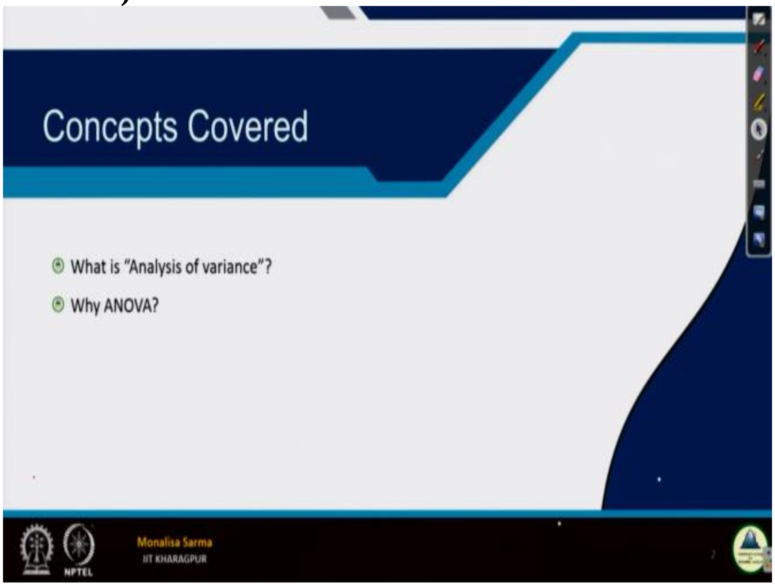
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 32
ANOVA - I

Hello everyone so, today we are starting a new topic that is analysis of variance in short we call it ANOVA. So, it is not exactly a new this is also another statistical inference technique, but here this is used for when we try to compare more than two population as sincerely then we use analysis of variance and the thing is that this topic before I start it is a bit complicated than the topics which we have covered previously.

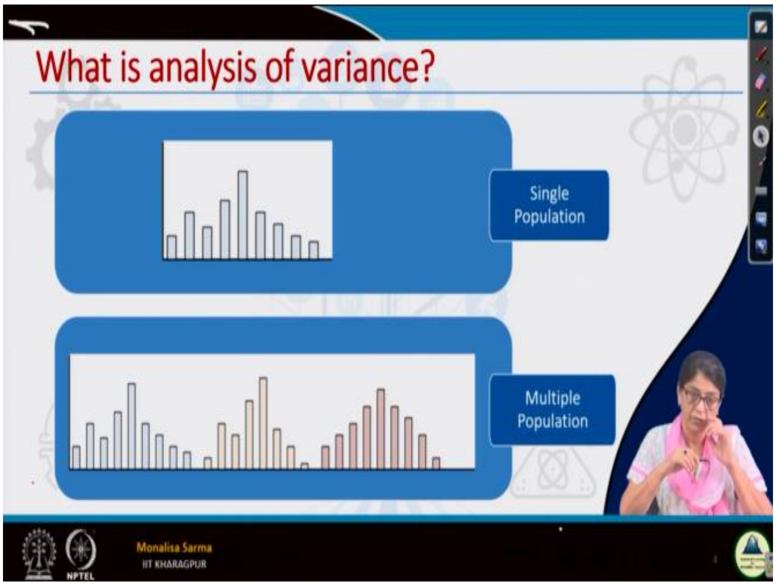
So, you will have to listen to it very minutely and you may have some doubts no issue that doubts can be cleared in the doubt clearing sessions there is no issue and even you can read through the textbooks also I have mentioned the textbook specific textbook from where I have taken this.

(Refer Slide Time: 01:15)



So, now, coming to the things what I will be covering in this class first is what is analysis of variance in short what is ANOVA and then second, why ANOVA we will be covering basically these 2 topics in this lecture.

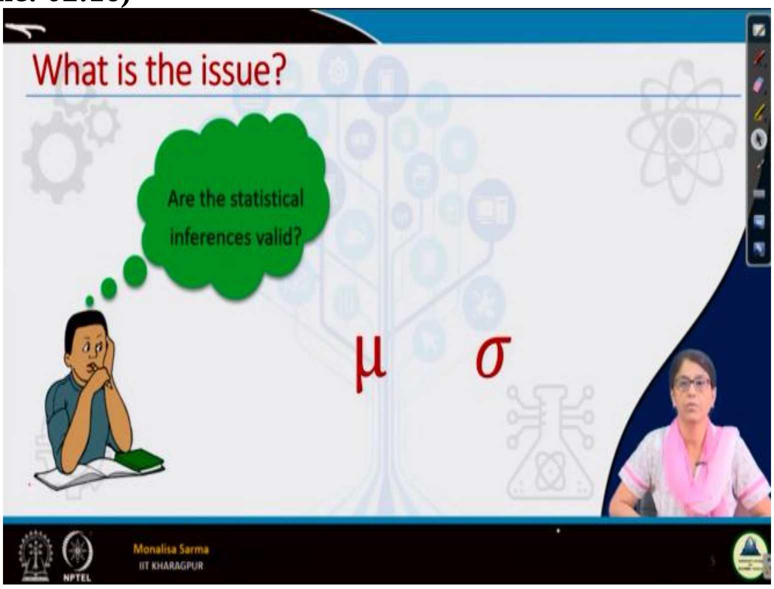
(Refer Slide Time: 01:31)



So, till now, what we have seen is that we have tried to infer what to say infer about a single population, we try to infer mean of a single population variance of a single population, then again we have tried to infer the mean of two populations and variance of two population whether I we did not try to infer what mean of two population we have when we used two population.

Basically two population we used to compare with the mean of two population are same greater less and then similarly, variance also we used, we tried to compare the variance of two populations for that we use app distribution if you can remember now, what we will do is that we want to compare multiple proportion more than two populations.

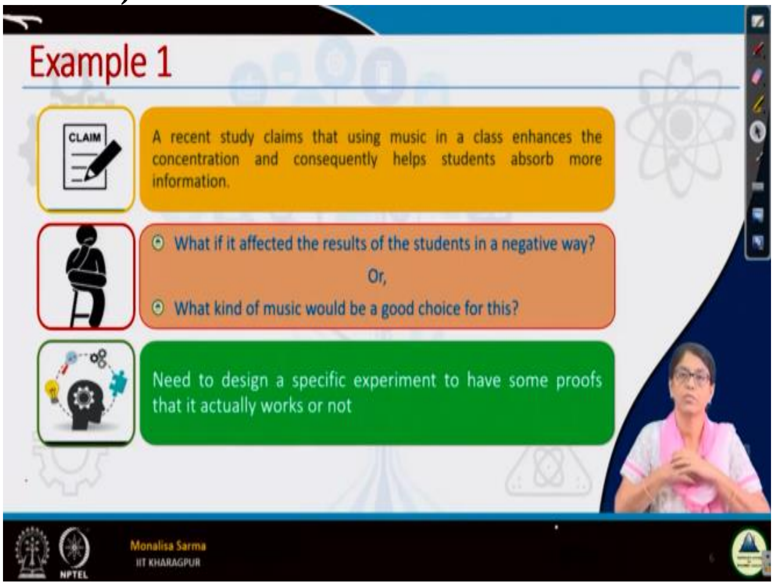
(Refer Slide Time: 02:16)



So, whatever concepts we have used while comparing two population or single populations like we have we tried to find out the mean of the sample from there we try to infer about the population mean similarly, we try to find out the variance of the population and from there we try to infer about the population variance all those the way we have done is it now valid we can we do it the same way.

But we have done for single population and two population can we do it the same way? Let us see that now.

(Refer Slide Time: 02:50)



So, for that let us start with an example first. So, what is the example? A recent study maybe in the newspaper it has come or somewhere it has come like, a recent study claims that using music in a class enhances the concentration and consequently helps students to absorb more information. So, study has been done study means on a sample of students basically, some organizations is trying has done a study.

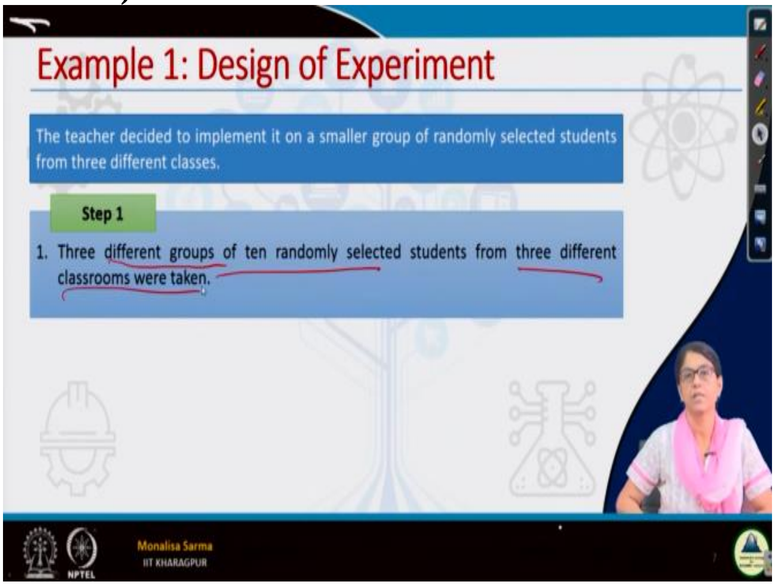
And from there what it has found is that the conclusion is that so, if we use music in the class, it enhances concentration and when it increases the concentration then it consequently helps students to absorb more information students performs better. So, study has concluded there concluded that now, we are not convinced with this study, we want to test it, is it really true? So, we are also interested in finding out what if it affected the results of the students in a negative way?

The study has told that this music has improved the concentration increase the concentration and students have started performing better, but then is it really true? Is it we if we put music in the class, will it really affect the performance? It is affecting the performance but is it effective in a positive way or negative way? That is the first thing. Second thing, what kind of music would be a good choice?

Music means there can be any music. So, what kind of music will help the students will have an impact positive impact or what kind of music will lead to a negative impact. So, for that, we need to design a specific experiment to have some proof that it actually works. So, study claim that we are not satisfied with the study, we are not sure of the study so we wanted to find out.

So, for that we really need to do an experiment to have proof that actually music actually increases the concentration.

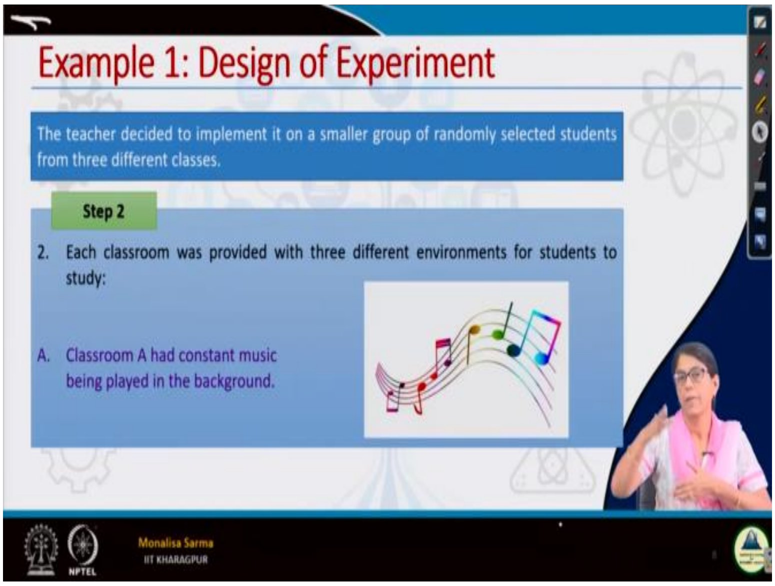
(Refer Slide Time: 04:49)



So, what is the design of experiment the teacher has planned? The teacher decided to implement it on smaller groups of randomly selected students from 3 different classes of the same level. So, what is teacher how he or she decided to carry out the experiment he has designed experiment in such a way what he has decided what he or she has decided to implement it on a smaller group of randomly picked students from 3 different classes are the same level.

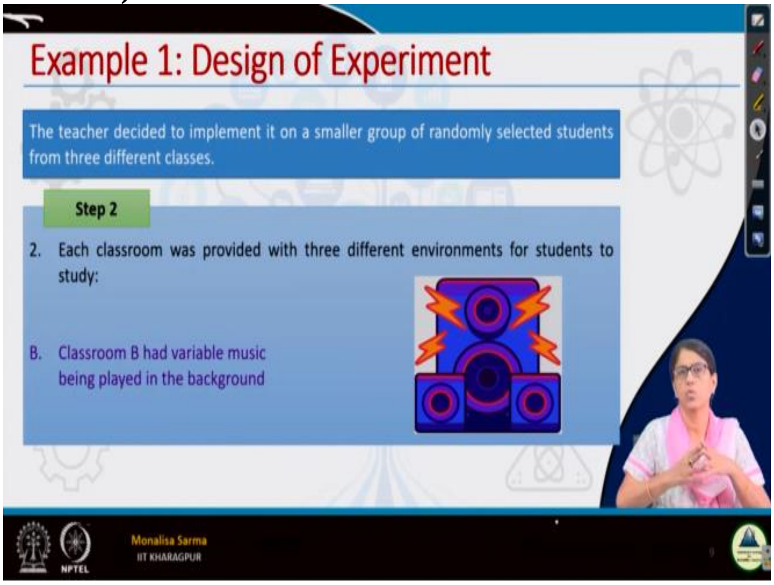
So that is the first step what is the first step? First step is create different groups of 10 randomly selected students from 3 different classrooms were taken, that is the first step.

(Refer Slide Time: 05:39)



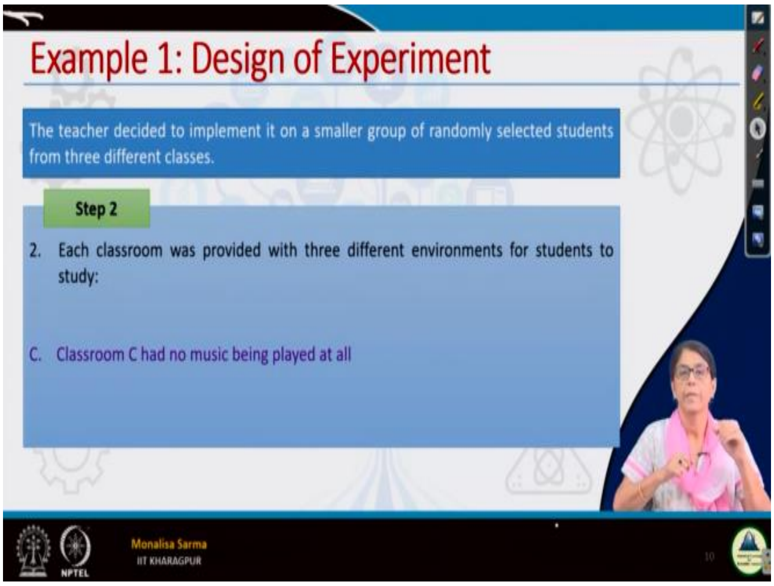
Then and so, we have taken 10 students from 3 different classes in 1 class, what the teacher do? The teacher has constant music being played in the background, while the classes are going on, at a certain scene that a constant music is constantly playing in the background. So, that is the first group he has total 3 groups. So, in the first group that is done.

(Refer Slide Time: 06:07)



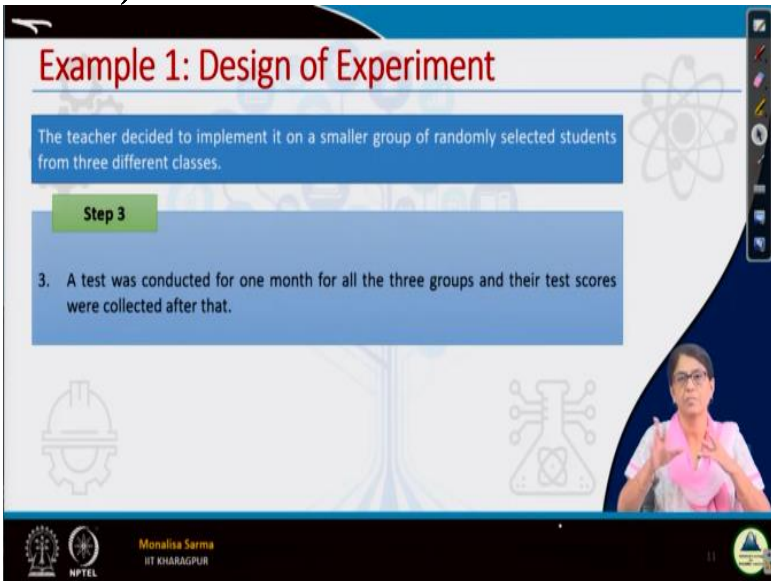
And in the second group, the class B had variable music in the background, sometimes very slow, sometimes loud music sometimes very soft music variable music is played in the background that is the second group of students.

(Refer Slide Time: 06:22)



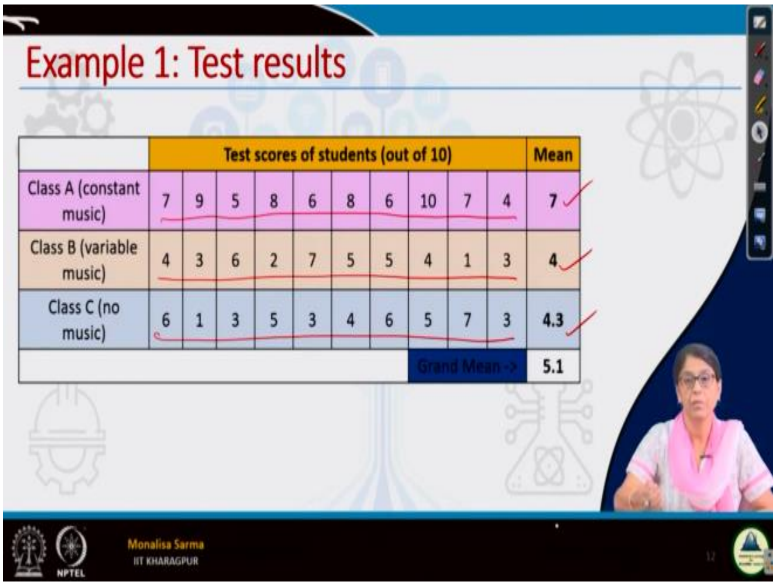
And third group of students had no music played at all. So that is how the teacher has designed the experiment, the teacher has picked randomly picked 10 students from 3 different classes of the same level. And the first group he has played the music constantly in the background, and the second group, the teacher has played random music and for the third group, the teacher has played no music at all.

(Refer Slide Time: 06:49)



So, test was conducted for 1 month, for 1 month testing is continued. And after that, the teacher has taken some sort of tests and for all the 3 groups and the test scores were collected after that.

(Refer Slide Time: 07:05)

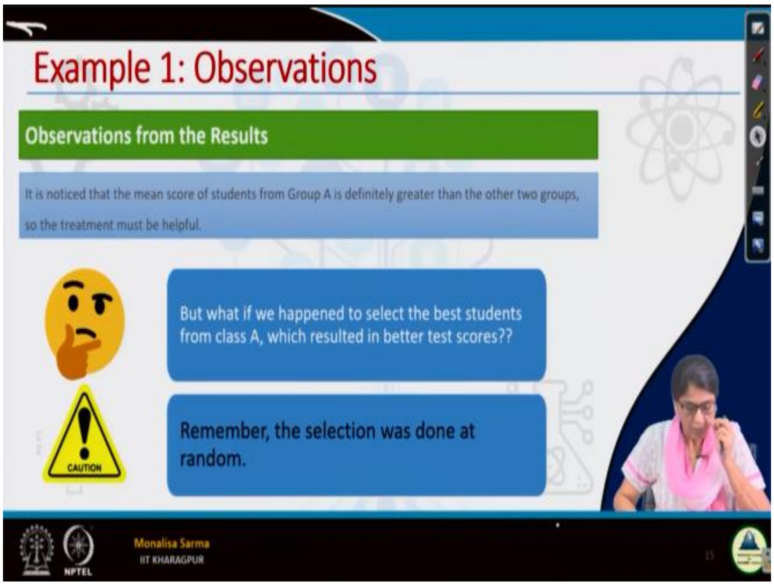


So, this is the test scores for 10 different students, this is for classroom A, these are marks out of 10 teacher has taken a test conducted a test out of 10 and this is the marks obtained by classes with the constant music is played these are the marks which is obtained by the class B where variable music, and this is the these are the marks which the teacher got where no music has been played.

And we see if we find out the mean, for the class A it is 7 class B it is 4 class C it is 4.3. So, you know how to find out the mean, we have already seen it, it is basically the average. So, from this mean, what we have seen is that directly if we try to infer from the mean, what we see is that class C has performed much better than class B and class C, class B and class C, we can see there is not much of difference, no significant differences there.

But class A has performed significantly different from class B and class C, if we see the mean from the mean, we can give that conclusion.

(Refer Slide Time: 08:13)



So, it is noticed that a mean score of students from group A is definitely greater than the other 2 groups so the treatment must be helpful. We have seen the results from the results, we have seen class A where we have played a constant music constant score, what they got the students what they got the mean score is significantly better than the other 2 groups and the class B and class C.

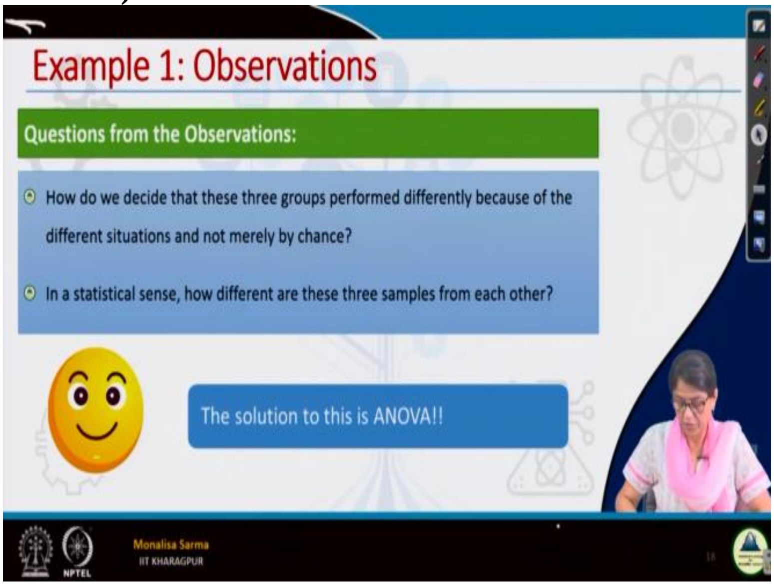
That means is the music really helpful? Means it makes us think, maybe the treatment is actually helpful. If we play music background in the thing, then maybe it is really helpful, maybe it is true. But what if we happen to select the best students from class A, because we have randomly picks the students is not it? Random means it is the probabilities is not it? We have just picked.

So, when we picked might be that we have selected all the best students for class A, that may happen, is not it? Or the other way around might be we have picked all the weaker students for class B and C might be the music is not having any effect at all just that we have picked very good students, some of the very good students for class A or maybe for class B and C, we have selected some very poor students, there may be some outliers in class B and C some students who are very, very poor.

That might have had the effect of bringing the mean to such a low value or class A we have there may be some outliers or some 2 or 3 students have scored really good that is again an outlier maybe which has bought as mean quite high, is not it? While we can trying to find out the mean we do not analyze the marks what they are getting, we do not analyze all the data points. We have seen that when we find out a mean or variance.

We do not analyze each and every data points we directly look at the mean value or the variance value wherever we are interested, whichever we are interested. So, that may happen that may be the case our selection may not be proper in a random selection, so, since the selection was done very much at random.

(Refer Slide Time: 10:18)



So, how do we decide these 3 groups perform differently because of different situation and not merely by chance. So, how do we decide that these 3 groups that have performed differently it is therefore, perform differently it is because of the different treatments and not because of mere chance, when I will call it what to say this groups that perform differently merely by chance.

Maybe when the students which are picked are not a uniform kind of students, it is actually the treatment that did not have any effect, we are getting different score because the students that we have picked may not be uniform students maybe that is why we got these different results, is not it? So, we just by seeing the mean we cannot be sure that the difference is because of the different treatment that we have done or because of this simple random chance.

In a statistical sense, how different are these 3 samples from each other there is a very important thing we need to find out. So, the solution to this is ANOVA, so, how can we find this this sort of things, how can we find out? The solution to this ANOVA like let me give you a 1 more example suppose we are interested in finding out the there are suppose there are 3 different treatment type available for a certain disease.

Like if let us take cancer, for cancer, there are different treatments for 3 different types of treatments are available. Now, what we and we do not know which treatment is better, which among this, which treatment is better, we do not know, we need to find out which treatment is better. So, maybe how will you do from the population of all the people who are suffering this and suffering the particular disease for which the treatments are there.

3 treatments are there for we will randomly pick 3 different samples, 3 different groups in each group, we will apply these 3 different treatments, and maybe the treatment which is efficient maybe we can find that by curing time, but now the time it takes to cure the disease, that can be our criteria to find out the efficiency of the disease. Time it takes to cure make the person free from the disease.

So, now, we have given this treatment, we have suppose this treatment, we have used the treatment for say around 2 months and after that, we are seeing the curing time how much percentage has been cured or it is totally cured or not whatever and based on that, we can now the condition is based on that directly suppose treatment a treatment B and treatment C similar to whatever example we have seen just now, this music example.

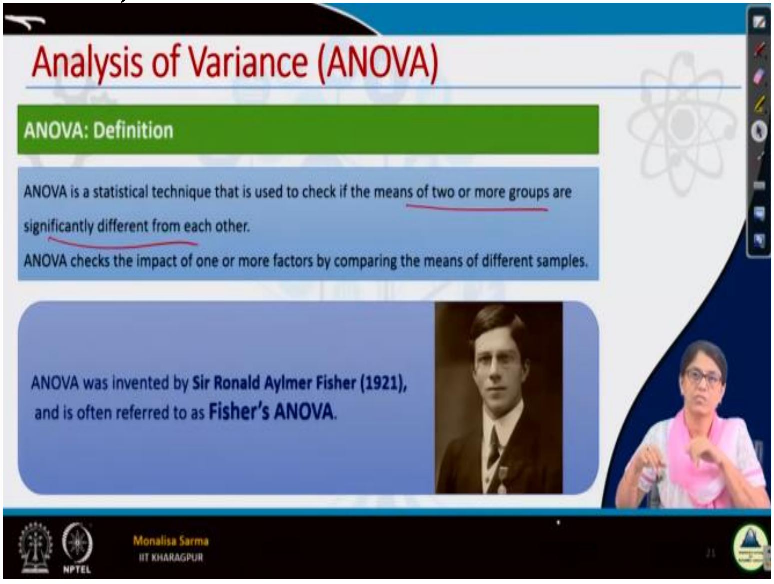
So, now suppose we have seen for treatment A maybe the person the curing time was quite less for treatment B curing time may be bit higher for treatment, C maybe a bit more higher or maybe same as B or whatever it is. Now, just from this, can we directly say that treatment A is better than treatment B and C similar to the example that we have discussed? Maybe yes, maybe no? Why maybe no?

Like when we have picked the people, there may be chances that among that, there are some patients who are very old, who have some other ailments. So, then for them, disease it really takes time to cure. So, maybe for B or C whichever has taken a long time to cure or so they are maybe they are there while picking the patients maybe we have picked such patients who has already some other ailments.

Because of that, maybe the curing time it is taking a longer time or maybe in the sample A for the sample for the group A for the first group where we have used treatment A maybe in that sample, there are some patient maybe who are already taking some other medication because of that maybe their what is the disease is almost partially cured before starting this treatment. If it is partially cured before starting this treatment, then what happens when the treatment starts they will it will cure soon.

I am not talking about all maybe some because some will have an effect on the overall mean. So, this sort of things may be there. So, directly if we see the mean time it takes to cure we cannot directly come to the conclusion that this mean this time is different only because of the treatment. There may be other factors to that. So, this we can find out with the help of ANOVA.

(Refer Slide Time: 15:06)



Now, the question is what is ANOVA? ANOVA is a statistical technique that is used to check if the means of 2 or more groups are significantly different from each other significantly different I have used this term significant before when I have discussed statistical inference I will not be repeating that. So, ANOVA is a statistic that is used to check if the means of 2 or more groups are significantly different from each other.

Now, what are the difference what we have seen in the recent music example? Are they are significantly different, now, what does significantly different here means we will have to see. So, ANOVA basically tries to find out if the means of 2 or more groups are significantly different from each other. So, what do we mean by significantly different we will come to that. So, ANOVA checks the impact of 1 or more factors by comparing the means of different sample impact of 1 or more factors now, what is this factors?

It checks the impact of 1 or more factors by comparing the means of different sample here in the music example what is the factor? Factor is the music we are playing different type of music, the constant music, very random music and no music the factor is music different type of music no music can also be we can call it as a null music, so that is a factor. So, again with their maybe the second factor.

Second factor maybe suppose in this example, instead of taking the class of the same level, if we take some say class, some students have class 10 and some student have say class 2, so means how music affects elder students how music affects younger students. So, there are 2 factors maybe music and age. So, that is what ANOVA checks the impact of 1 or more factors by comparing the means of different samples.

It was invented ANOVA was invented by Sir Ronald Fisher and is often referred to as Fisher's ANOVA by his name it is also referred to as Fisher's ANOVA, so, now, the question is why we ANOVA means analysis of variance very well. Now, here we have already spoke that, that used to check the means of 2 or more groups ANOVA is trying to check the means of 2 or more groups are significantly different from each other.

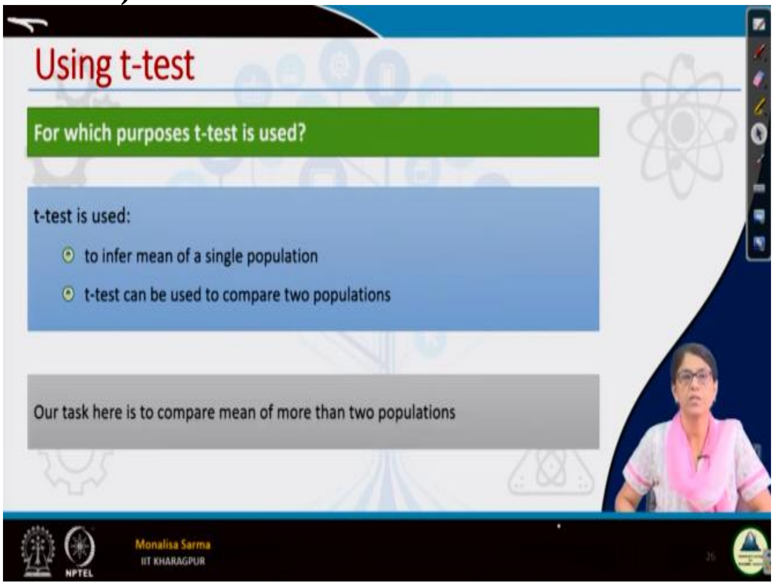
So, we are talking of means we are talking about comparing means of 2 or more groups. Then, at the same time, we are talking analysis of variance, how by analyzing variance, we can talk about the difference of 2 or more groups of mean that is something we need to see and we will gradually come to that now, before going to that first thing is that, now, this question when our objective is to compare the means of different more than two population that is our objective.

We are not from an objective we have nothing to do with a variance we just have to find out the we had just to compare the means of more than two populations and music example there were 3 population even the disease example what I have taken there also there are 3 population we had to compare 3 population mean of 3 population then why not use t test?

(Refer Slide Time: 18:35)

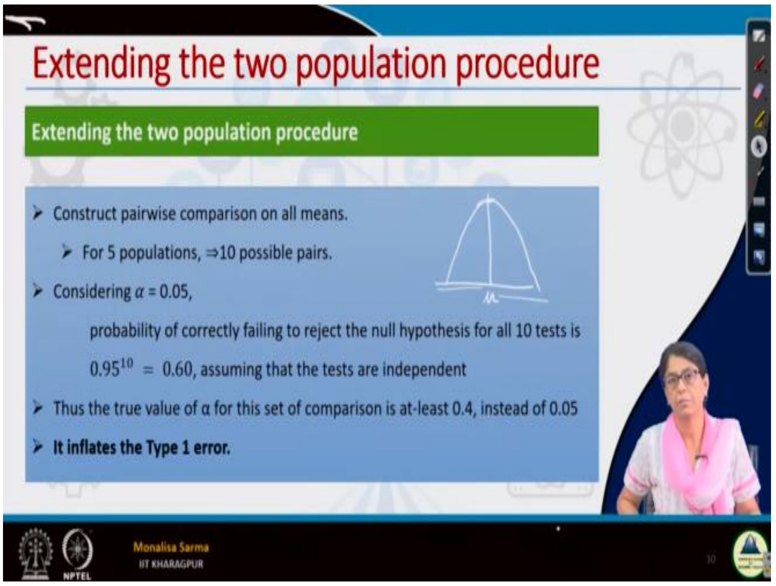
We have used details say we have seen t test details we could use for comparing 1 population details we have used for comparing two population as well is not it? So, why we are using t test? Why we are using analysis of variance? For comparing means this is something very confusing.

(Refer Slide Time: 18:55)



So, t test where we use t test is used to infer meaning of a single population t test can be used to compare two populations. Now, t test we can use to compare more than two population as well. What is the big task here?

(Refer Slide Time: 19:09)



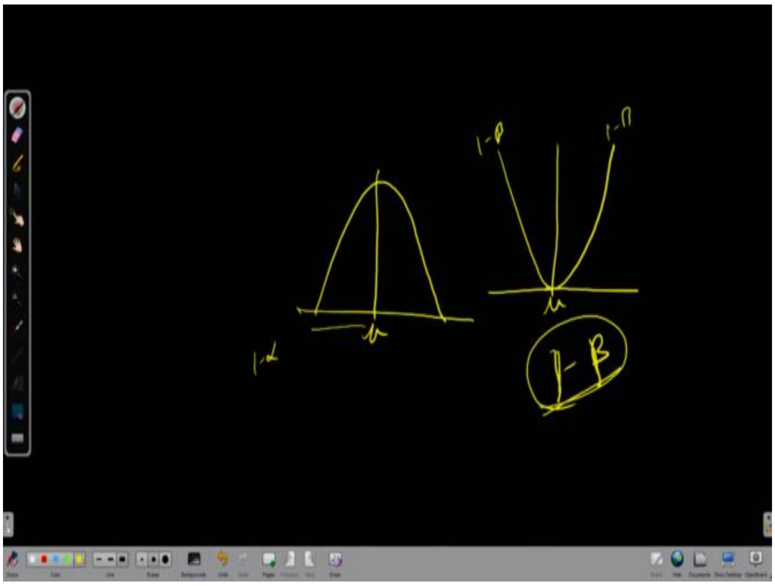
So, we will see what is the big task here? Suppose if you see if you are interested in comparing 5 population then how because t test we can use for comparing 1 population or at the most two population not more than that t test is designed in such a way now if you are interested in comparing mean of 5 population, so 5 population if you see all possible combination, it will be total 10 different possible pairs.

So, if you want to compare 5 population means of 5 population then we will have to do total 10 t test that is okay. If we have enough time it has, we will do 10 t test, that is not a problem, if we have enough time, but there are some other problems. So, what is that problem? Suppose we consider α is 0.05 that means our significance level is 0.05. So, if the significance level is 0.05.

Then what will be my β what will be my what to say maximum β maximum because β we have seen we have seen the β curve remember β curve is something this is not it? Here at this is basis of the mean where we will just see what is my β curve? B curve was something like this is not it? B is maximum at this point, when it is very near to the null hypothesis value.

That at that point this is β is maximum β will be minimum that is goes further away from the null hypothesis value, we have seen that and what is this? This is power curves.

(Refer Slide Time: 20:44)



So, this is my β curve so, if I talk about my power curve, this is my power curve. So, this is that this equals to μ 1 my value is very near to μ and what happens my β is highest, but when it goes further away from μ then my β goes gradually reduce and it goes to around 1 - α is not it? So, that is why higher β will also not happen it will not have a very negative impact we have seen that.

And this is my power curve my power curve my power is very less when it is very near to when my hypothesis value is almost equal to the actual value then my power is less and gradually my power increases is not it? So, at this point what is my power curve what is my value of the power it is 1 - β this is my maximum power I am gradually it will sorry, this is my maximum power my maximum power is 1 - β this is 1 - β and gradually it is coming down.

So, now, if we see here, so, if we consider for 5 pairs if there are total 10 possible pairs for each pair if the α is 0.05. Then what happens what is the power of each test power of each test will be 0.95 is not it? Power of each test will be point 0.95 so, for all t tests is so then power of all t tests total there will be 10 tests or what will be the power of total t tests all these tests will be independent is not it?

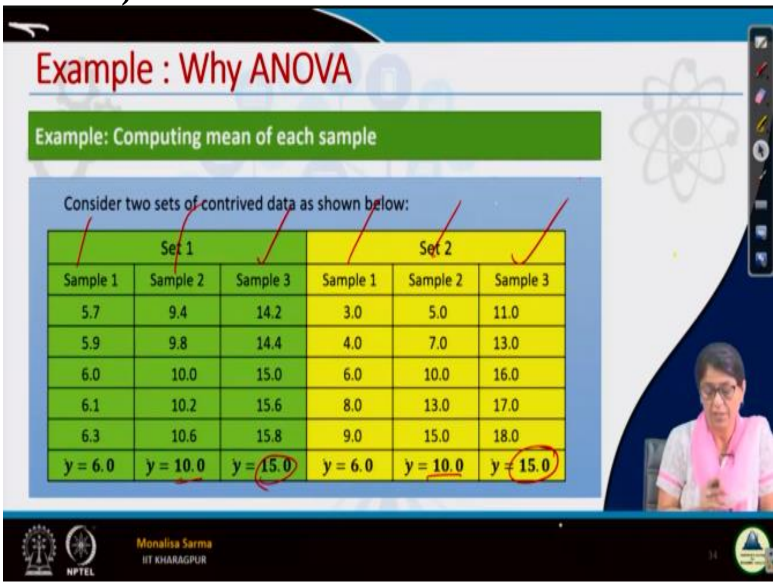
All these tests are independent we are independently trying to test different tests. So, since all the tests are independent, so, when we are interested in power of the total test, because the total test requires us 10 possible 10 tests the power of each test in total tests will be 0.95 to the power 10 and that is 0.60 and that will be the 0.60 is the power of my whole test. Now, if this 0.60 is the power of the whole test, then the true value of α is at least 0.4.

Because my power will be more than this when my power be more than this my value will come this will be this is at least 0.5 when α is at least 0.4. Then that means my α can be more than this also it can be α can be 0.5 also 0.6 also, when it will be this when it will become more when this will become more or less is not it? When the power curve is as I saw new power curve is like this.

This way it comes, comes, comes and power is very less at this point when the power is very less what happens my α will become more, more than 0.4. So, whereas I have started with an α of 0.05 see, that is the how my type 1 error gets inflated type 1 error a 0.4, it is too much to bear. And that is at least if my value can be more than that, because this 0.6 power is the maximum power and as I have seen power curve it this is the maximum at this point and gradually it comes down.

So, my power can come down to when my power will be lesser than that my α will go more and more up it will inflate the type 1 error and for 5 population we need 10 possible pairs for 6 population again more number much more for 10 population you can see imagine the number of pairs. So, you can imagine the type 1 error that we will have. So, t population will not be useful at all.

(Refer Slide Time: 24:34)



Set 1			Set 2		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
5.7	9.4	14.2	3.0	5.0	11.0
5.9	9.8	14.4	4.0	7.0	13.0
6.0	10.0	15.0	6.0	10.0	16.0
6.1	10.2	15.6	8.0	13.0	17.0
6.3	10.6	15.8	9.0	15.0	18.0
ȳ = 6.0	ȳ = 10.0	ȳ = 15.0	ȳ = 6.0	ȳ = 10.0	ȳ = 15.0

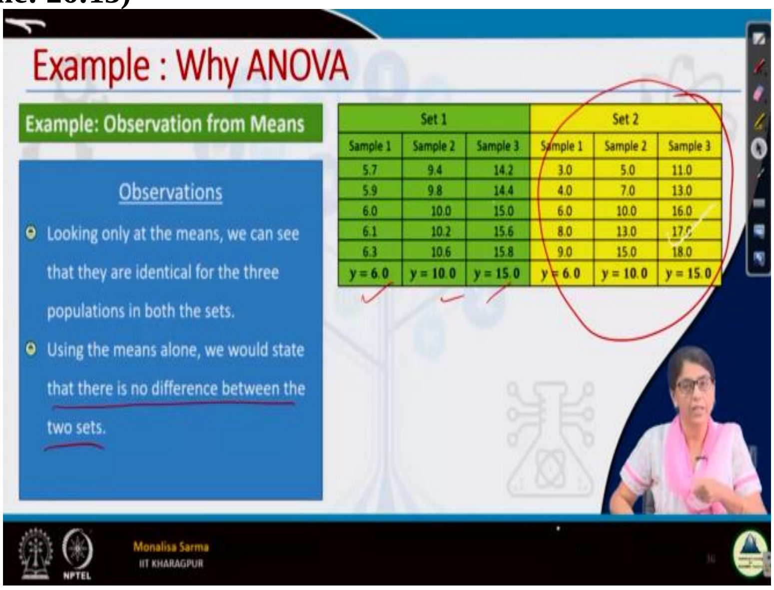
So, now that is why we have seen the t test is not at all suitable. We have seen that t test is really not effective when we try to compare more than two populations because it really inflates our type 1 error. So, what is the next so for that, we will try to see an example for that we have developed some data. So, the 2 sets of contrived data contrived data means we have made some data just for this example purpose only just consider there are 2 sets of data this is set 1 this is set 2.

So, set 1 has some from this set 1 we have 3 samples of 3 different populations. Similarly, for set 2 again we have 3 samples of 3 different populations. So, now what we have seen if we have tried to find out a mean of this sample, you see the mean of sample 1 of set 1 is 6 and mean of sample 1 of set 3 is also 6. Similarly, mean of sample to of set 1 is 10 here also it is 10 it is 10 here so, it is 10 here if you see this 15 here also it is 15.

So, if we just see the mean of all the both the samples, that means, we can do maybe we can take that both the sets, we have picked the sample from the same population means sample 1 of set 1 and sample 1 of set 2 we have picked on the same population again sample 2 of set 1 and sample 2 of set 2 also we have picked from the some other population. So, population B again sample 3 from set 1 sample 3.

From set 2 we have picked say from other population C this is this 2 are from the same population these 2 are from the same population, these 2 are from the same population if we just see the mean we can come to the conclusion, but is it actually true.

(Refer Slide Time: 26:13)

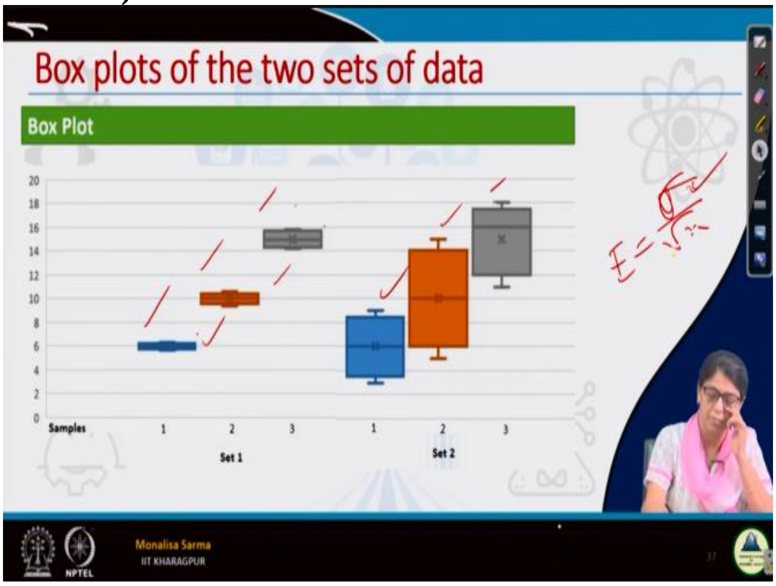


See, looking only at the means, we can see that they are identical for the 3 population in both data sets. Using the mean alone, we would state that there is no difference between the 2 sets. If we just see the main we can say that there is no difference between the 2 sets our intention while we are taking these 2 sets of data, our intention here is to find out is this forget about this data set.

If we just in 1 set our intention here is to find out whether this 3 populations are different the mean of 3 population are different, if we just see the mean here we can see that mean of this 3 population are really different 1 is 6 10 15 it is it looks to be significantly different, some say 2 also we are interested in finding out the mean of the 3 populations are different here also we are getting similar to set 1 we can say this 3 populations are actually the mainstream significantly different.

So, maybe that is 3 populations are different. If we can conclude this direct listing from the mean and from the mean we can say set 1 and set 2 are identical if we just see the mean.

(Refer Slide Time: 27:17)



Now, what we have done, I have drawn a box plot of this 2 set of graphs I have drawn the blocks plot, remember how we draw the boxplot I have discussed in the first lecture box plot, you please go see the lecture again if you have forgotten that thing so I will not explain it here. So, we have drawn the box plot of this 3, 2 sets of data. So, this is the box plot of set 1 for 3 different samples.

This is the boxplot of set 2 for 3 different samples. Now here we proceed from the box plot if we see the means are same here also the mean is 6 here mean is 10, here mean is 15, 2 sets mean are same, but you see here in set 1 the data's are bunched together it is very near or very close the variance of this data is very different here. But here it you will see that data's are very much variant remember, when we have discussed central limit theorem.

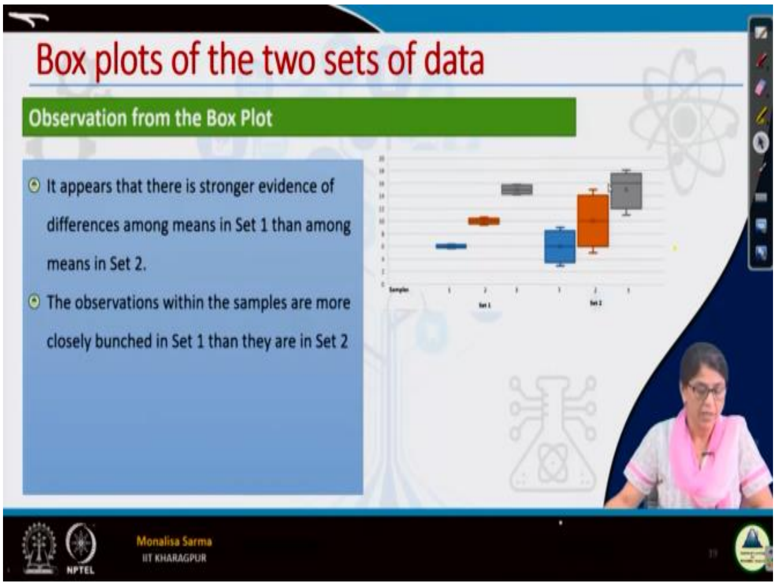
So, what I told you what was the if we take a sample from a population with mean μ and variance standard deviation σ, if you have taken a sample from a population with mean μ and standard deviation σ than the if we have the sampling distribution of the sample when we see the mean of the sampling distribution is the mean of the population mean and what is the variance of the sampling distribution variance of the sampling distribution was $\sigma / \sqrt{n}$ is not it?

And here also we have pointed out if the variance of the parent population was is quite more if the variance of the parent population is very high, then what happens that mean what we can that the mean is the on an average sample mean is equal to the population mean that particular statement is not a precise statement for that population statement, we cannot say this is at all a precise statement, there is real reliability of that statement is very, very low, when our standard deviation is very high it means when a variance is very high, is not it?

When a variance is very high, this is the standard error. So, the standard error of σ is very high means my error will be very high. What does the variance of the sampling distribution of mean indicates variance indicates the variance between the different sample means, if we pick different samples from the same population, and if we try to find out the mean the difference between this means is quite high.

If the difference between this mean is quite high, that we cannot just tell mean of all this mean is equals to the population mean we cannot tell claim that so when the variance is high, that is reliable of that statement comes down is not it we have seen that. So, here if we can use this in this 3 data you will see it is punch to get a variance is very less but variable here variance is very high.

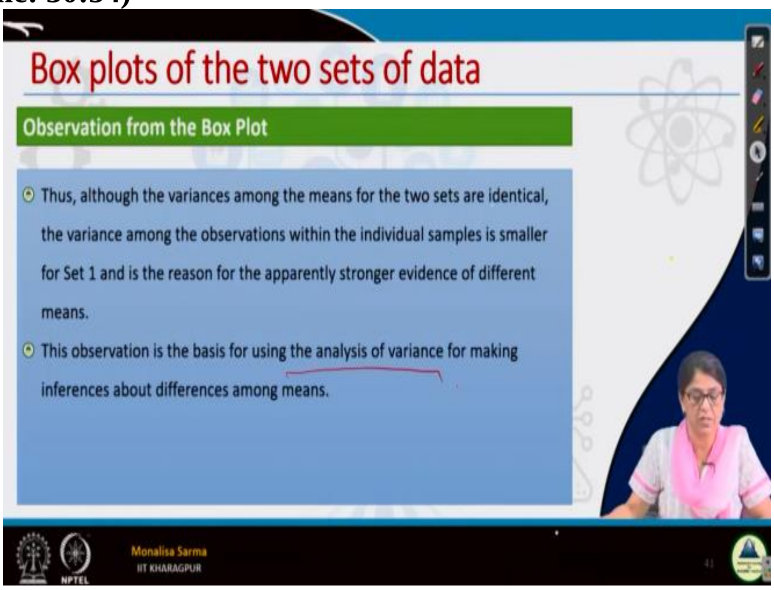
(Refer Slide Time: 30:03)



So, if we see this sets of data, it appears that there is a stronger evidence of difference amongst state 1 then among the means state of 2 in set 1 we can directly see that means of this data, there is a stronger evidence that the means of state 1 are really different with stronger evidence that this district population means are really different, but here though the mean is same to the first set.

But we cannot say that there is no strong if not at all a strong evidence that a mean of this 3 population are different, maybe this 3 populations, this 3 data belongs come from the same population, because you see the variance the observation within the sample are more closely bunched in state 1 than they are in state 2.

(Refer Slide Time: 30:54)



Thus, although the variance among the means for the 2 sets are identical, the variance among the observation within the individual sample is smaller for set 1 the variance among the observations within the samples, the variance within the observation it is very small, is not it? The variance among the observation within the individual sample is smaller for set 1 and this is the reason for apparently stronger evidence of different means.

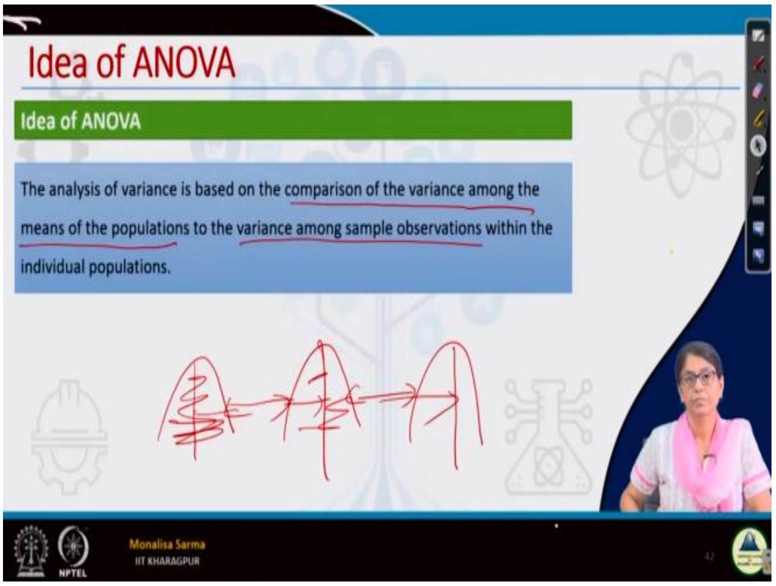
Now, we can understand why we need to use analysis of variance when we try to find out the compare the means of more than two population this observation is the basis for using the analysis of variance for making inferences about differences among means, we have seen this in the first set from the because of the variance, but what does it mean such thing, but in the first set, because the variants are very less.

The data's are really placed apart the from if we see the boxplot from the box plot, we could see that as a really plays a part it has stronger evidence that this the means that this 3 population, this 3 samples come from really 3 different populations, because the variants are very less from that only we could find out that this 3 cannot be from a single population, but in the other case set 2 it maybe we have it is the same type of population.

We have picked the data from the same type of population, but the data we have collected that is giving a different mean maybe because if we see the mean, if you see the variance it is so much the data there is so much varied as if it is it belongs to all the data belongs to 1 single population. So, we use analyzes we have used these variance within simple sample to find a way to tell something about the mean of the population that is the concept behind.

So, this observation is the basis for using the analysis of variance for making inference about the difference among means.

(Refer Slide Time: 33:01)

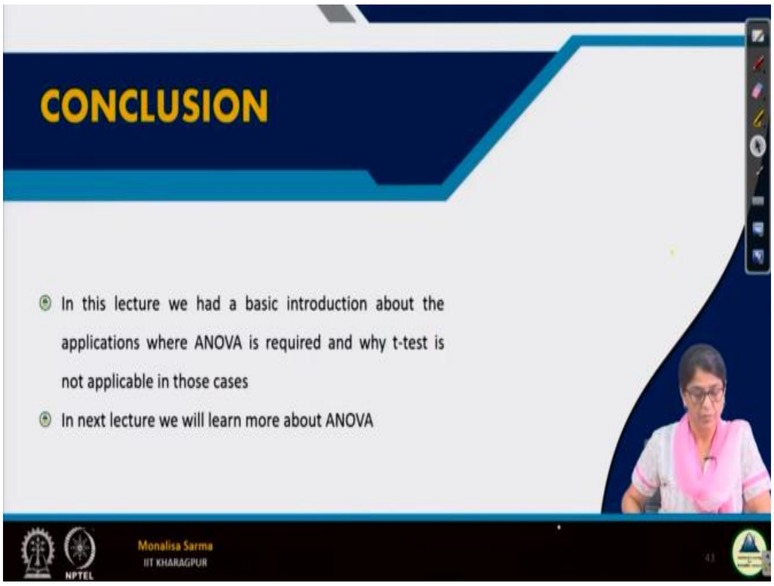


So, basically, what is the idea of ANOVA analysis of variance is based on the comparison of variance among the means of the populations, so, this is 1 population this is 1 populations. So, comparison a variance means, what is the variance among this mean suppose, this is this mean is here, this means here, this means here, so, difference between this main difference between this mean.

So, is based on a comprehension of the variance among the means of the population to the variance among sample observation, variance among the sample observation among the samples. So, well rank for in ANOVA what we do we try to find out 2 different means, or 2 different variants, one is variance within the different means and one is variance within 1 sample.

See, this is one sort of variants within the different means, and one is variance within the samples there is variance within this sample.

(Refer Slide Time: 34:20)



So, long way to go, so, now, it is almost, we have already exceeded the time, so I will stop this lecture here. So, in this lecture we have a basic introduction about the application way ANOVA is required and why I test is not applicable in those lecture and in the next lecture we will learn more about ANOVA.

(Refer Slide Time: 34:40)

And this is the reference which I talked about and thank you guys.