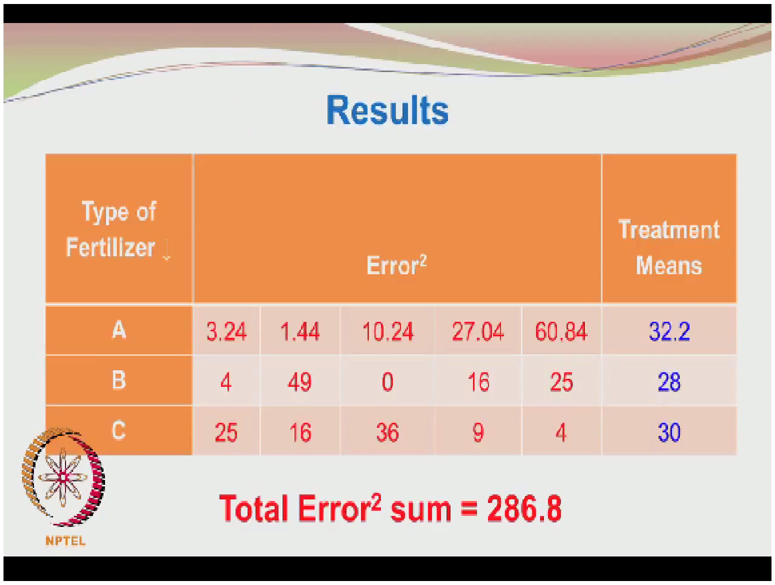


Error Sum of Squares

Error sum of squares: Difference between individual observation within a treatment and the treatment mean is taken and is squared. All such quantities are added.

These are shown in next slide.

(Refer Slide Time: 43:39)



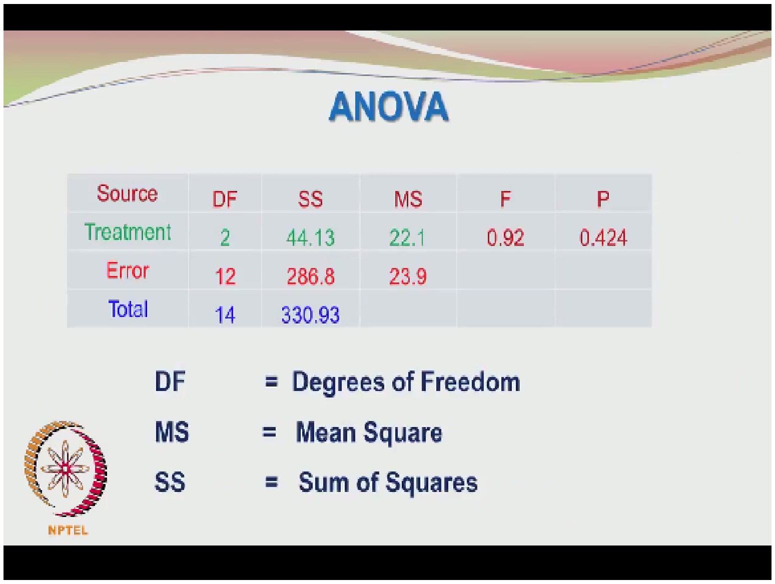
Results

Type of Fertilizer	Error^2					Treatment Means
A	3.24	1.44	10.24	27.04	60.84	32.2
B	4	49	0	16	25	28
C	25	16	36	9	4	30

Total Error^2 sum = 286.8

So when you add all of them, the error sum of squares comes to 286.8.

(Refer Slide Time: 43:45)



ANOVA

Source	DF	SS	MS	F	P
Treatment	2	44.13	22.1	0.92	0.424
Error	12	286.8	23.9		
Total	14	330.93			

DF = Degrees of Freedom

MS = Mean Square

SS = Sum of Squares

So we can put up a summary table as shown in the slide. We have the source, the analysis of variance table, the source of variation, treatment sum of squares, error and the total, the source of variation is the treatment, the source of variations due to error and this is the total variation. The degrees of freedom for the treatment is 2, 3 treatments are there and so you will be having 2 degrees of freedom.

Similarly, for the error, you will have a*n-1, a=3 and n-5, n-1=4. So 3*4 will give you 12 degrees of freedom for the error. The total degrees of freedom would be 14. Total number of experiments-1, total number of experiments is 3*5 15, 15-1 is 14. The sum of squares we just now found doing those calculations. I hope you enjoyed doing those calculations. Anyway and I hope you also got the correct answers, 44.13, 286.8 and 330.93.

I am just adding 286.8 with 44.13 to see whether it is indeed matching with 330.93, right. So I am going to divide the sum of squares by the degrees of freedom. So 44.13/2 would be approximately 22.065 and 286.8 where this sum of squares for the error will be divided by 12 which will give 23.9, right. Is that answer correct? 12*2 24 and 12*4 48. So approximately 24, okay that is correct. Now we can find out the F value. How do you find the F value?

We have to divide the mean square treatment by the mean square error and we will get an F value of 0.92. P value of 0.424 is also reported, that is quite interesting. What is this P value? We will see when we continue after a small break. So I hope you have understood these problems. The problems were done or created by me, okay. They are all fictitious problems. They have absolutely no implication to any incident or real person. Anyway we will see how to go for the next part of the problem. How to interpret the results and what conclusion we make? So see you shortly.