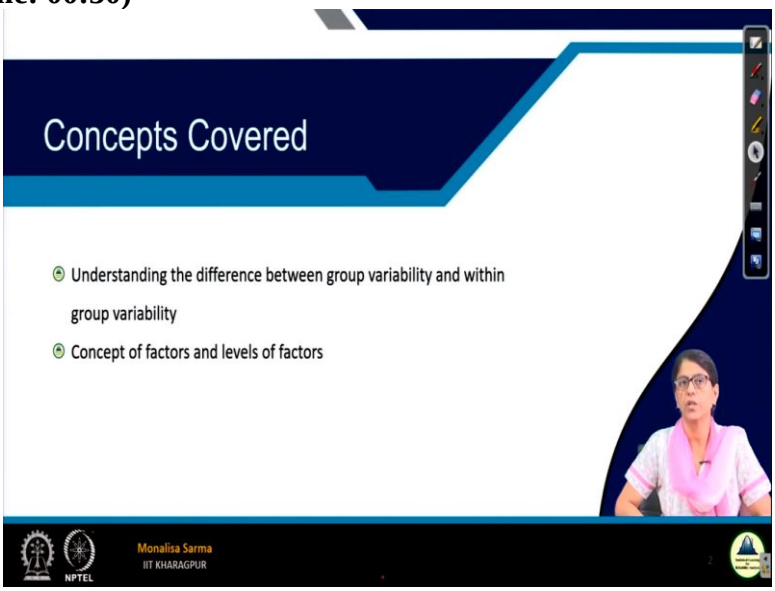
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 33
ANOVA-II

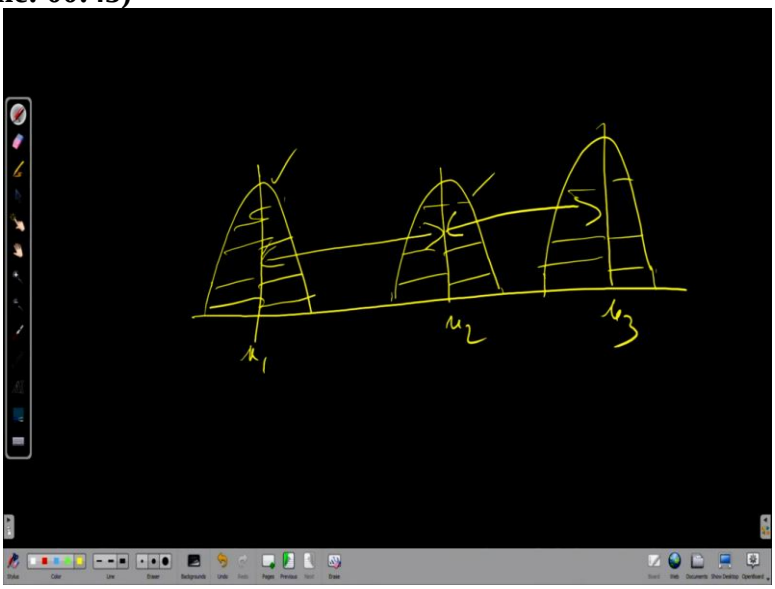
Hello guys, so in continuation of our lecture on analysis of variance, so today what we will see?

(Refer Slide Time: 00:30)



So today, we will see basically, we will try to understand the difference between group variability and within group variability and the concept of factors and level of factors, why I am talking about group variability and within group variability.

(Refer Slide Time: 00:45)



So, remember when I talked about last class when I talked about this, what is an ANOVA basically means, what ANOVA based on what I discussed that it is based on a comprehension of the variance among the means of the population suppose we have 3 different population from the my 3 different populations, this is my 3 different populations is sampling, sampling of 3 different populations.

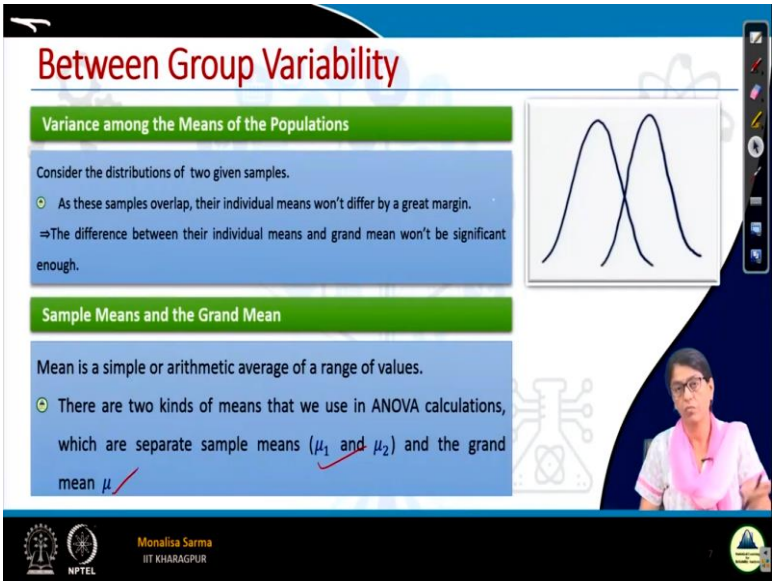
So, what is the mean of this my diagram is very bad and very poor in drawing actually. So, this is the mean of this population 1, this is μ 1, this is μ 2 this is μ 3. So, if when we are interested in the we are talking a variance when variants we are interested in variants within this treatment, because, why I am calling it treatment like, let us recollect this example, what we have taken the first example of music being played in the background.

So, what is the treatment? Treatment is the music and how in the music that means, one is with constant music another one is random music another one is no music, suppose this is with constant music. So, this is the mean this is with random music, this is the mean this is with no music, this is the mean. So, mean of this, this is the variance for this variance of 2 different samples, this is the variance of the sample 2 and sample 3.

So, we are interested in this variance as well as we are interested in variance within one samples, this is the variance this variance here also this is the variance. So, basically in ANOVA we say we are interested in finding out a variance within the treatment means as well as the variance within the samples, you know, what is variance you know, right I do not have to discuss that so, now, coming to that, so basically for that we will understand the difference between group variability.

What does group variability indicate that means a variant between 2 different treatments is not it. And within group variability within group variability means variance within a sample.

(Refer Slide Time: 02:59)



So, now, so, consider this distribution of these 2 different samples, this is one sample this is the second sample consider this 2 different samples, but we have seen this samples are quite overlapped. As this individual means would not differ by a great margin, these samples are overlap if we try to find out the individual means, what is this mean? This means, it may not be very much different means.

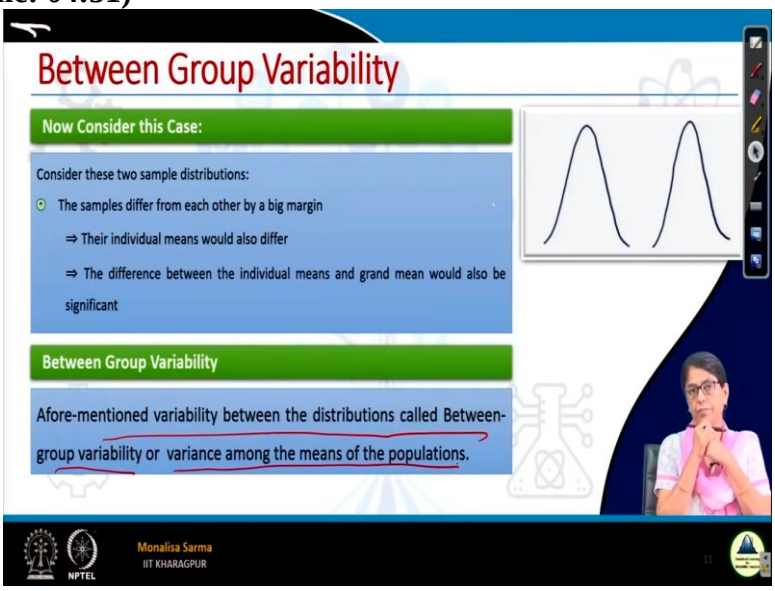
Because it is overlap there will be difference definitely, but it will not be much difference the difference between the individual means and grand mean on be significant enough now, what is grand mean? It is again a new term we are hearing here, individual mean we know what is the mean of a sample you know that grand mean when we let us go to the same example where I have discussed about music being played in the school.

So, for each class I have taken 10 data. So, 10 data for class A constant music is being played and 10 data for class B well random music, 10 data for class C but no music. So, totally how many data we have? Totally we have 30 data, so, mean of all this 30 we call it as a grand mean. So, if our data is in this form, where it is overlap and the samples overlap, then what happened there individual mean it will not be much different from the grand mean, grand mean understand mean of all the datas.

When the samples overlap there we will find that is obvious the grand mean will indivisible mean will not be much different from the grand mean. So, some what is simple mean we already seen. So, there are 2 kinds of means that we use in ANOVA calculation, one is

separate sample mean that is μ 1 μ 2 μ 3 mi 4 based on the number of population and the grand mean μ grand mean is the mean of all the data.

(Refer Slide Time: 04:51)

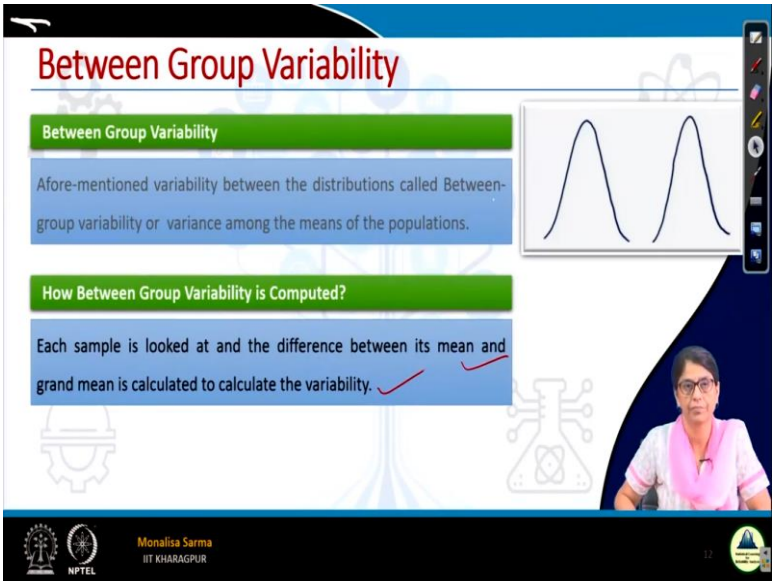


So, now you considered these 2 distribution sample distribution and again these 2 are the 2 different sample these are the distribution is, so these are distribution that quite apart earlier it was overlap, now it was quite apart no now since it is wide apart the mean what I get from this data, the mean what I get from this data. So, this mean will be different and moreover this mean will be quite different from the grand mean as well, grand mean will be sum of all this.

So, their individual mean would also differ, the difference between the individual mean and grand mean would also be significant in these cases. So, this is there is always a between group variability now, we are discussing between group variability is not it, this is one group this is another group that we are trying to find out the variability between this group, how we are finding out the difference of means is a variable variance upon the groups is not it?

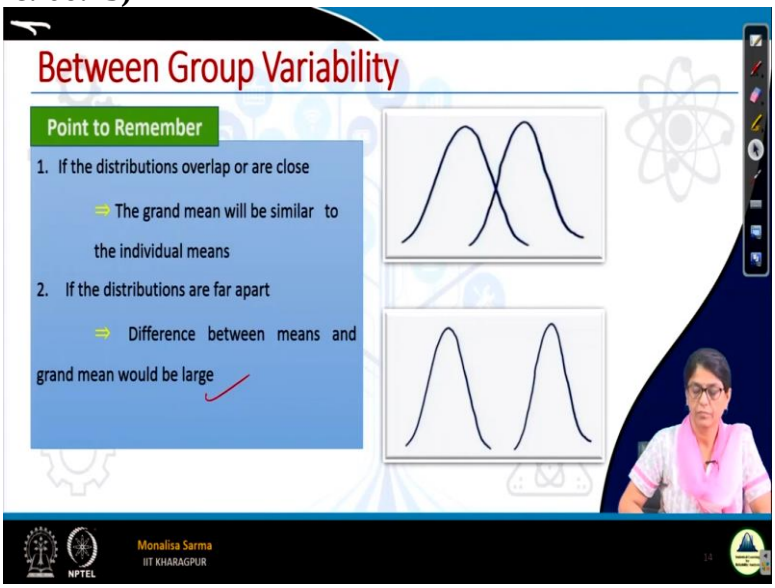
So, afore-mentioned variability between the distribution is called the between group variability or variance among the means of the population. So, this is called between group variability or variance among the means of the populations.

(Refer Slide Time: 06:00)



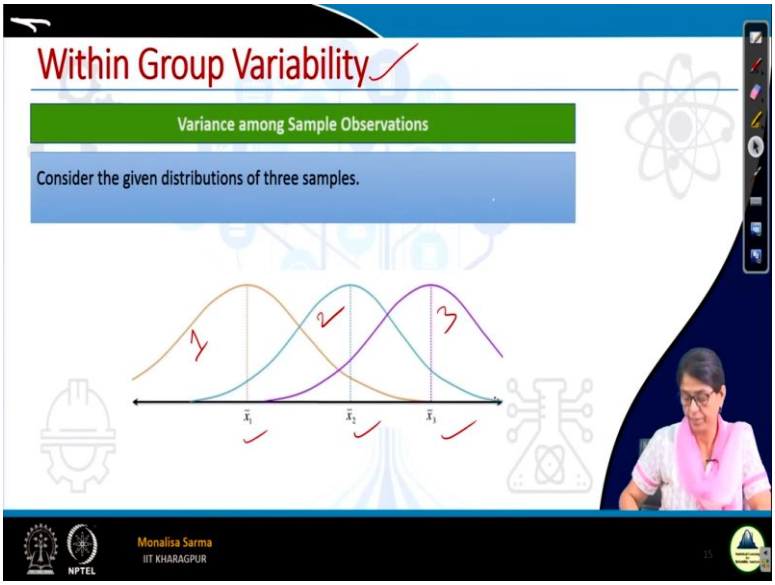
So, now how between group variability is computed when we are interested in finding the between is sample is looked at and the difference between its mean and the grand mean is calculated to calculate the variability, how we calculate the variability? We find out the grand mean from the grand mean we subtract each mean that is the variability is not it, that is called between group variability.

(Refer Slide Time: 06:23)



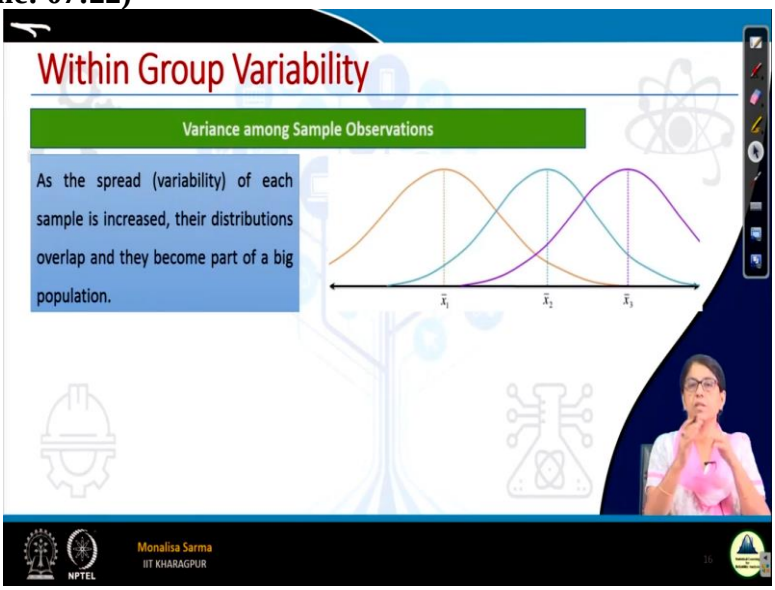
Now, so some points to remember if the distribution overlap are very close the grand mean will be similar to individual mean not exactly similar but similar. So, if the distribution of far apart distribution difference between means and grand mean would also be large.

(Refer Slide Time: 06:47)



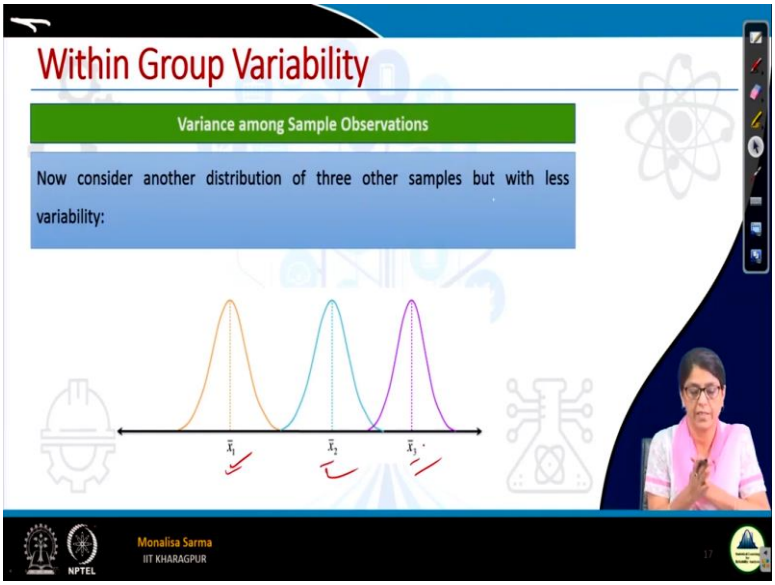
So, now, we are interested in finding out within group variability we have seen what is between group variability between means between 2 groups. Now, we will see within group variability, so, consider given distribution of 3 different samples here we have 3 different samples, this is the distribution of sample 1, this is the distribution of sample 2, this is the distribution of sample 3 see and see the mean of first is x 1 bar, mean of second is x 2 bar, mean of third is x 3 bar.

(Refer Slide Time: 07:22)



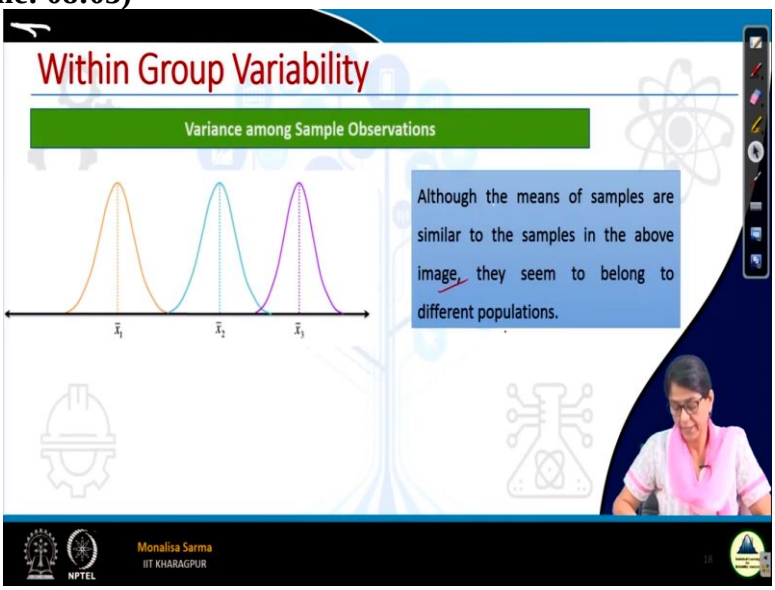
As the spread of each sample is increased, since the variability is quite high this so, what we see that distribution overlap, their distribution is overlapping and it is as if we have got it from a single population it is as if they are not a different populations. But it is as if it is from a single populations.

(Refer Slide Time: 07:46)



Similarly, now, consider another distribution of 3 other samples, see the means same x 1 bar, x 2 bar and x 3 bar. Here also what we got mean x 1, bar x 2, bar x 3 bar here variability is less but the mean is almost same.

(Refer Slide Time: 08:03)



So, although the means of the sample are similar to the samples in above the means, they seem to belong to different populations, here it is as if we got it from one population as if all the 3 samples belong to sample because there is overlapping there is no demarcation we cannot really distinguish that this population is different from this, again this is different from this we cannot really demarcate here, it is so much overlap.

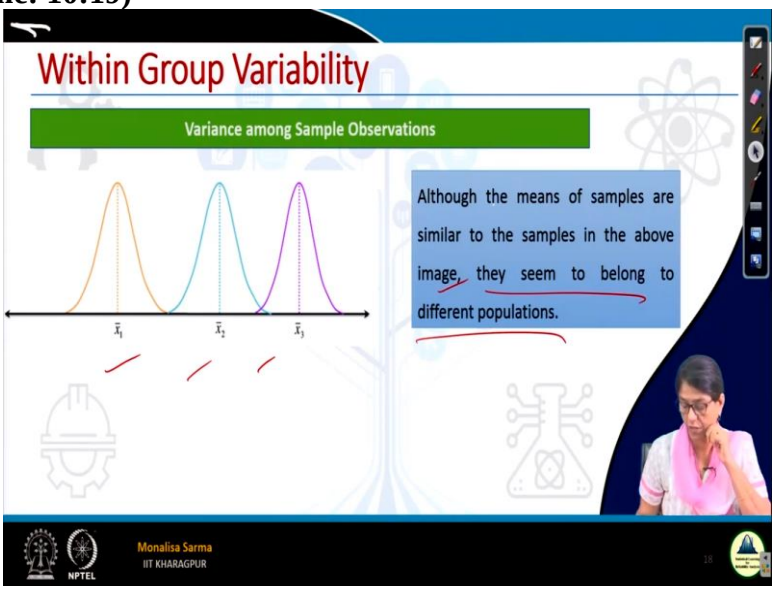
Why it is overlap because of lots of variants within a sample means the samples are very much varied like form very much varied example, let me give you an example like suppose, we are interested in finding out the mileage of a car same type of suppose we are interested

we are trying to see the miles of Verna car for one car we got suppose it is mileage is 20 kilometre and for another we got it is 15 for another we got a 10 another we got it 5 another we got it 8 is variance, it is a very much varied, the data is very varied.

That is what we call it is high variance. So, similarly, we took an example for another car for, for another make not this time to get Verna next time say we took it for another make say I20. So, again here also we try to see them and see the mileage we got a 25 20 15 10 22. So, as if this is belonging to the same population as if it is not 2 different cars, but the same car because the data are so varied, it is not that for Verna.

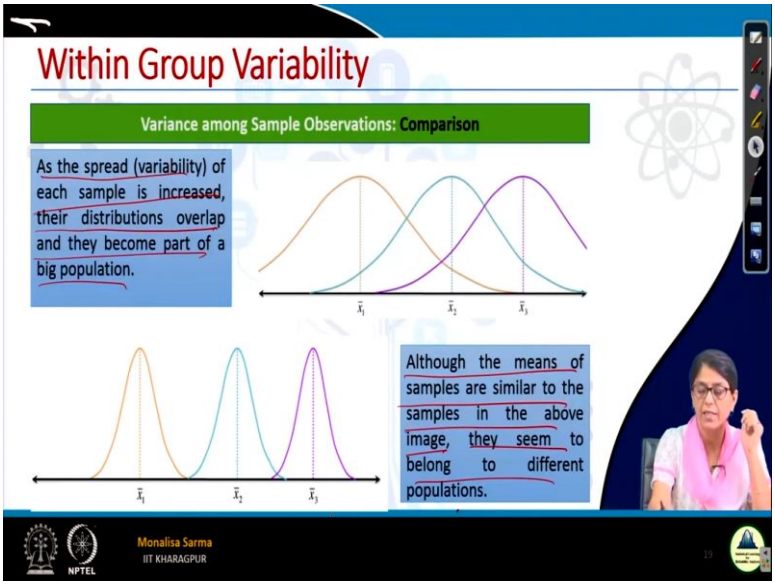
We are getting that lies within and around 18 say 16 17 18 19 15 very much nearby and for say I20 we are getting say 24 25 23 21 26. It is not that but it was varied data as it derived overlapping as if we are looking for, as if we are considering the mileage of a single make of the car it was we got that feeling for the data. So, this is the case when the data are very much varied. Similarly here when the data says see here we got the same mean, but the datas are not varied.

(Refer Slide Time: 10:19)



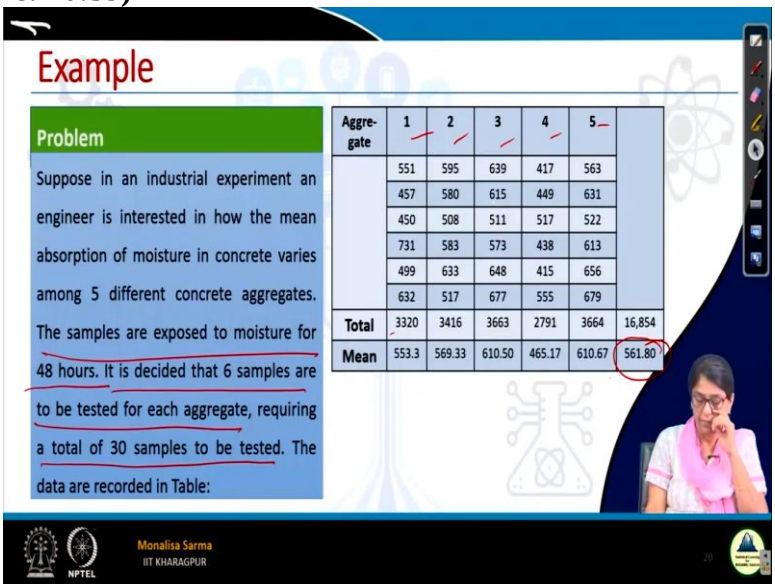
So, this particular this belongs this seems to belongs to different populations, but the previous one it does not seem to belong to a different population. So, this is within group variability.

(Refer Slide Time: 10:31)



So, although the spread as the spread of each sample is increased their distribution overlap and they become part of a big population, although the means of the samples are similar to the samples in the HAVOC means, they seem to belongs to different population here as if they belong to different population different A, B and C here as it the belongs to simple single populations. So, this is within group variability.

(Refer Slide Time: 10:59)



Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.3	569.33	610.50	465.17	610.67	561.80

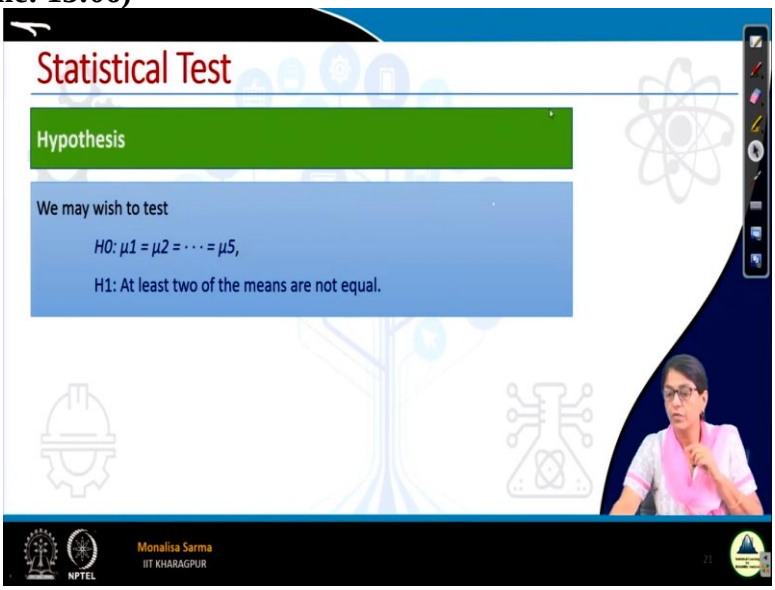
Now, let us see an example, suppose in an industrial experiment, an engineer is interested in how the mean absorption of moisture in concrete barriers among 5 different concrete aggregates, so, what it has is it has 5 different concrete aggregates basically 5 different concrete aggregates means there are 5 different concretes in 5 different aggregates it has used different maybe different chemicals, 5 different chemicals to the aggregates, concrete aggregates.

So, use different chemical store on their total same, we have taken some concrete aggregates, and they we have used some chemicals, we use 5 different chemicals, that is why we have got 5 different concrete aggregates. And what we are interested in finding out the samples are exposed to moisture for 48 hours it is decided that 6 samples are to be tested for each aggregate requiring a total of 30 samples to be tested the data are recorded in the table.

So, we are interested in finding out what is the moisture absorption rate of each aggregate. So, more the moisture observation more washes that we do not want that kind of concrete. So, absorbed more and more moisture if it absorb more moisture than what happened with the dampness will be there in the walls. So, we do not want that type of concrete way which absorbs more and more moisture.

So, that is why we have done treatments in 5 different treatments 5 basically different chemicals, we have used we are interested in finding out is the by using this chemicals which one will give lesser absorption of moisture. So, we have tested for that accordingly we have got this data, these are the data for different 5 different types of concrete is concrete 1 2 3 4 5, these are the moisture absorption of moisture in 48 hours. We found the mean and then we have found the total and then we found out the mean now this is the grand mean as I was talking about.

(Refer Slide Time: 13:06)



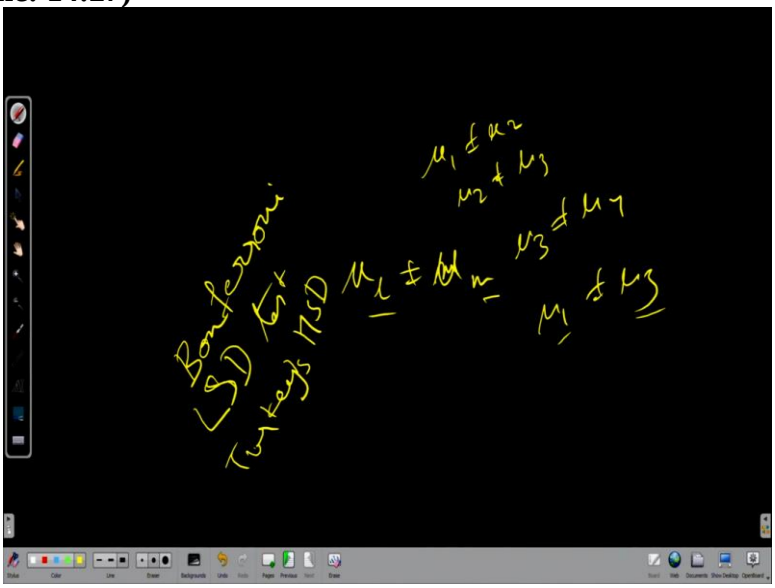
So, now we are interested in finding out in this question we are interested in finding out whether the absorption of moisture in all these 5 concrete types is different or not, whether any because of this different type of mixture we are using because of that will absorption

moisture will be different or it is the same. So, when we are interested in finding out this sort of thing definitely we will do hypothesis test what we have already seen when we are doing for single population or double populations.

Now here what will be the hypothesis test? The hypothesis test the null hypothesis is that all the population has equal mean that is our null hypothesis all the population will have equal mean now, then what will be the alternate hypothesis? Now, alternate hypothesis is at least 2 or more means or at least 2 or more means are not equal that means, here or null hypothesis we are telling that all the means are equal and alternate hypothesis at least 2 of the means are not equal there may be 3 naught equals, 4 naught equals, 5 naught equal.

But at least 2 of the means are not equal this will tell that there is some difference in the treatment that we have done.

(Refer Slide Time: 14:27)



Now, the question is, if we asked me here the question is like if we asked me what we are trying to find out at least 2 of the means are not equal. That means I can write it as that is μ l not equals to μ m, where l and m maybe any of the variables any of the samples that we have considered, l n maybe may come from any of the samples that belong to any of the 2 sample means, like say μ 1 ≠ μ 2 or μ 2 ≠ μ 3, or maybe μ 3 ≠ μ 4 or maybe μ 1 not equals to μ φ like that any l n m can take any value from 1 to 5.

So, now, the thing is that in this test what we have studied we will be able to only find out if at least 2 of the means are not equal, but then we will not be able to find out which mean is

not equal or which 1 or which 2 or which 3 means are not equal, that is basically for doing that, there is a different type of experiment for that, which is beyond the scope of this lecture, I will not be discussing that.

Just so, you can just note it down the different what to say techniques for doing this is one technique is called Bonferroni approach than another it is let me I will just write it Bonferroni approach if I am writing it f e double r r, Bonferroni approach another is a least significant this at least significant difference test least significant difference test least significant LSD least's significance different tests.

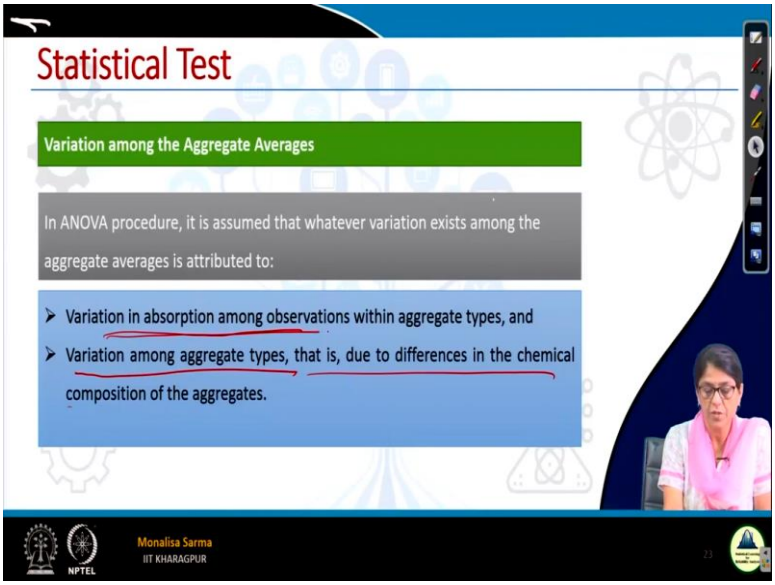
Another is tukeys t u k e y tukeys HSD, HSD represent honest significant difference. So, these are the 3 diff there are more there these are different type of tests by weeks, we can find out among the tests like among the different means, which mean is actually different or which 2 are which of the means maybe one main difference of 2 or 3 which are the means are actually different in analysis of variance we will not be able to tell which are the means are different.

We will be just able to tell where all the means are equal or there is any difference among them means, we will analyse the variance over just limited to that whether all the means are equal or any of the means are not equal. If you are some specifically find wants to if we found out suppose in this case, our alternate hypothesis is true that is at least 2 other means are not equal.

Then we will have to further do some further exponent which I have just mentioned, I have written it down also you can just note it down. So, we will further some tests to find out exactly which of the means are not equal anywhere. In most of the experiment that is necessary then we will do have to do that and most of experiment when we really do not need to do we are just it is sufficient enough that all the populations are same or all the populations are different.

Analyse ANOVA restrict to that just finding out whether all the populations are equal or all the population are not equal. So, that is why this is our H 0 is a null hypothesis alternate hypothesis, at least 2 of the means are not equal.

(Refer Slide Time: 17:44)



So, in the variation among the aggregate averages, we have seen in this example, among the aggregate averages, there is quite a variance that is 553 569 610 465 there are quite variance so, if the variance among the aggregate right is variance exist and another procedure it is assumed that whatever variance exists among the aggregates where we have seen variance among the aggregates it may be attribute due to first one is variation in absorption among observation within aggregate types.

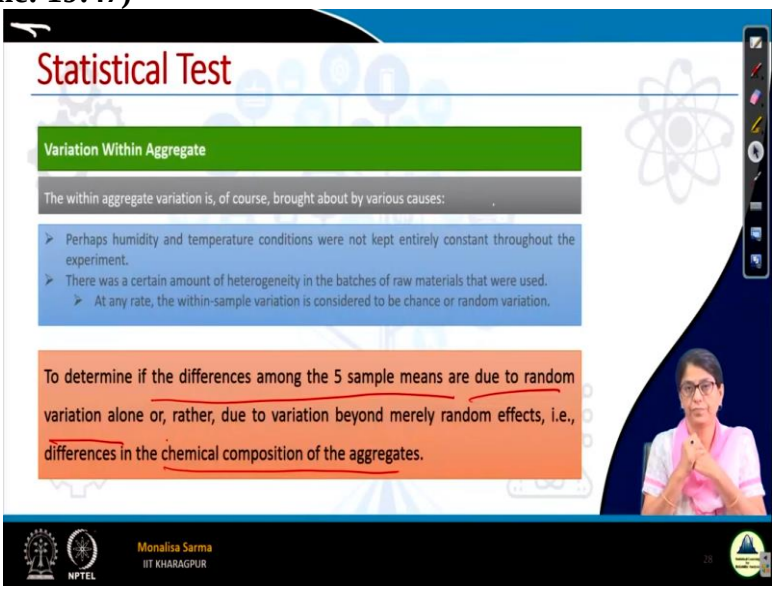
In one aggregate type only chemical suppose we have used chemical A, B, C, D, E in one set of samples we have use chemical A only but there also in one sample only where we have used chemical A only there supposed to we have 10 products using chemical A we have use a sample size of 30, 30 or 5 whatever we have used here suppose here we have used 30 samples when 30 samples we have used chemical A.

Among these 30 also then all moisture absorption may be different some countries may have absorbed more moisture, some countries may have absorbed less moisture, but why variation among aggregate type that is due to the difference in one. This difference in variation may be due to the variation among observation and all variation among aggregates type that is due to the difference in a chemical composition.

This variance may be due to 2 different factors 2 different reasons one is variance because of the different chemical composition, in one type of population we use A another B, another C, another D and another E this variance may be due to the different chemical composition you have used that means 1 variance that is we call it between group variability and another

variance may be within the same group. Variation in absorption among observation may be different, why?

(Refer Slide Time: 19:47)



There are reason for that I have written I think I have the causes here. The within aggregate variation is of course brought about by various causes perhaps humidity and temperature conditions were not kept entirely constant for the experiment. Maybe because of that some concrete has absorbed the same chemical, chemical A we have used for all the 30 but some concrete have used less moisture, some concrete have use more moisture.

Maybe because the condition was not constant because we have conducted experiment for 48 hours condition maybe not constant throughout or maybe some concrete I have kept in one room. So, another concrete I have kept in another room and maybe the temperature is not, environment is not same in both the room and there may be certain amount of heterogeneity in the batches of raw materials.

The raw materials that we; observe the sample that we have taken up 30 samples for each different 5 categories 30 x 5. So, this 30 samples what we have taken so, 30 samples well we have used maybe the cement we have bought similar type of cement for different company. So, the different companies so there will be some slight variance.

So, there will be certain amount of heterogeneity in the batches of raw materials for that maybe chemical absorption may be more or less by using the same chemical treatment at any rate that within rate sample variation is considered to be chance or random variation. So,

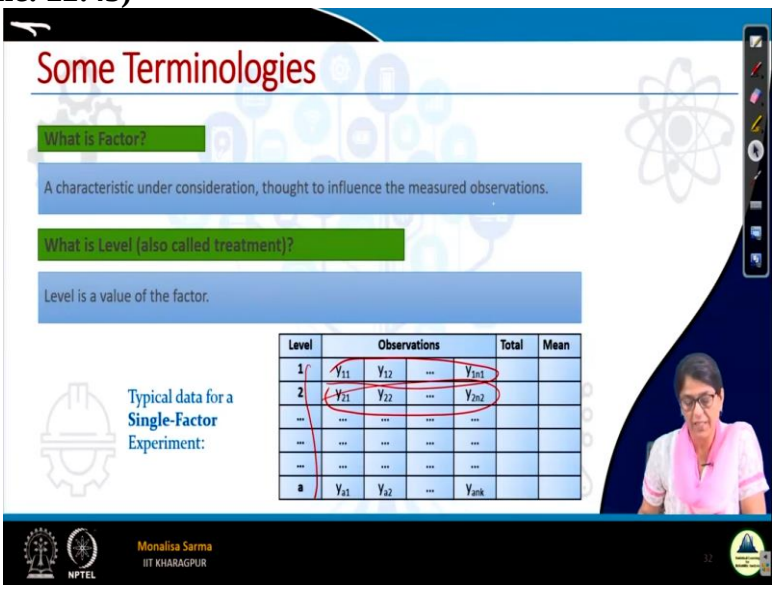
within sample variation we call it a chance variation or random variation. And variation within the variation between groups we call it what it is? It is a variation because of the treatment different treatment.

So, to determine now, what we have to determine if the difference among the 5 sample means are due to random variation alone or rather due to variation beyond merely random effects that is difference in the chemical composition of the aggregates. Consider the music example, the difference in the test scores is only because that we have picked some intelligence students in group A.

And in group B and C we have might be we have picked some dull student, is it the differences because of that or the difference is because of the different type of music we have played in the different classes. Similarly here, so is this difference in the mean that we have seen is due to the random variation alone, random variation what we have discussed, random variation maybe the moisture content with the same type of chemical may be different.

Because of the heterogeneity of the raw material, because of the environment where we have put. So, is it a variation what we have seen is the variation is because of the random variation or because of the different treatment that we have given. So, we need to determine that. So that is what we will do in ANOVA. ANOVA does essentially that.

(Refer Slide Time: 22:43)



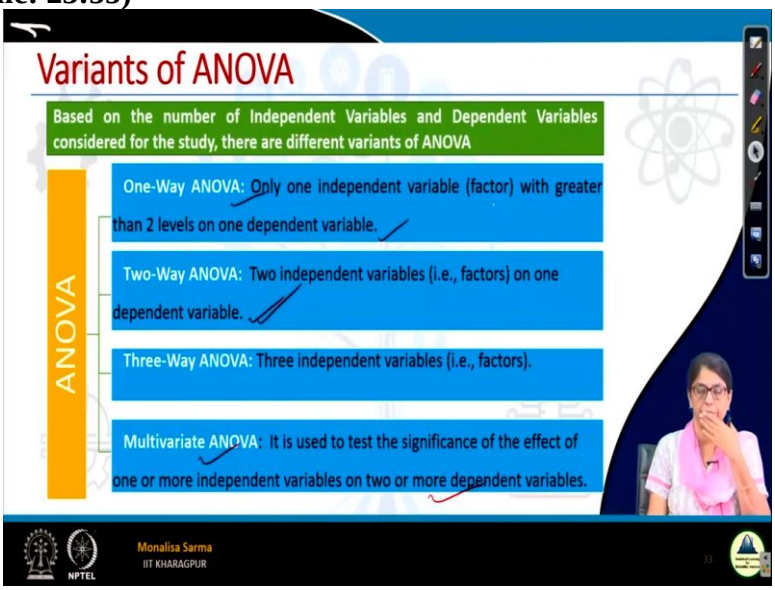
So, when we discuss and what the first thing we need to know, what is a factor, as I have already told you that this music, no music, constant music that is what that is factor we can

call it also treatment or else we are changing one factor that is music, either we are playing constant music, random music or no music. And this just recent example, we are using there is one factor, factor is what can we use it type of chemical.

Chemical A, B, C, D, E. So, in the first case music example, there is one factor how many levels are there, there are 3 levels 1 2 3 levels no music, constant music, random music 3 levels here, there is one factor that is we have used chemical, chemical is one factor, how many levels are there? There are 5 different levels 5 different chemicals we have used level is this basically the value of the factor.

So, if we see a typical data set for a single factor experiment, so this is what we have discussed is always a single factor. So, single factor, we can have different levels for each level will have different set of data, so this is a typical data set we can say.

(Refer Slide Time: 23:55)



Now, the ANOVA we will go to details of ANOVA first before that, there are basically different variants of ANOVA based on the different dependent variables and independent variables. Now in our example on music, which is the dependent variable and which is the independent variable, our dependent variable is the test score. The test score is dependent on something the dependent variable is a test score for independent variable what? Independent variable is the music.

Music we are changing, we are giving this, we are trying to we are playing on that, that is independent, that is not depending on something we are not changing depending on variances

independently changing that. So that is the independent variable and dependent variable is our score. So, in that example, what we have? We have the music exam example we have one independent variable that is the music and one dependent variable that is the test score.

So, this is called one way ANOVA. Only one independent variable or factor with greater than 2 levels, we have 3 levels there and one dependent variable that is called one way ANOVA, two way ANOVA 2 independent variables on one dependent variables 2 where 2 independent variables. Suppose the same example music example I have to, as I have already mentioned before.

Suppose I used A is also as a factor, I have tried this music experiment on the older students as well as younger students, here also A is independent, I am just changing the A's, I am bringing in whole class 10 students or class 2 students I am just saying it is not dependent on anything. So, likewise, music age is also has a independent factor. So, based on these 2 music, and A s how it is affecting the test score, test score is dependent on these 2 music as well as age.

So, this is the one dependent variable, 2 independent variables. So, this is called two way ANOVA similarly, we have three way ANOVA, 3 independent variables and one dependent variable. So, similarly, we have multivariate ANOVA it is used to test the significance of the effect of one or more independent variables on 2 or more dependent variables and it is called multivariate variables 2 or more dependent variables.

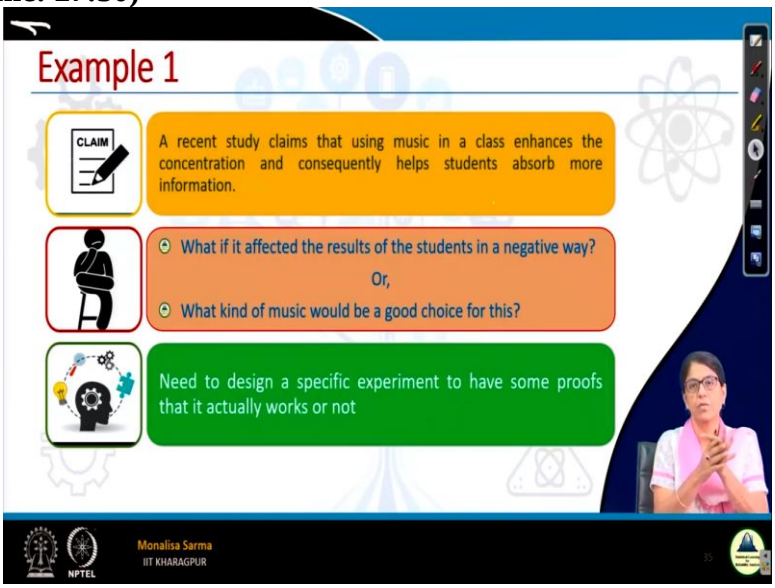
Let us take the recent example what we have taken that concrete mixer. Now here we are interested in, we are mixing some chemical and we are interested in finding out the absorption of moisture suppose, we are also interested in finding out the setting time, setting time of this concrete we are mixing some chemical, chemical A B C D and E how this chemical is affecting the moisture absorption as well as how this chemical is affecting the setting time.

So, here independent variable is 1 that is the chemical treatment that we are using, but we have 5 different chemical treatment level and dependent variable is setting time as well as the moisture absorption. So, similarly there can be more than one independent variable and more than one dependent variable. So, that is called multivariate ANOVA. In this lecture, basically

in this course, we will be discussing only one way ANOVA, two way ANOVA and three way ANOVA it is very complicated to discuss in this class.

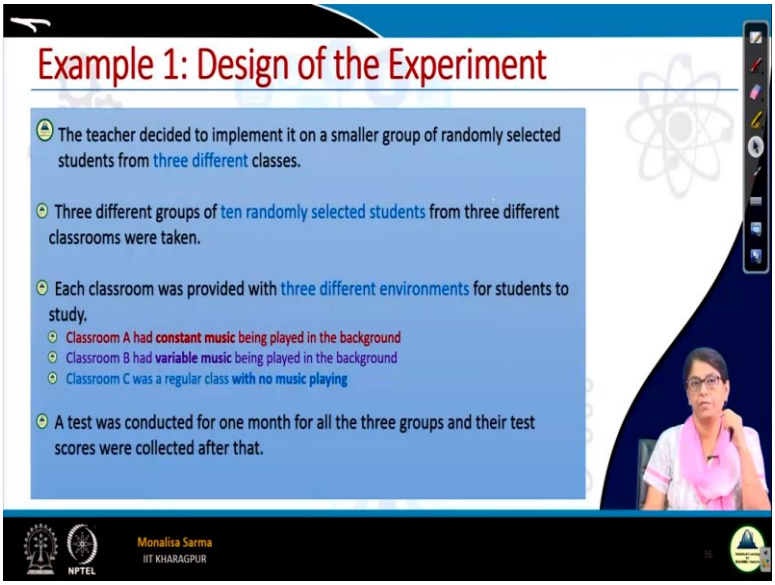
Once we know the integrity of one way ANOVA it will be very easy for you to just Google it and learn by yourself actually two way ANOVA and our 3 way ANOVA it is very difficult to do it manually also you will let it really need to do using computer. So, that is discuss in this class two way ANOVA, three way ANOVA and multivariate ANOVA which I assure you once you know one way you will be able to learn it by yourself.

(Refer Slide Time: 27:30)



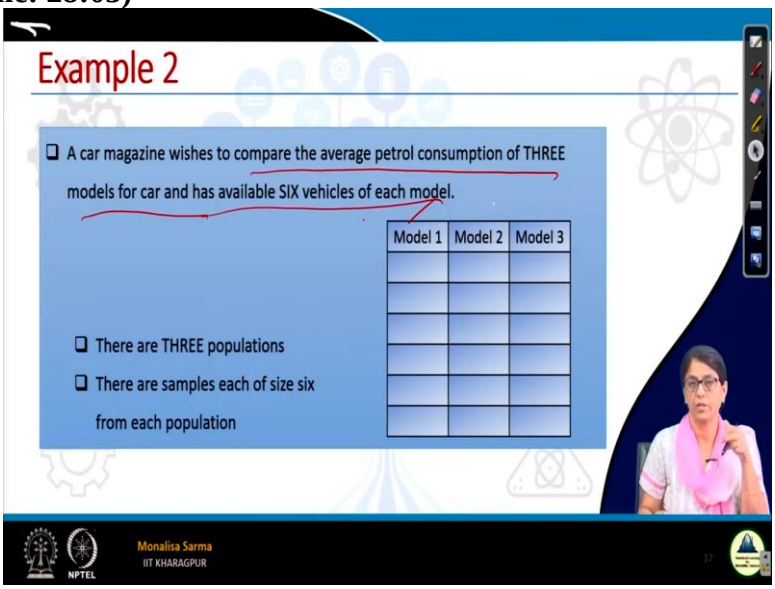
So, now, we will just quickly revisit some example, in the context which have already just mentioned in the context of one way, two way and three way, the recent study about the music what I told. So, I have already explained here like what is if I call it is a one way ANOVA does what we have discussed now that is a one way ANOVA that is music as an independent variable with 3 different levels.

(Refer Slide Time: 27:51)



And if I consider A is also then that is again one way ANOVA but sorry that is two way ANOVA where I have 2 independent variable and one dependent variable.

(Refer Slide Time: 28:03)



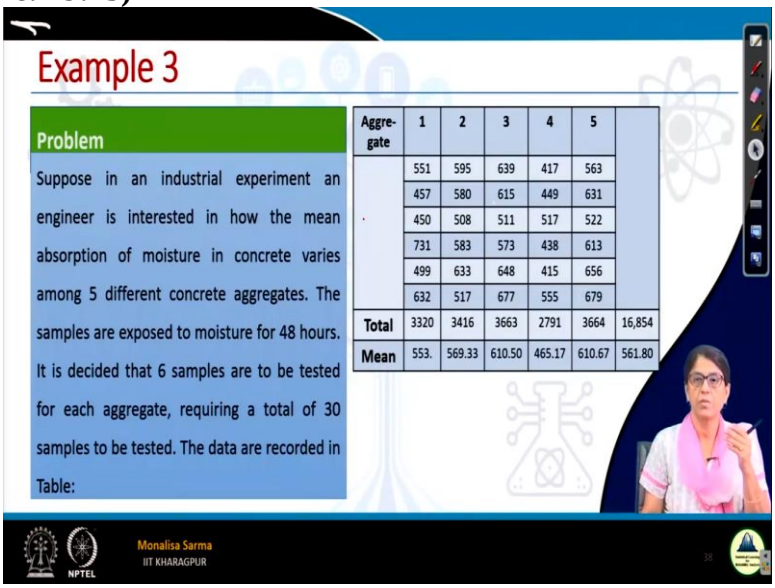
Similarly this if a car magazine wishes to compare the average petrol consumption of 3 models of the car and has available 6 vehicle of each model. So, we are interested in comparing the petrol consumption of 3 different models model 1 model 2 model 3 petrol; consumption we are interested in finding out the petrol consumption. So, what is the dependent variable here? Our dependent variable is the petrol consumption.

How much is the petrol consumption that is our dependent variable, it is depending on what? It is depending on the model, model 1 model 2 model 3 3 different models. So, here it is again one way ANOVA our independent variable is the model dependent variable is the petrol consumption suppose, again here again I have used 3 different drivers one is a very

young and another is sitting driver one is competitively old driver, one is a very old person basically and another is a very new driver for drive.

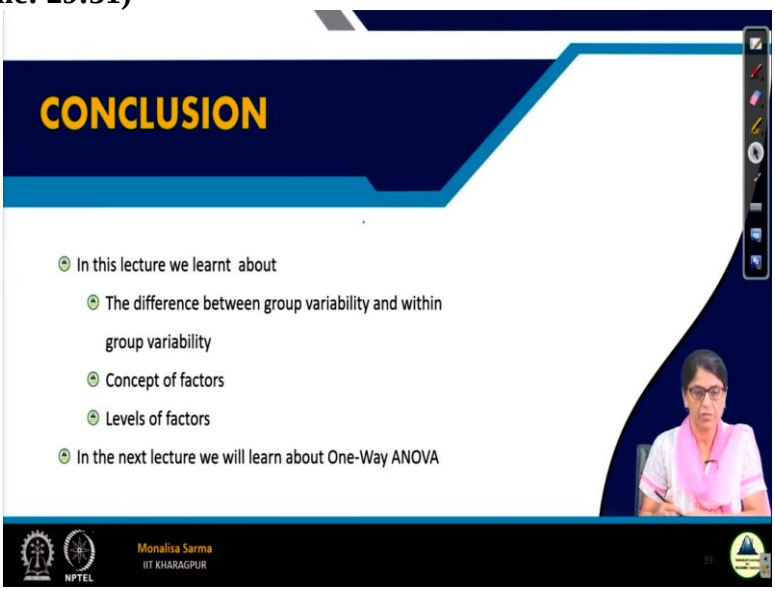
So, there again I have introduced 2 independent variables that drivers have 3 different levels, so models and driver becomes 2 independent variable and the petrol consumption is one dependent variables. So, this is again a two way ANOVA.

(Refer Slide Time: 29:15)



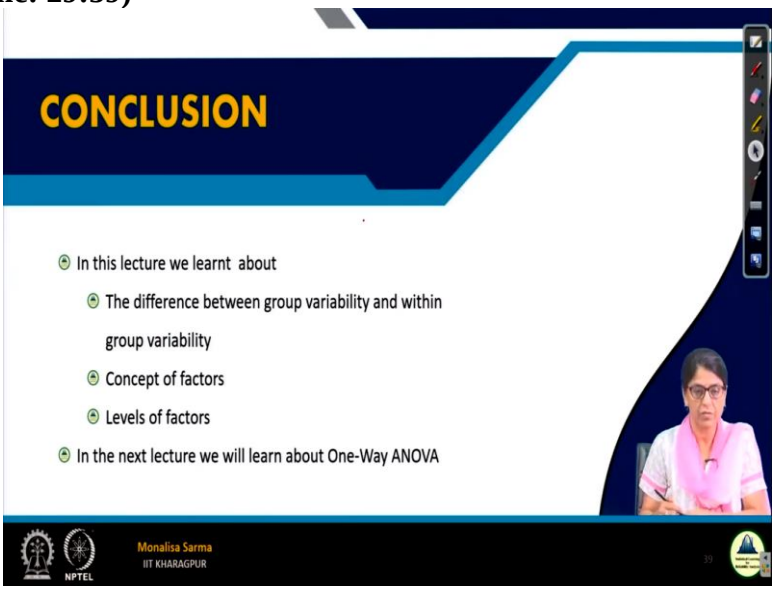
Similarly this example which I have already discussed in the context of one way ANOVA as well as the multivariate ANOVA, if I consider setting time as well as the moisture absorption it becomes one independent variable, 2 dependent variable So, we call it a multivariate ANOVA.

(Refer Slide Time: 29:31)



So, in this lecture we have learned about difference between group variability and within group variability we will also learn so, what is a factor, what are the independent variable, dependent variables and levels of the factors also called treatment mind it. In some book may get it as factor in some Google get it as treatments. So, it is a concept of treatment or factor and the different levels of this treatment are different levels of factor.

(Refer Slide Time: 29:59)



In the next lecture, we will learn about one way ANOVA. So, there is the reference and thank you guys