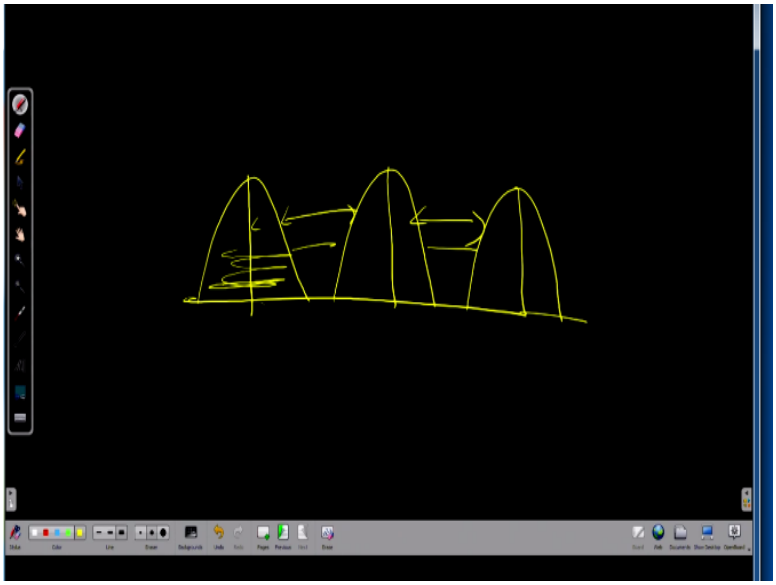
So, remember what we have seen the analysis of variance is based on let me write it is basically based on 2 factors we have seen what are those 2 factors? Let me write it here one is the comparison of the variance among the means of the population first let me write it and then the second one is like variance among sample observation within individual population we have seen this analysis of variance is based on what one is the comparison of the variance among the means of the populations, this is what we call it this thing?

We call it between group variability or between treatment variability, anything this we call is between group variability and this second one that is variance among the sample observation within individual population, this is what this is called within group variability.

(Refer Slide Time: 03:06)



So, if we basically draw the diagram, if I draw the diagram say this is one sample this is another sample, I am giving the distribution of the sample basically this is one sample. So, this is the mean this is the mean this is the mean. So, first one is within when I tell is between group variability as between group variability means, this variability between this group variability between these 2 groups, variability between these 2 groups this is called between group variability and another is within group variability.

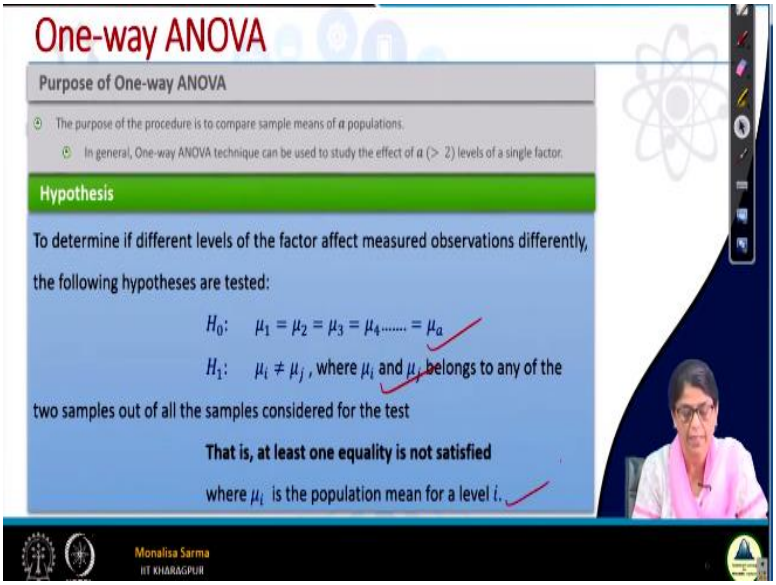
Within group variability this variability within the group that mean then what is the standard deviation basically, standard deviation is what is the how much it deviates from the mean each and every data how much it deviates from the mean. So, this is called within group variability and these things are called between group variability. So, we had analysis of

variance is basically based on this comparison of these 2 variance, we have discussed till here and we will gradually proceed from there like. So, based on this also we have seen.

But in ANOVA what is the concept of dependent variable, what is the concept of independent variable we have also seen that. So and now, in today's class, we will be discussing one way ANOVA one way ANOVA means as I told you if you can remember, so, there is one independent variable and one dependent variable which depends on one independent variable. Like again if I can come to my previous example for the class which the performance of the student based on the music.

So, the performance of the student was the dependent variable independent variable was the music so, that is called one way ANOVA where we have both dependent and independent variables this one, so, we will be discussing today one way ANOVA.

(Refer Slide Time: 04:59)



So, now as we have already discussed and this is just a quick recap the purpose of the one way ANOVA is to compare sample means of a population, a population means when I am telling a population means that means there is only one factor since we have one way ANOVA there will be only one factor. For one factor there are how many levels if there are a levels that means total will have a populations like for the music, there are one factor that is music is factor, how many levels whether there are 3 levels like.

And even I have given an example of medicines, medicine a treatment, different types of treatment for a particular disease. So, treatment a, b, c so that was also one factor and there

were 3 levels. So that was a population's means 3 populations. And again, we have given a single example of concrete mixer, the moisture observation of a concrete mixer, if we add different types of chemicals; I have used 5 different types of chemicals. So, there is 5 different types of chemicals that means total there are 5 levels.

So that means here 5 population a is 5 a is the number of level basically. In general one way ANOVA technique can be used to study the effect of a levels of a single factor where a is greater than 2, but at the same time for even if a = 2 still we can use ANOVA I will not give you the answer you can thing yourself for two populations, can we use ANOVA? If yes, why do not we use ANOVA why you use T test, if you want to compare two populations.

Very simple answer you also yourself will be able to get the answer for that. So, now, coming to that, how we go about the test, how to go about the whole thing? So, our main aim here is why do we do ANOVA, what is the reason behind? First of all we are interested in trying to find out compare the means of a population. So, basically we want to find out because of a particular treatment, is there any difference in the different populations we have given some treatment.

Because of the treatment is there any difference in among the population or there is no difference, that is what we basically want to find out. So, now, so, we will like previously for what statistical inference what we have done similarly, here also will become a carrying out hypothesis test. So, our null hypothesis is that all the μ's are equal implying that the treatment has no effect on the populations.

So that implies the null hypothesis, null hypothesis all the means are equal means, we have done the treatment like we have played the music, but music or no music or constant music, whatever it is the it did not have any effect on the performance of the students, it totally depends on the student's own calibre. So, that is my null hypothesis. So, for the mixing chemical to the concrete, they are also the mixing with the form.

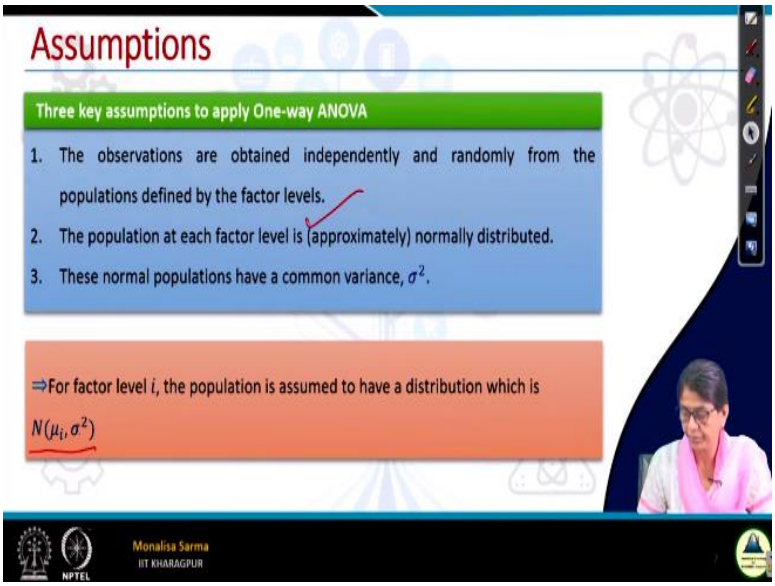
If I mix different chemical also the moisture observation there is no change in the moisture observation that is my null hypothesis. Now my what is my alternate hypothesis, alternate hypothesis is that at least 2 of the means are not equal that means the factor is having some or the other effect because of the treatment, we are getting some effect there is a change in the

mean and at least 2 of them means, there can be change in more than 2 there can be 3 are different 4 are different or 5 populations, all the 5 are different.

But at least 2 means are different that is the alter means alternate hypothesis means that is all the means are not equal, basically, that is my alternate hypothesis. Now, as I told you in my last lecture, so if I am interested in finding out among if among the between the 2 hypotheses in my alternate hypothesis gets accepted the means when my null hypothesis gets rejected and alternative hypothesis gets accepted if my alternate hypothesis gets accepted I can only say that all the means are not equal at least 2 are not equal.

But through these tests through ANOVA I am not being able to say which of the means are not equal for that different there are separate tests which I have already mentioned in my last lecture. So, now carrying with this, so, this is my null hypothesis this is my alternate hypothesis. And the way μ i is the population mean for level i.

(Refer Slide Time: 09:28)



Now, when we do one way ANOVA we have to take certain assumption, like for our different statistical inference also, you see, we have taken some assumption what was the assumption when we have taken T distribution, chi square distribution, F distribution that our population is normal. We have taken the assumption that our parent population is normal with that assumption only we have to work on a statistical inference though T test is quite robust.

But still very much away from normal is not acceptable. χ^2 and F is very much sensitive to normal. So, here also similarly we have certain assumption. So, what are the assumptions? There one of the first of them assumption is that the observations are obtained independently and randomly from the population defined by the factor levels. So, defined by the factor levels the different population from the different factor levels the different population from the levels randomly we take the samples.

There is no dependency among picking samples from here and there is no dependency randomly we pick the samples from the different factor levels populations. That is the first assumption that assumption is valid in all the tests what we have done. The population at each factor level is approximately normally distributed here also we need approximately normally distributed like T, χ^2 and F. It has one more assumption this normal population have a common variance σ^2.

So, we are trying to compare different populations a populations where a is the number of the level now, all these populations are approximately normal further this all this population has a common variance σ^2. So, these are the 3 assumption now, like remember way we have taken a common variance and while we are comparing two populations to for independent case when you are comparing two populations and the population standard deviation is not known.

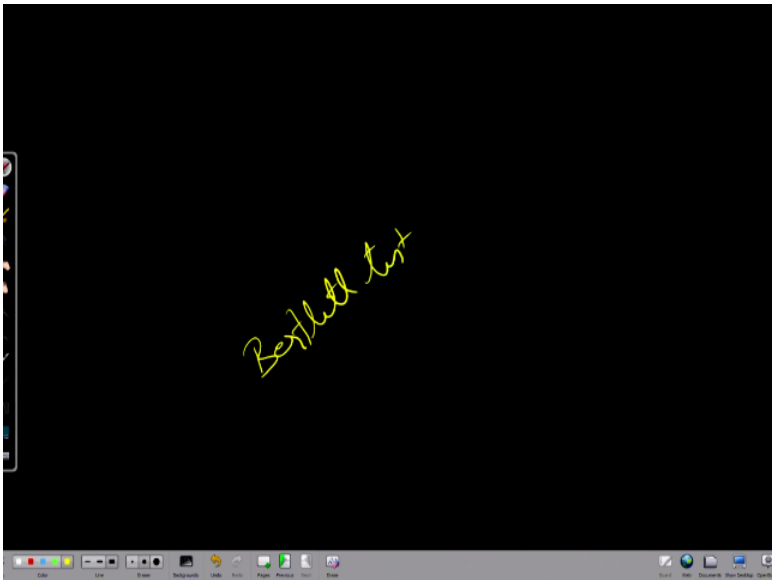
Remember then what we have taken another two populations, we are trying to compare the mean of two population and the population standard deviation is not known than what we have taken we have assumed it to be equal there is one case assumed to be equal, when we have assumed to be equal, then from there we try to find out a pooled variance estimate. Remember if you cannot remember please go to the slides again. So, from there we try to find out the pooled variance estimate what is the pooled variance estimate?

It was nothing but a weighted sum of the variance of both the populations both the samples basically weighted sum of the variance of both the sample that gives us the pooled variance estimate that it what we have already done in the while comparing two population. Similarly, here also we will assume that it has a common variance σ^2 and like in two population case, we have assumed that σ as the both the population has a common variance that is the pooled variance that we have seen.

But again they will test also 2 test whether both the variance that we have assumed If I can we assume the variance of both the populations is equal for that we have seen F test remember similarly, here also we can check because here there is no two population there will be more than two population 3 4 5 6 7 8 9 10 anything. So, we have assumed all the population has a common variance σ^2 and plus at the same time, here also there is a test that this assumption that it has a common variance σ^2 it is true or not.

So, this is beyond the scope of this course, I am not going to the test and in fact if you are interested, you can go to it yourself only it is available in the textbook and you can Google it also.

(Refer Slide Time: 13:28)



I will give the name of the test. So, it is called basically Bartlett test at this stage, you do not require basically so, I have not kept this on. Now, for our convenience and this ANOVA actually we need to be assumed that all of our population has a common variance σ^2, if we do not have a common variance σ squared and can definitely we cannot use just one way ANOVA in this case, so that still if you are not sure that they may have different variance, then we can go and check whether this Bartlett's test.

Actually, the thing is that it is very natural to assume that it will have a common variance why you see how what type of population we are not trying to compare apples and oranges. What we are doing on the same populations? We are basically doing different treatments. Like for the case of students, students score, we are basically same level of students only is

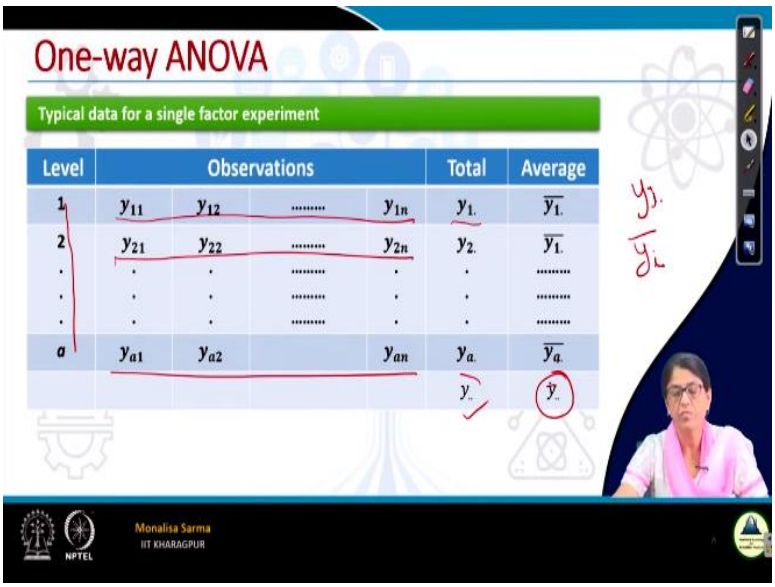
not it? Either, I am talking of one level, same level of students; we are taking it from different class different section.

And what was the variance of their performance? The variance in the same level of class and student level, the variance of the performance will be same level, is not it? You are supposed to have given some treatment because of some treatment. It is not that some people will have very good result and some people will have a very bad result that is very unlikely whatever earlier what was the variance of performance before any treatment after treatment definitely their score will improve.

So, if the score improves at all the score will improve, but the variance will remain same earlier suppose before the suppose we have concluded that giving music helps the concentration level earlier suppose, a student in a class he got out of 50 his score is always around 40. So, after introduction of music maybe he got 45 another students say his performance is 35 maybe after introduction to music he got a 38 the variance basically remains the same before our level and after variance usually remains the same.

That is why it is quite intuitive to assume that the it will has a common variance that is σ^2, but still if you have doubt you can go and do the Bartlett's test. So, for factor level i the population is assumed to be have a distribution which I can mention it in this related in this way because this is what that means, it is a normal population with mean μ i and variance σ^2.

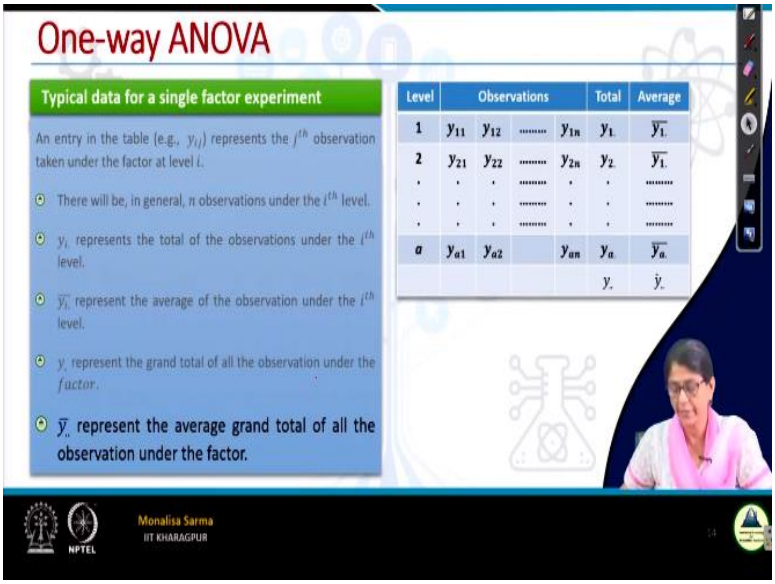
(Refer Slide Time: 16:17)



So, typical data set for a single factor experiment will look something of this sort these are the different levels we have 1 to a levels. So, this is the first data set or for level 1 the different data is for level 1, this is the different data for level 2 similarly, these are the different data for level a now, that is sum total of this I am using an expression that is y i dot see some total is y 1 dot sum total is y 2 dot that means, y i dot that not written properly y i dot then mean of this is y i dot bar.

Similarly, the grand total grand total means total of all the data how will, I get grand total this total will also give me the grand total. So, this is y double dot this one y double dot that is grand total and this is the grand mean y double dot bar is the grand mean, grand mean is the mean of all sorts all set of data this also I can get by the mean of this means, is not it?

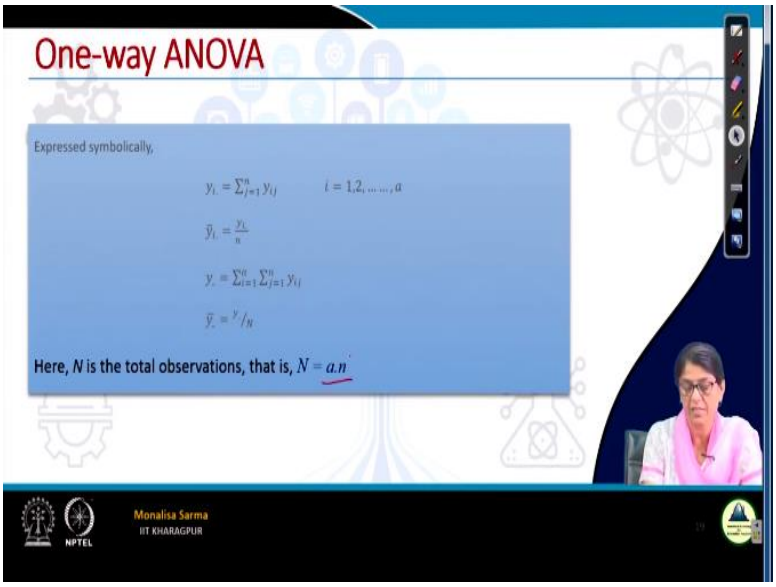
(Refer Slide Time: 17:37)



So, an entry in the table example y ij what does it represent any entry suppose y ij say y 2n. So, here i is 2, j is n. So, represent the jth observation taken under the factor level i so, this is the nth observation taken under factor level 2 so, any entry table y ij that is a representation there will be in general n observation under the ith level here first we are considering that all the sample sizes are equal we will also see when the sample sizes are not equal.

So, here we are considering that we have taken from each population we have taken an equal sample size that is n, y i dot represents a total of the observation under the ith level as I already told you, then the y i bar is the mean y double dot is the grand total and y double dot bar represents the grand mean average total means average grand mean yesterday we have come across this what grand mean remember.

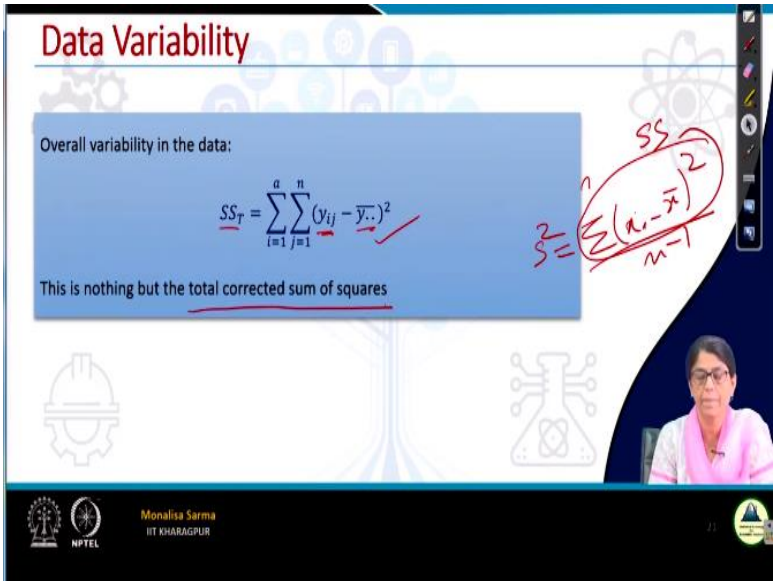
(Refer Slide Time: 18:51)



So, this is expressed in this; what is y i dot, y i dot bar and all there is nothing you know like at this y i dot is the sum of all data. So, that is sum of all data basically for which will for ith level we are summing up all the jth data how many data's are the n data. So, for the ith level all the y 1 + y 2 + y 3 upto y n for the ith level that gives me my y i dot and y i dot bar is divided by n. Similarly y double dot is the summation of all data then grand mean is the y double dot by N, N is the total number of all data.

So, N is the total observation where N = a n because sample size is n total a level so, that will be N = a dot n. Now, we will see given this table we have seen this data this data this is the whole data is not it? From this whole data, now we are interested in finding out the overall data variability. We know what is variability and statistics when you are talking about statistics, variability is a term which will come across each and every step. So, you should know clearly what is that. So, now I am interested in finding out the overall data variability.

(Refer Slide Time: 20:21)

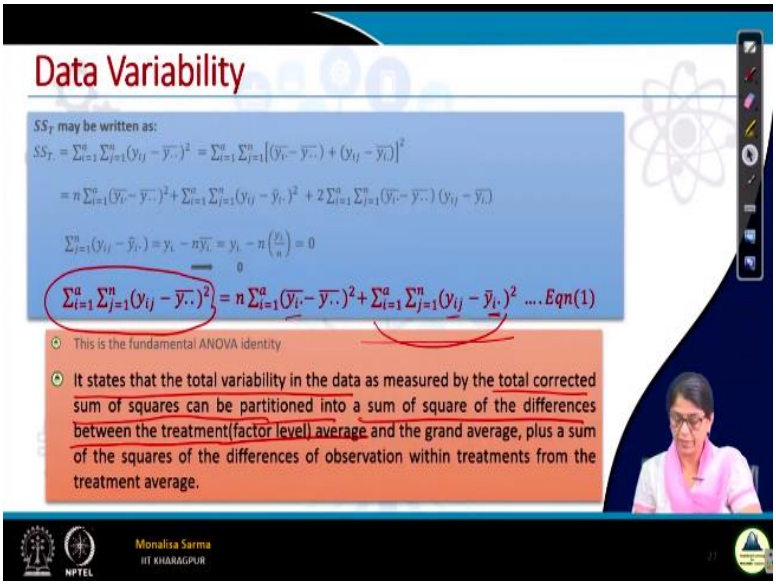


So, overall data variability means total and let me take a small example like the music example, there are total 10 students who have taken from each class so total there are 30 students. So, first I found out the mean of all these 30 students, so that I call it grand mean. So, overall variability means how might each and every data varies from the grand mean that is made overall variability of the data, is not it?

So here, so this let me term it as an SS T, so this is each data I am subtracting it from the grand mean, this is nothing but the basically the sum of squares. Remember, what is sum of squares we have seen this is if I am interested in finding out S^2 for a sample, what is S^2? S^2 is $(x_i - \bar{x})^2$ this is i = 1 to n, then this is divided by n - 1. So, this factor we call it sum of squares is not it? Sum of squares SS so, this is nothing but the sum of squares only for all the levels.

So, that means, y i means we are considering all the data's all data we are subtracting from the grand mean. So, we are finding out this is the total variability of the data not from one sample from all this sample from across all the sample the whole set of data, for the music example, there are total 30 data's so, we are trying to variability from across all this 30 data by subtract each and every data from the grand mean, mean of all this data. So, that is we are calling it SS T this is nothing but a total corrected sum of squares, is not it? This is nothing but the total corrected sum of squares.

(Refer Slide Time: 22:17)



So, what we got is, again the same so whatever this expression, this expression is SS T. I am writing it again, this is my expression, you see, now this expression, I will do a bit of twisting I want something else. So, I will twist the expression a bit, but my expression should remain same, I should not change the expression in any way. But somehow I will do some sort of twisting. So, what I have done here, I have introduced one new term that is y i dot bar.

So, if I introduce one new term, then if I add one term, I will have to subtract the term as well, is not it? So that is what only I have done, I have added this and I have subtracted this also. So, by doing that, I have simplified it in this form earliest y ij - y double dot bar that is the grand mean. Now, I have introduced y i dot double bar what is y i dot bar remember, that is the mean of for one treatment one level mean of one level have introduced that new data mean of one level.

So, this I am adding it and subtracting so this is nothing but now it is in the form of $(a + b)^2$. So, $(a + b)^2$ is $a^2 + 2ab + b^2$. So that is what I got it. So, first, this is the term this term a^2 this is a, this is b, this is a^2. So, first I got this is my a^2. Say why I have written n for in this term, you will see there is no j there is only variable is only i so when I am writing j = 1 to n.

So, since there is no variable for j that means what it will be summation of those terms j times is not it? The n j = 1 to n time so that that is why I have written n into this from this I got n into this from this again here, y i and j both the terms are used. So, I am keeping both the summation. So, I got this term that is b^2 now 2ab so this is 2ab this term. Now see, now in

this 2ab you say 2ab this term. So, I have taken a portion of this term portion of this 2ab which I have given here by red colour, you see I bought this term here, what do I get?

If so what is ∑ j = 1 y ij ∑ j = 1 y ij gives but they basically total of one level is not it? Total of ith level so it is y i dot and similarly, ∑ j = 1 y i dot bar somebody here no j term is there. So, basically it will be n x y i dot bar again what is y i dot bar? y i dot bar = y i dot divided by n. So, n n get cancelled, so, this term becomes 0 that means my this whole term becomes 0.

So, what is remaining is only $a^2 + b^2$ term. So, this is the equation to finally, I got this equation this is what my total variability of the data my overall variability of the data which is also called total character variance. So, this is equals to this is my a^2 term this is my b^2 term I got 2 components here and this equation let me call it as equation 1. Now, this equation 1 this is the fundamental ANOVA identity.

Now, we have seen what is ANOVA? ANOVA is something to be conscious of 2 variance one is variance between the group another is variance within the group another is based on that we have seen that but exactly in what way it is linked nothing we did not see now, we are coming to the ANOVA identity. So, this is the fundamental ANOVA identity. What does this state this expression if you look it by yourself, you do not have to look into the size just look at the expression a will be able to say.

What does it mean see what this is the total variability this portion? This is also red, my pen is also red, it is becoming difficult to this thing. So, now this portion is the total variability in the data. This is the total variability data this total variability of data is measured by these 2 what is this? This measured by the total corrected sum of squares can be partitioned into a sum of squares of the difference between the treatment average.

Difference sum of squares the difference between the treatment average seen in the difference between the treatment average what is the treatment average is? y i dot bar is the treatment average is not it? And y double dot bar is the grand mean. So, what is that is the sum of squares of the treatment average is not it? Basically it is the between group variability we are trying to find out the variability of treatment average from the grand mean exactly not variability.

Because when we find out variable it has to be divided by n - 1 term is the n - 1 term is missing just the sum of squares, this is the sum of squares of the difference between the treatment average and the grand average. There is a sum of squares between the treatment average and the grand average this is one term and what is the second term? This is the second term, what is the second term is a sum of squares of the differences of observation within treatment from the treatment average.

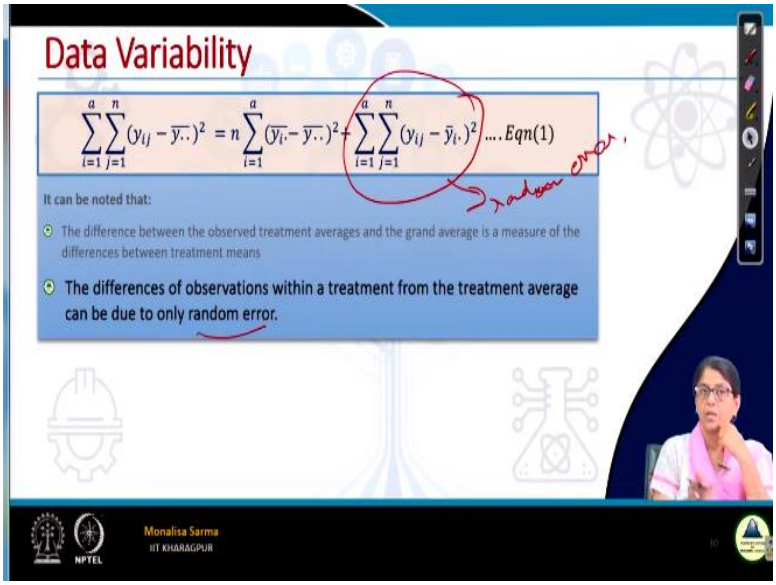
Say this is the treatment average from the treatment average what we are substituting each observation of the treatment. So, there are 3 treatment levels we will get 3 treatment means let me call it y 1 bar, y 2 bar, y 3 bar. I will get 3 treatment mean y 1 bar, y 2 bar, y 3 bar. So, now from the first treatment each value from each value I am substituting from y 1 bar what I am getting sum of all the plus again sum of each value from the sample 2 - y 2 bar.

That is a variance within that group, plus again variance within the third group within so that is my; this equation, this expression. So, understood what we got in this fundamental ANOVA identity. This total variability in the data which is measured as the total corrected sum of squares this I am working on the total variability in the data, is not it? I am let us not talk in terms of variability. Let us talk in terms of sum of squares.

So, what is this the sum of squares total collected sum of squares, because I am trying to find out the difference of each and every data from the mean, that is in my left hand side, in my left hand side, I am trying to find out the difference of each and every data from the mean. So that is my total corrected sum of squares. And this total corrected sum of squares I am partitioning it into 2 parts, what are the 2 parts what I did not partition this equation, I got the equations what is this 2 parts?

One part is that it is the sum of squares of difference between the treatment average and the grand average that is one part. That means here basically, I am finding out the difference between the between different groups and other is the sum of squares between each data within a group from the group mean.

(Refer Slide Time: 30:41)



So, this expression here again, I have written it the same expression. So, now, this expression, this difference between the observed treatment average and the grand average is a measure of the difference between treatment means, this portion, it is called a measure of the difference between treatment means, each mean and if each treatment means I am subtracting from the grand mean. Then, this is the difference of observation within a treatment from the treatment average this portion.

Within a treatment from the treatment average and this portion, why we get the variance within a sample why do we get a variance we have seen while doing the example for this concrete mixer, when mixing the concrete mixer, we are mixing with different chemicals to see the absorption ratio absorption amount basically. So, we have seen why there may be difference within the samples difference within the sample maybe the while doing the test, the environment was not the same for throughout the time that may be one of the reason.

Or there may be certain the heterogeneity in the material itself or there may be a while the testing there may be testing equipments, we have used different testing equipment to test the whole lot of samples that different reasons. So, this is basically all these reasons we call it due to random error. So, this within the treatment from the treatment this whole thing, we can call it as random error is not it? Because variance within a treatment, it can be due to that only is not it?

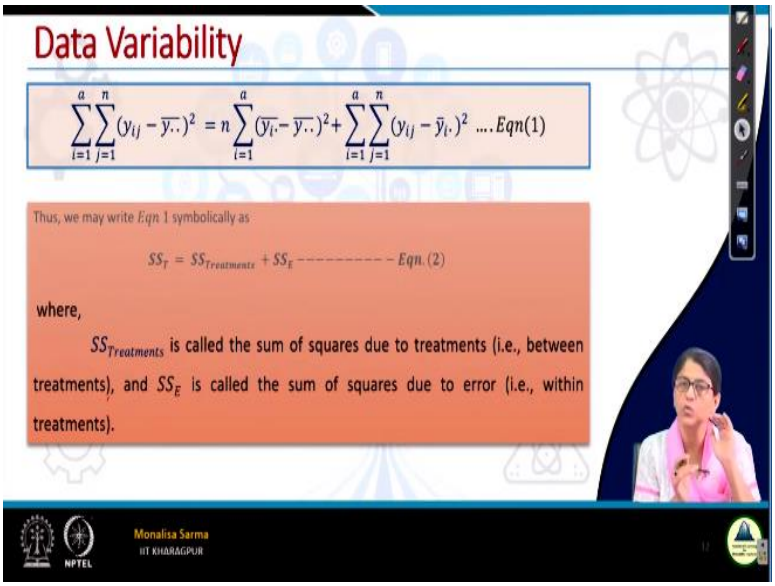
Due to any of the reasons, and when was not same products for experiment heterogeneity of the material, heterogeneity of the measuring material, anything in any of the example if you

take any other example things may be different, like when we are talking of different types of treatment for a particular disease, suppose for treatment a, treatment a group of people who are getting treatment a there also we might get variance among the data the curing time for some people may have cured getting to the 30 days.

Some people are getting cured 25 some 26 some 27 some 29 or maybe some 32, 35. So, they are maybe variance among themselves on the same treatment why many reasons maybe one of the reasons maybe the some people maybe immunity level is quite low some people immunity level is quite high and maybe think some people may not be taking the medicine properly very weak strong many other many other reasons like so, all those we can call it as a random error.

So, total avoidable variability in the data we are classifying it into 2 partitions one is maybe due to the random error and another is due to the treatment difference due to the treatment.

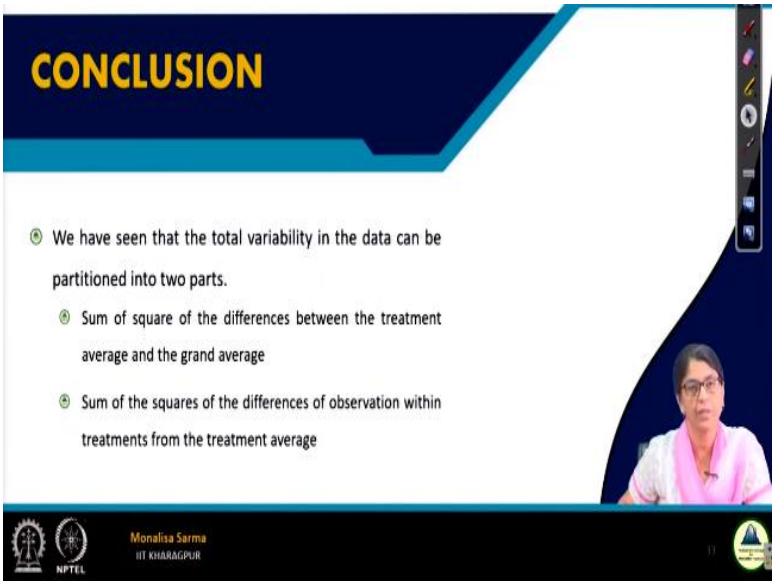
(Refer Slide Time: 33:55)



So, equation 1 we can write it symbolically as in this pattern. So, this is SS E, E in the subscript basically. So, SS T is the overall variable in the data SS treatment SS T meaning total variability in the data and SS treatment means variability among the treatment SS E is variable to due to random error. So, SS treatment is called a sum of squares due to treatment that is between treatment and SS E is called the sum of squares due to error that is within treatments, this is not the end of one way ANOVA.

So, we have till now we have come to this stage, where we found that the total variability in the data is partitioned into 2 parts. One part is the sum of squares due to treatment. Another part is the sum of squares due to errors. Still, there are many other things to discuss for this class. I am stopping my lecture here.

(Refer Slide Time: 35:00)



So, see here if I want to conclude this, what I will say is that we have seen that the total variability in the data can be partitioned into 2 parts. Sum of the squares are the difference between the treatment average and the grand average sum of the squares of the difference of observation within the treatment from the treatment average.

(Refer Slide Time: 35:19)

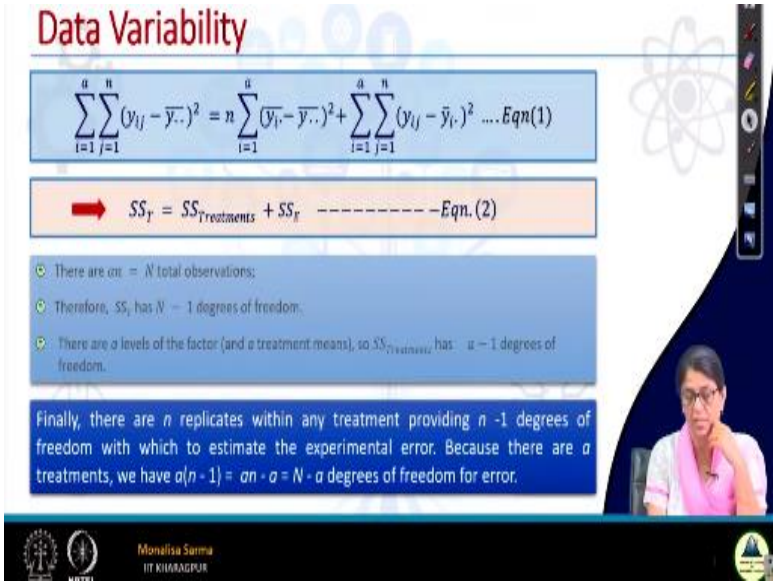
So, these are the references than thank you guys.

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

Lecture – 35
ANOVA - IV

Welcome guys. So, we were in discussion on ANOVA we were discussing one way ANOVA basically.

(Refer Slide Time: 00:37)



And in our last class we have stopped here, we have stopped at this point. So, our SS T that is the overall sum of squares, the total sum of squares we have seen it can be partitioned into 2 parts, these are the 2 parts one is SS treatment and another is SS E. Whereas SS in treatment is the sum of treatments sum of squares basically, that means here we are trying to find out the difference of treatment average from the grand average.

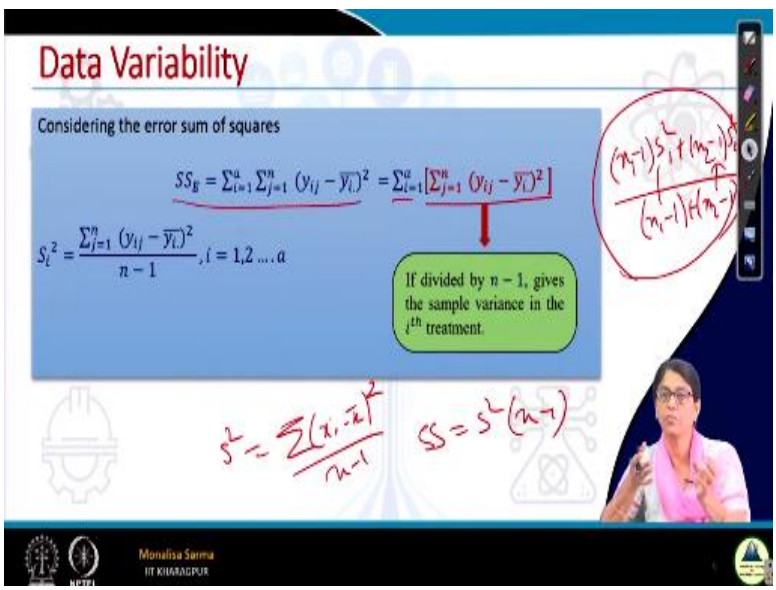
And SS E is the due to random errors where we are trying to find out now difference of all the data from the grand mean sorry, within a sample difference of data from the treatment mean. So, that is basically that is the variance within the treatment. So, that is SS E that is due to the random error. So, now, we will take pick up each and every both that terms together and the first of all, so, total how many, absorptions are there? Total there are N observations why?

Because there are total a level and each level there is small and data, so total is a into n = N. So, now, what is the degrees of freedom for SS T degrees of freedom will be because they are we are considering the overall the total so, degree of freedom will be N - 1. So, now tell me what will be a degree of return for SS treatments. So, by dividing by the degree of freedom basically sum of squares divided by the degree of freedom will get a variance.

So, what is the treatment variance to find out the treatment variance the degree of freedom for SS treatment is basically a - 1. So, now, what is the degree of freedom for SS E that is of course, when we know this term and you know this term definitely we can find out this term that is definitely one way next other way also we can also find out see, there are total n replicates within a treatment and each provides a n - 1 degrees of treatment.

In each treatment, we have degrees of freedom is n - 1 and total there a such populations. So, total degree of freedom will be a into n - 1 that is N - a is the degrees of freedom for SS E.

(Refer Slide Time: 02:44)



So, now, first thing, so, here there are 2 terms SS treatment + SS E So, we will take one, one term. So, first let us take SS E. So, this was my SS E terms is not it? Now, here see in this term, you just see this term which are market with red if this term if I divide by simple n - 1 what do I get this term y ij - y i dot bar that means, within a level, I am trying to find out the difference

between each data from the mean is not it? So, I will get the variance of that treatment variance of the ith treatment.

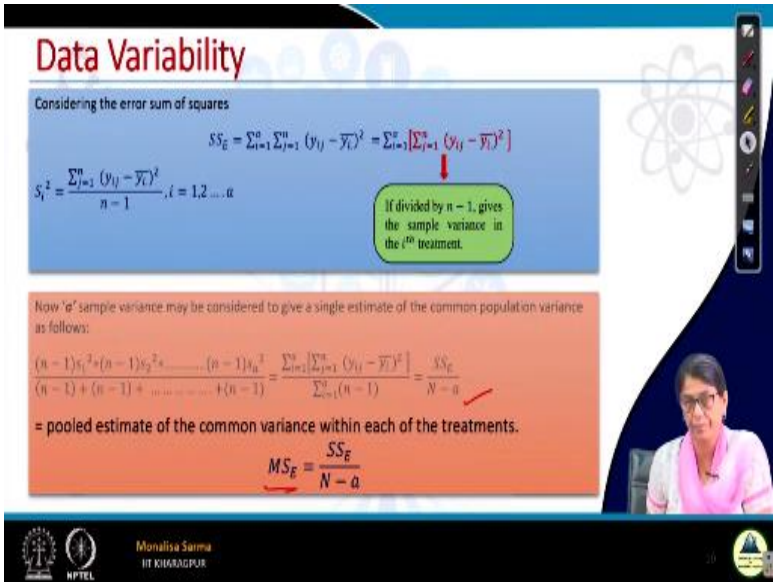
So, now, this is again $\sum$ i = 1 to a. So, if I tried to sum of all the ath level basically what I will get that is nothing but the same thing what we have got a pooled variance estimate for the two population remember our two population pooled variance estimators what was that if you remember n - 1 s 1 ^{2} + n - n 1 - 1 n 2 - 1 s 2 ^{2} / n 1 - 1 + n 2 - 1 this is the degrees of freedom for first one this is the degrees of freedom for the second one, remember this was our pooled variance estimate.

And if you forgot how do you how did you get this simple variance formula S ^{2} is equals to this (x i - x bar)^2 / n - 1. So, if this is sum of squares, this is nothing but some of square that is sum of square is nothing but s ^{2} into n - 1 is not it? So, this is the pool variance estimate. So, similarly, this term if we divided this one time if you divided by just one and minus one I get the variance of the ith level.

Similarly, if this of all these levels if I divide by the n - 1 correspondingly then basically it is nothing but a weighted average of weighted sum of all the variances of all the levels. So, that is nothing again the pooled variance estimate because here see, we have considered the initially only we have assumed that the population variances are same. So, population variances are same. So, we need to find out what is the population variance. So, how did you do that?

We found out the sample variance from the sample variance, we come out with the pooled variance estimate of the population.

(Refer Slide Time: 05:20)

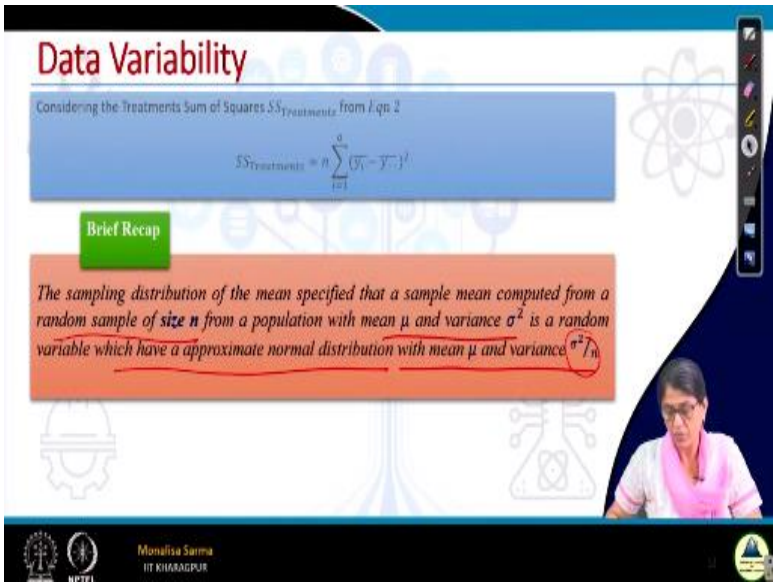


So, basically, if we do this, this particular expression from this particular expression, if we extend it, so, and then just I am dividing by this degrees of freedom of each level. So, this is what degrees of freedom of each level if I get this, this is nothing but this equal to N - a and what is this term this term is nothing but my this term. And what is this term? This term is SS E, to SS E / N - a. So, SS E / N – a, what did they get this? This is nothing but my pool variance estimate.

This is the pool estimated a common variants within each of the treatment. See here, I am going slowly SS T is equals to I got SS treatment + SS E. Now from SS E, I could see that SS E is nothing but if I just divided by the degrees of freedom, I will get the variance of all the pooled variance estimate of the population. So that is what I have done. So, that means my distance add just added one extra one that is divided by N - a.

And that becomes my pool estimate of the common variants within each of the treatment and this term SS E / N - a, I am using a different expression now, for that let me use the expression MS E. Our latest SS E now SS E divided by N - a let me use this term MS E, MS E is nothing but a pooled estimate of the common variance. So, each population has a variance of this SS E / N – a, this we got it from SS E

(Refer Slide Time: 07:00)



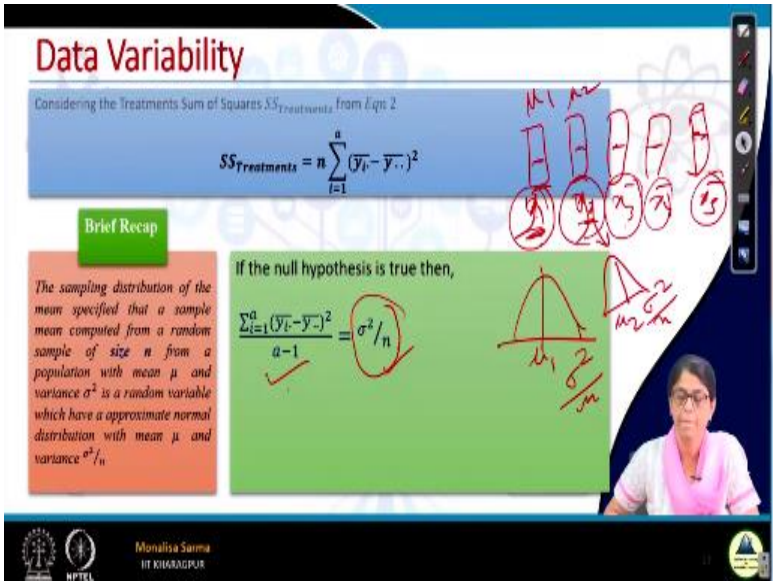
Now, we will come to SS treatment. So, this was my SS treatment formula what is this? the sum of squares of the treatment average from the granaries? Now, you if you recall the sampling distribution of the mean which we have learned when we have learned sampling distribution similar to central limit theorem only central limit theorem just talks about the normality. So, what is the sampling distribution of the mean? Sampling distribution means specified that the sample mean computed from a random sample of size n.

From a population with mean μ and variance σ^2 is a random variable whichever approximate normal distribution with mean μ and variance σ^2 / n what does this say? So, if we take a sample from a normal populations with mean μ and variance σ^2, so, the mean of the sample we have taken a sample there is a population, population has mean μ and variance σ^2 from the population we have taken a sample.

So, now, from the sample we have found out what is the mean of the sample? Now, this mean of the sample let me call it as x bar. Now, this x bar it will have a sampling distribution because x bar different sample we will get different x bar. So, it will have a sampling distribution now, sampling distribution which is approximately normal. So, normal is 2 parameters that is mean and the variance what is the mean and the variance.

So, mean of the sampling distribution is same as population mean that is μ and its variance is population variance divided by n. So, variance of this sampling distribution is population variance divided by n that is σ^2 / n. So, this we know from the sampling distribution of the mean now, we will be utilizing this.

(Refer Slide Time: 08:53)



Now see here so, now try to think and the analogy now what we have we have total a levels a populations from this a population we have taken some sample total a samples we have taken okay each of size n now, this sample has mean μ i each of these sample has mean of μ i and each of the sample has what is the what does this sample has variance each of the sample will have variance σ^2 / n why? See, suppose there are 5 population we have taken 5 samples this is 1 sample 2 sample 3 samples 4 sample 5 sample we have taken 5 sample.

So, from each one suppose this is x 1 bar, this is x 2 bar, x 3 bar, x 4 bar and this is x 5 bar and this x 1 bar will have a sampling distribution which is normal and what will be the mean of this distribution mean will be population mean? Suppose the mean of this population that is the first population that mean of this population is μ 1. So, it is mean is μ 1 and what will be the variance of this population? Variance of this population is population variance divided by n and what is the population variance?

Population variance is σ^2, so it is variance will be σ^2 / n where it is mean is μ 1. Similarly, for this x bar if we have a sampling distribution, what we have a sampling distribution what will be the mean of the sampling distribution population mean suppose this population mean is μ 2 and what is the sampling variance of this distribution σ^2, and here also it is σ^2 n why?

Similarly, if you take x 3 x 3 mean maybe μ 3 variance will be again σ^2 / n, x 4 also μ 4 σ^2 / n, μ 5 σ^2 / n why σ^2 / n for all the case because we have assumed all the populations we have assumed that they have the same variance that is σ^2 all the population initially only we have assumed they have the same variance that is σ^2, is not it? And we have also computed what is σ^2 from the ANOVA entity we have computed what is that.

That pool variance estimated we have computed is not it? So, pool variance estimate we have computed that is I think in the last slide we only computed, this is the pool variance estimate this one SS E / N - a what is this? This is nothing but the σ^2 that is, so, all the population has this variance. So, since, all the population has this variance, so, some variants of all the sampling distribution also will be σ^2 / n. Now, as you mean that null hypothesis is true.

Now, let us assume that a null hypothesis is true if the null hypothesis is true depends what happens all the μ are equal μ 1 = μ 2 = μ 3 is equal to say is μ if we assume that all the null hypothesis are true that means, all that means, all the population that means, all the population has same it means as if all the population that is as if we have taken a sample from a single population only that is all the population has same it is just an extended population there is no difference between the population.

If we assume that null hypothesis is true all the samples as if we have taken it from a same extended populations. So, assuming that all the population will mean will be same μ, μ 1 = μ 2 is equals to 3 to 4 to 5 say this μ. So, if in that case now, suppose the different samples that we have taken the means, now, suppose this is if considering that the null hypothesis is true that means, that means, we as if it is a single population from all the population.

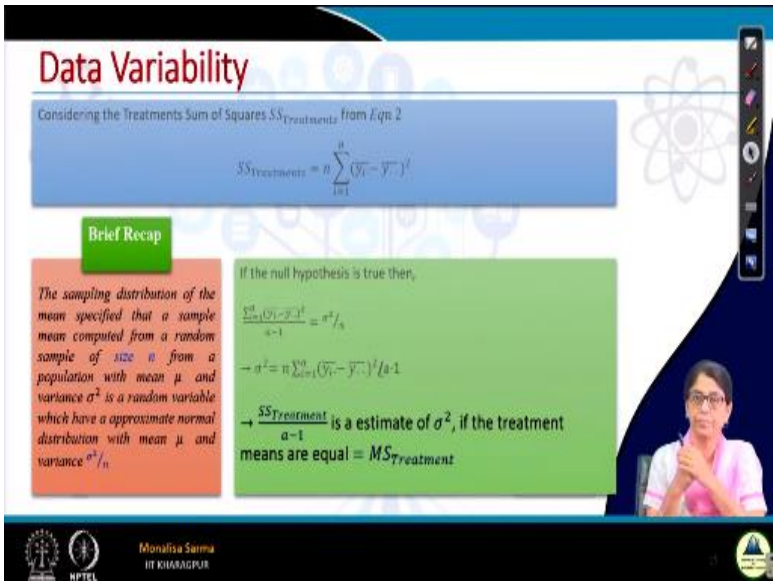
Suppose, we have taken this 5 samples and this suppose this is one is x 1 bar one value I got x 1 bar another value I got x 3 bar x 4 bar and x 5 bar. Means if from one population, we have taken some samples, 5 samples, 5 samples, I have got this data now, if I want to find out what is the variance among these samples, variance among these samples is nothing but this is the case I want to find out the variance among the samples.

So, how will I find out the variance among the sample? Instead of by x here I have written y so, so, consider it is y same thing y or x whatever it is, so, variance among the sample is y bar minus that is the grand mean is not it? How do I find out the variance? Variance is I try to find out the each and every data from the mean is not it? So, when I am taking this so, that means my mean is the grand mean so, divided by how many data total n level so, it is a - 1. So, this is nothing but this is equals to σ^2 / n.

Because all this population we have seen all the samples it has a variance of σ^2 / n is not it? So, we have taken only these samples. So, if this is the case, that means the variance if we now consider this particular nth samples and if we try to find out a variance among these a samples, so, this variance among the; a samples I got this and this is nothing but σ^2 / n is not it? Because I know the variance among this data is σ^2 / n the variance among this data is σ^2 / n.

Variance among these data is σ^2 / n, variance among these data is σ^2 / n, variance among these data σ^2 / n. if that is the case, then variance among this will also I will get σ^2 / n. So, this value is nothing but equals to σ^2 / n then by simplifying it a bit.

(Refer Slide Time: 14:57)

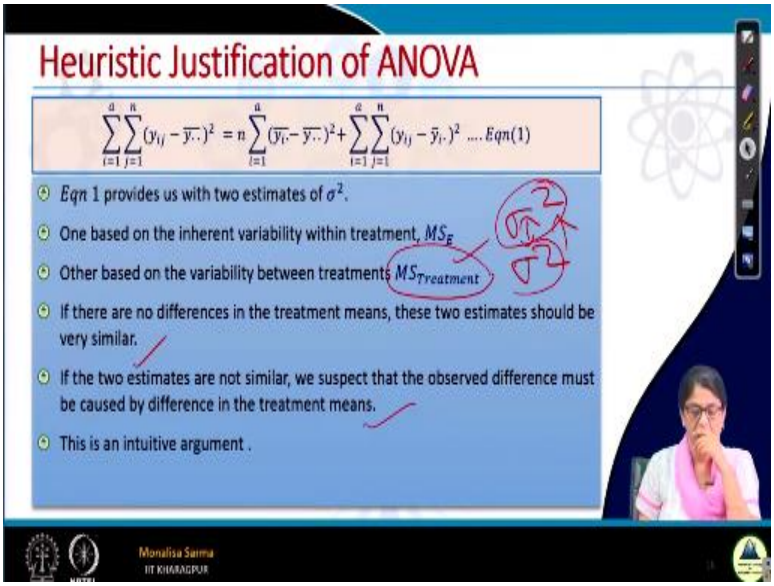


So, what did I get σ^2 you see, so, my σ^2 is nothing but this divided by a - 1 and I have taken the side so, my σ^2 is this by a - 1 and how this value you see this portion this portion I this is my this portion. So, now what is the σ^2 that means SS treatment divided by a - 1. So, SS treatment is the estimate of σ^2 if the treatment means are equal so, if the treatment means are equal I have considered null hypothesis if all the treatments means are equal.

Then I have got this if all the treatment means are equal that means, there is no change there is no effect of the factor if all the treatment means are equal then my estimate of σ^2 is this will SS treatment / a - 1. So, see I got 2 estimate of σ^2 one estimate of σ^2 I got it from the error SS E. So, that I found it for the random error what maybe the variance that is from that also I got one estimate of σ^2 is not it?

That I called the pool variance estimate pooled variance estimate of all the population that is when estimate of σ^2 that is one estimate of σ^2 I got another estimate of σ^2 I got assuming that all the population means are same assume that all the population means are same I got another one estimate σ^2, another one estimate of σ^2.

(Refer Slide Time: 16:36)



So, equation one provides us with 2 estimates of σ^2 one based on the inherent variability within the treatment that is the MS E what we saw SS E divided by this one. So, SS E / n - a. So, that is one variance estimate we got another variance estimate I got is SS treatment / a - 1. So, one based on the inherent variability within the treatment that is MS E other based on the variability between treatment other another estimate of σ^2 I got it by finding out the variability within treatment considering that all the means are equal.

Here I have considered all the means are equal by considering all the means are equal again I got an estimate of σ^2 that is MS Treatment in the earlier post estimated treatment here, I was not bothered whether the means are equal not equal, I was not bothered I was just trying to find out an estimate of the variance by taking all data and the second case I tried to find out the estimate of the variance by considering that all the means are equal.

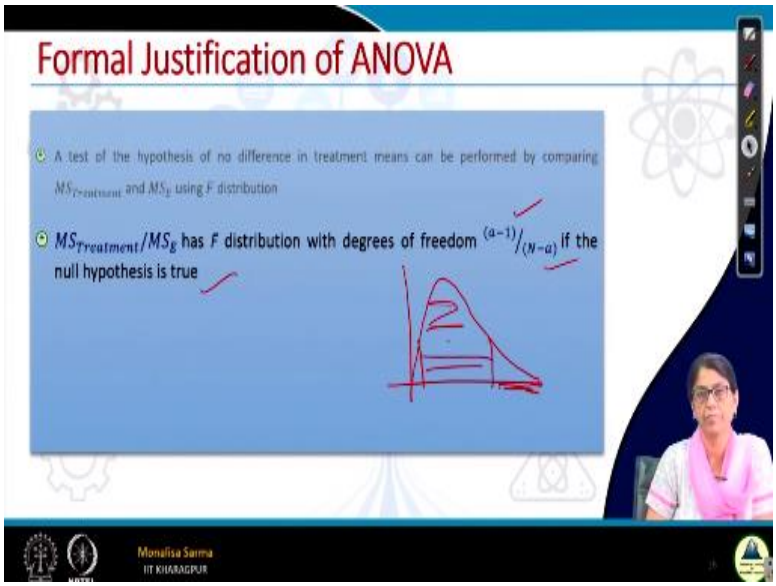
So, if there are no difference in the treatment mean these 2 estimates will be very similar. So, my second estimate of treatment σ^2 I have as you that is also an estimate of σ^2 this is also an estimate this is also a σ^2 this is also σ^2 this also gives σ^2 and MS Treatment also gives σ^2 MS E also gives σ^2, what is σ^2 variants of the populations. Now, this σ^2 and the σ^2 will be equal when because from this I got an estimate of σ^2.

How did they get an estimate of σ^2? Assuming that all the population means the same. So, it may all the population means are same that means, my both the variance estimates are same if all the population means are not same, this MS treatment whatever value I will get this will not be same as MS E, so, if the 2 estimates are not similar, we suspect that the observed difference must be caused by the difference in the treatment mean, so, as you mean that both that all the means are same I got an estimate of σ^2.

Now, I have assuming that all the means are same and I found it an estimate of σ^2 if all the means are not same that the estimate of σ^2 which I find that will definitely not be equal to σ^2 that is the pool variance estimate. So, if the 2 estimates are not similar, we suspect that the observed this difference must be caused by the difference due to the treatment means, one difference is already there due to random error.

So, if the 2 estimates are not similar that means the difference we are getting the difference because of the effect of the treatment whatever treatment we had for the music because of the effect of the music we are getting the difference in values because of the effect of the different treatment we are getting in different values, so, this is an intuitive argument we can also see those in formal approach.

(Refer Slide Time: 20:03)

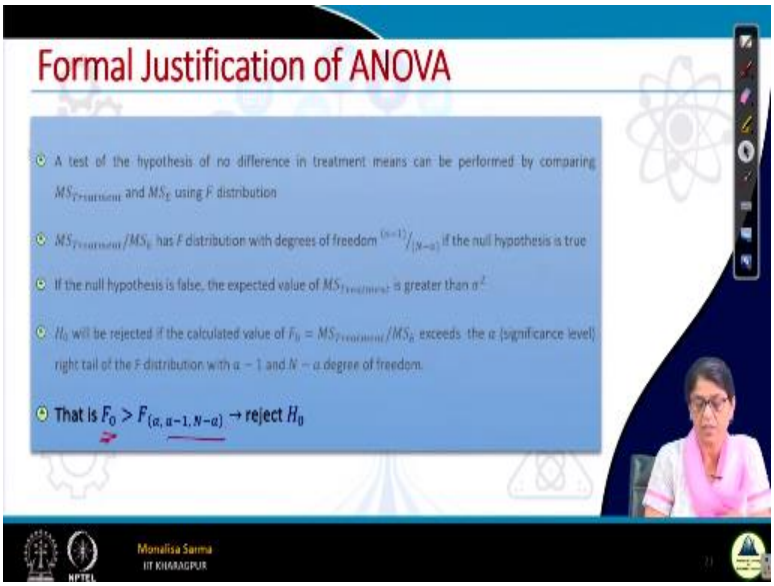


Now so, the test of hypothesis of no difference in treatment means it can be performed by comparing MS treatment and MS E. MS treatment gives us an estimate of σ^2 MS E gives an estimate of σ^2 to estimate of σ^2, how do you compare 2 variance for comparing 2 variants we have seen we use F distribution. So, if Ms treatment / MS E has an F distribution, what will be the degree of freedom MS treatment degree of freedom is a - 1 total number of levels - 1 MS E degree of freedom is n – a.

We have seen why it is n - a is not it? So, if the null hypothesis is true, then this MS treatment / MS E will have F distribution. F distribution means remember the F distribution is something of this sort. So, f distribution means it will be in the acceptable range and MS treatment / MS E it will be the acceptable range. MSE is something I am trying to find out a variance of all the data. So, MS E will not change. Now, what happens if the means are not equal. So, this MS treatment that this variance I will get a very bigger value.

When they made numerator value is bigger. So, my f value f is nothing but MS treatment / MS E, so, in the numerator value is bigger I will get a very high f value that means I am well will lie in the critical region in the rejection region.

(Refer Slide Time: 21:41)



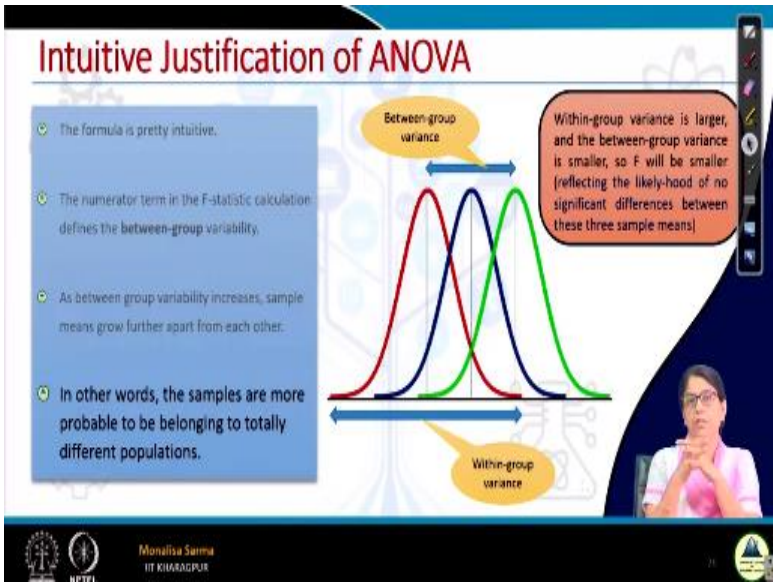
So, here is the null hypothesis is told false the expected value of MS treatment is greater than the σ^2 we will get a high value first. So, H 0 will be rejected if the calculated value of F 0 exceeds

the α significance level that is the right tail of the distribution here we will be using only the right t because our F value will be because it is our F is what am sorry MS treatment / MS E F will never be smaller than MS E it will never returned it is not possible to be smaller.

Why because already whatever variance is there it is there how can we get less than that. So, only a MS treatment will be more either variance within the treatment if it is more. So, Ms value will be higher value. So, MS value is a higher value that means with my acceptance rate my rejection region will only be in the lower tail that means in this region, so that is why it is always one way it is and that is we are trying to test this if f is greater than this value, this is all on this one tailed test and it is greater than a particular value.

We will not check for lesser than a particular value. So, the expected value of MS statement is greater than σ^2. So, if F 0 is greater than this if the value of F 0 what we calculate if it is greater than f (α) for this degree of freedom a - 1 and N - a, then we reject H 0 mind it, it is a one tailed test with the significance level α.

(Refer Slide Time: 23:31)



This formula as we as I told you, it is pretty intuitive. The numerator term that is the in the F calculations, the statistic defines the numerator terms it defines the between group variability okay. So, this is the between group variability this is this between group variable numerator turns

into dedicates its decisive between group variability and this term this is denominator trances within group variability, variability within the group.

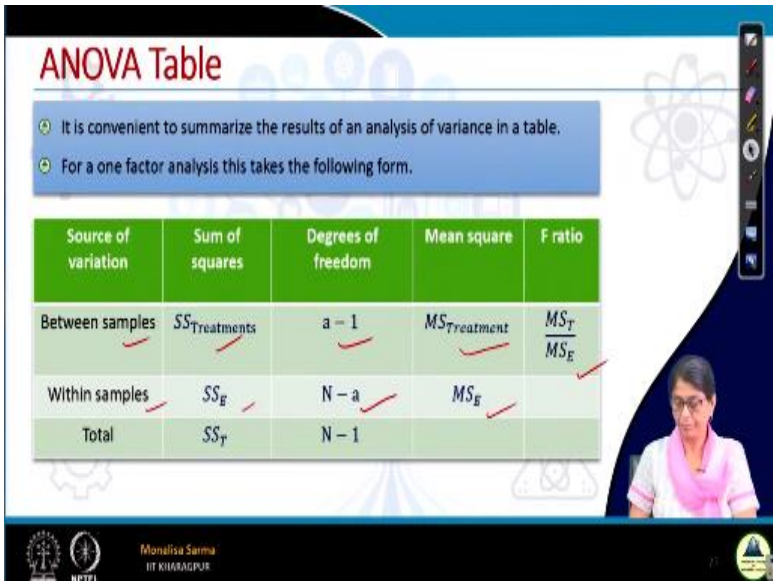
So, this if this variability between group variability is bigger that definitely between good variability is bigger that means what that means, it does not belong to same population that means it belongs to different population that means the population means are different. So, within group variance is large and between groups variance is smaller. So, f will be smaller reflecting the likelihood of no significant difference between these 3 sample means.

So, between group variance is smaller than what happens I will get a very smaller f value smaller value means it will lie in my acceptance region. So, the ath indicates that there is no difference between the population means of all the populations. As between whom were a group variable increase sample means grow further apart from each other in other words, the samples are more probable to be belonging to totally different populations, if the between group variable is quite bigger as if it belongs to a different population.

So, understood that means, from the what is the ENOVA entity? ENOVA entity gives us the 2 variance estimate. One is variance between the group, another is variance within the group. So, when I have to compare this to 2 variants, always I will be using t distribution. So, it might between group variability which is in the numerator, if that variability is a bigger value, if the between the 2 variable is a bigger value, then definitely I will get an F a very high value and if the value f false in the critical region or rejection region that means what?

That means when between group variability is very high, that is why I am getting such a high value and then we cannot say that the populations all the population are mean are same that means, because of the treatment, there is effect on the population. Now, because of the different treatments the population means of the populations are different.

(Refer Slide Time: 26:14)



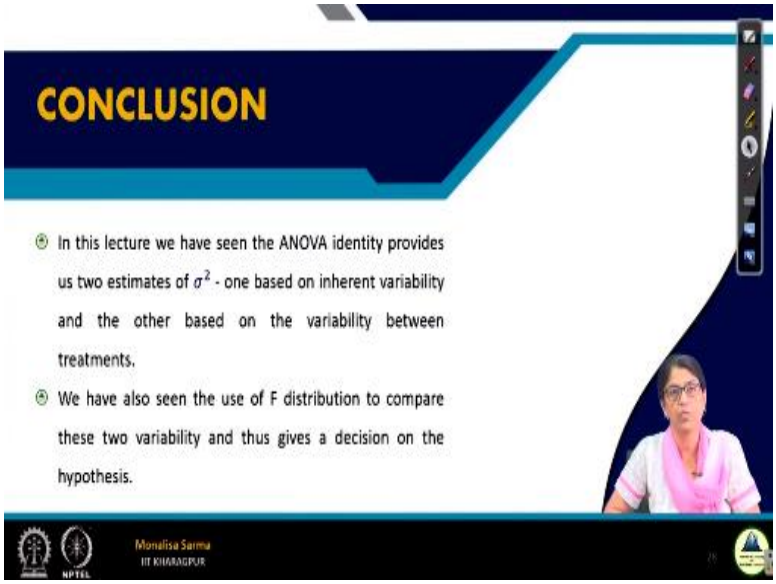
So, it is convenient to summarize the results of analysis in a table for a one factor analysis we can just give it in a table form. So, source of variance between sample that is a SS treatment with degrees of freedom is a - 1. So, then from that we can find it this MS treatment, what is MS treatment is MS treatment / a - 1 gives us MS treatment, then within sample it is SS E. And SS E / n - a we get MS E and F is nothing but MS T / MS E.

So, basically now to conclude so, when we are trying to do ANOVA, exactly what we are trying to do, we are trying to do 2 things, we are trying to find out the variability between different groups. We are trying to find out the variability between different groups that are one then second is we are trying to find out the variability within group. So, variability between groups, I am telling you this MS treatment variable the between group I am terming it as MS E.

Now, if both the terms are equal in both the terms are almost equal not maybe same, but if there is no much of difference in both that terms are similar MS treatment and MS E if they are similar, then we can say that the treatment did not have any effect on the populations. So, whatever variants we are getting there, then the variances may be due to the random error, but if both the terms are very different, that means if MS treatment is quite high, that means the variability between groups is very big.

When the variability between groups is very big that means it definitely belongs from the different population that means because of the treatment, we are having some effect and that is why the population means there is a change in a population mean.

(Refer Slide Time: 28:13)



So, in this lecture, we have seen that provided data provides us 2 estimates of σ^2 one based on the inherent variability and other based on the variability between the treatments. We have also seen the use of app distribution to compare these 2 variability and does gives a decision on the hypothesis.

(Refer Slide Time: 28:32)

So, these are the references and thank you guys

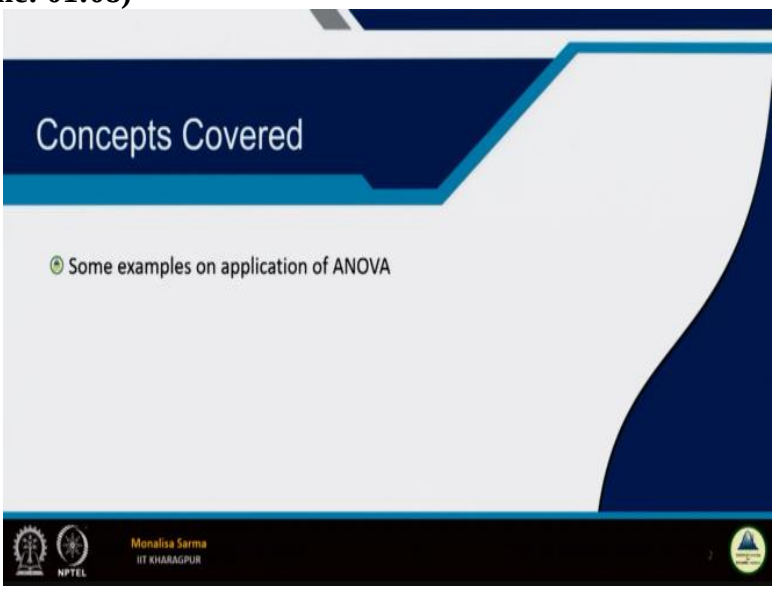
∑αβμσ√λ Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 36
ANOVA - V

Hello guys so, today to continue our discussion on analysis of variance, we have almost covered the required concept of ANOVA by that I mean only 1 way ANOVA because we have just discussed 1 way ANOVA the 2 way ANOVA 3 way ANOVA multi way ANOVA and all this I have not discussed as I told you, once you know that 1 way ANOVA you will be able to learn it by yourself only.

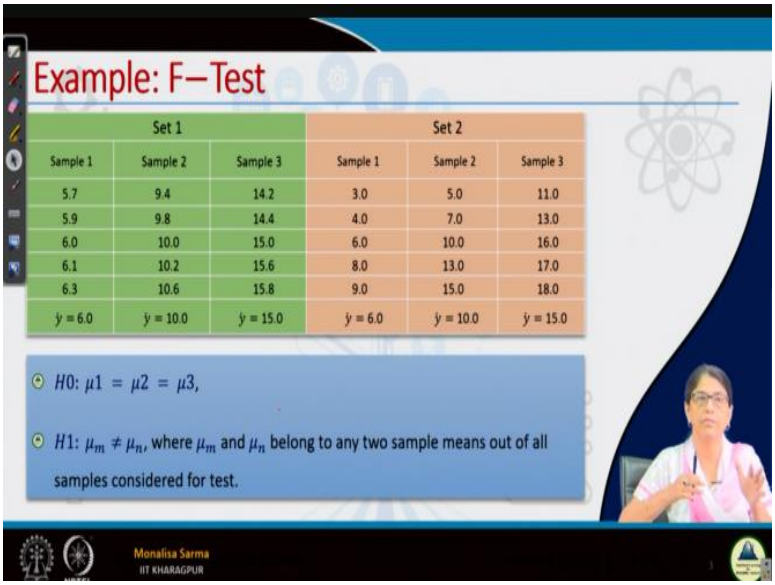
And moreover, it is really difficult to teach all those in a classroom and because intense computation is required for that, however, the concepts are the same with the same concept we will be able to do the rest.

(Refer Slide Time: 01:08)



So, now in today's class, what we will be doing is that we will be seeing some examples, we have fine we have learned applications, we have learned the concepts. Now, we will see how we will use those to solve the problem.

(Refer Slide Time: 01:17)



The first problem that have examples, the first example that I have taken is it is not an that contrary data that data we which we have just created. That is it is not an example as such, we have remembered this data, we have used to come to the conclusion that is why when we want to compare means a different population, why do we need ANOVA? We have used this example for that purpose.

And now, so first example we will try this only with this example, we will see what ANOVA gives and what the result ANOVA gives in this contrary set of data. So, if you can remember properly there in the last class not I think the first class on ANOVA what we have discussed is that there are 2 sets of data that is set 1 and set 2. And we see there are 3 difference we have taken 3 samples from 3 different population maybe we have taken from 3 different population or maybe not.

We do not know like we have to prove that whether all the populations are really different. That is what it means that the populations are really different that is what we wanted to prove that was our objective. So since we see here in the set 1 sample 1 we got some mean of 6 first let us see what we got in sample 1 we got mean 6 sample 2 we got mean 10 sample 3 we got mean 15 this is quite a difference mean quite a significant difference in the all 3 sample.

Now, if we just tell from the mean, you have we can see that and we can tell that, the means are really different that means these 3 samples are taken for the from 3 different populations having 3 different means, if we just take see the mean from the means we can see that similarly, the set 2 also set 2 we see the same set 2 the mean of sample 1 is same as the set 1

mean of sample 1 set, then mean of sample 2 of set 2 is same as mean of sample 2 of set 1, similarly for set 3, also.

So that is, that is why we could also tell for even for set 2, also, all the 3 samples are from 3 different populations whose means are different. And in fact, if we can just from the means we can say maybe that for both set 1 and set 2, sample 1, of set 1 and sample 2, maybe it is from the same population, again sample 2 have both set 1 and set 2, maybe from the same population.

Similarly, for sample 3 for both set 1 and set 2, maybe from the same populations, means 3 different populations from one population, we have taken sample 1 for both set 1 and set 2, for another population, we have taken sample 2 for set 1 and set 2, likewise. So from just looking at the means we could come to that conclusion. So now, remember, we were not very much convinced with that.

Because this difference; in means maybe attributed to many other factors. That is what we have discussed earlier. So, what we have tried to do remember we have plotted a boxplot for this. So when we have plotted the box plot for the set of data, what we saw is that for this sample 1, I do not have the figure right now with me it is in the first slide of the ANOVA and it was lecture first lecture basically.

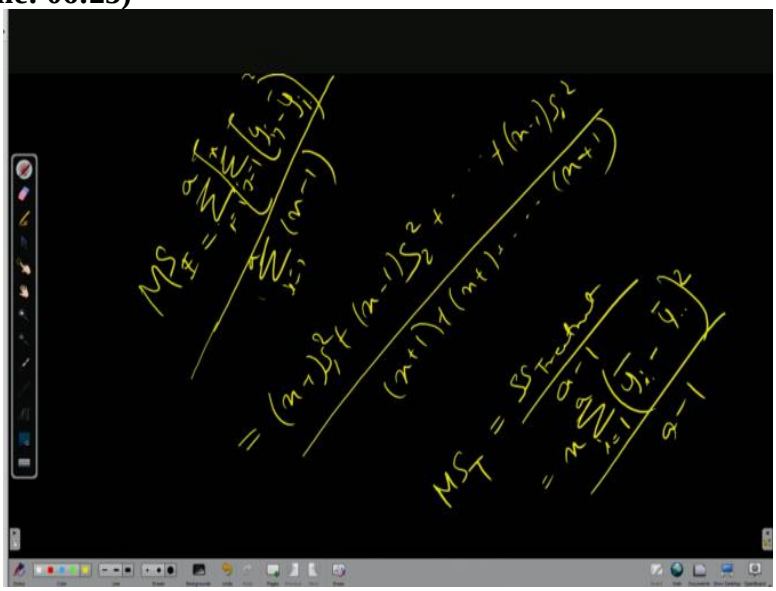
So what does the bar boxplot indicate boxplot indicate the 3 samples of set 1 there from the we could see they are very closely bunched together and that means the variance among themselves within variance, variance within the sample was very less and from the boxplot we could very easily say, this sample 1 sample 2 sample 3 these 3 may have come from different populations, but same we could not say for set 2.

Because it was it has been the huge variance it was as a from the figure as if we can say that as if it has come from a single population 3 different sample, but it has come from a single population it was its variance is so high for all the sample 1 sample 2 sample 3, that means, we could tell that for set 1 3 means are different, but for set 2, we could not say that a 3 means are different, why?

Because we know from sampling distribution of mean we know that if the variance is high, if we have a high variance, that the mean of the sampling distribution is equal to the population mean that statement is not a very precise statement reliability of that statement is not high, we cannot say with accurate precision that the sample mean is equal to the population mean when the variability is quite high variability indicates the standard error among different means.

Sampling distribution of mean means what? Sampling distribution of means, means that different means have different samples, isn't it? So, now, the same example now, we will use to do this, we will be using the same application to find out using ANOVA whether what we got from the boxplot is that the same result we get it or not, let us see that. So, first thing is that if we try doing the ANOVA first what we need to find out for that, you remember ANOVA thing.

(Refer Slide Time: 06:25)



So, if you can remember what was our F value of F we need to find out the value of F. So, what was F if you remember correctly, f is MS treatment. Treatment by MS error, this MS treatment was the difference of treatment mean from the grand mean and MS error is for all those different for all the samples basically we tried to find out the variance within the sample so if you just a quick recap.

So what was the formula for MS error we call it MS E basically. So, what was MS E? MS E = ∑ i = 1 to a indicating that there are total a levels than summation of in each level how many data days was to 1 to n. So, (y ij - y i dot bar) ^{2} and divided by the i = 1 to a n - 1.

Remember this was the formula for MS E basically if you can this also you can write it, it was nothing but a pool variance estimate also distinct.

We can also write it in this way n - 1 s ^{2} + n - 1 s 1 ^{2} basically s 2 ^{2} + n - 1 s i ^{2} then n + 1 + n + 1 + n + 1 this is equals to this we have seen it isn't it? What is this n + 1 up to total how many levels 14 A levels and this as 1 ^{2} is the variance of the individual level. So, n is the number of data in each sample here the samples all the samples are of the same size that is n. So, this is saying we have already seen it.

So, this is how we can calculate MS E we can either calculate MS E by this formula or we can calculate MS E by this formula whatever is this the same thing from this we got this and similarly, what was MS T MS treatment that is MS treatment is equals to SS treatment / a - 1 that is the level minus 1 and this is what was that n ∑ i = 1 to a it is the sum of squares of this treatment means from the grand mean so, it is why i dot - y dot dot this is the grand mean y dot dot whole square / a - 1.

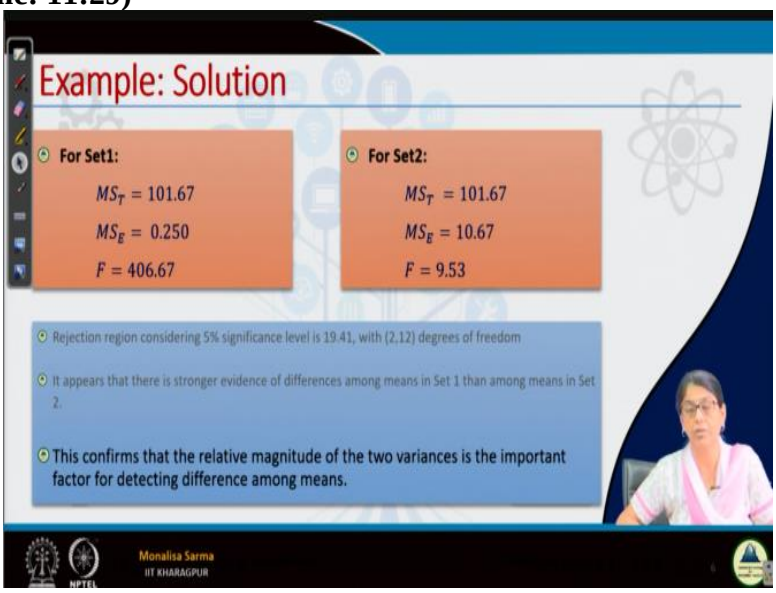
So, this was our formula remember, so, now, for this sample set, if we first thing is that we will calculate what is that MS treatment as well as MS error. So, if you want to calculate MS treatment, what do you need to do? I will just find out a variance from this from each sample I will find a variance So, n - 1 very this is suppose these variances as suppose this variance is s 1 ^{2} this variance of this is s 2 ^{2} variance of this is s 3 ^{2}.

So what is how many 1, 2, 3, 4, 5, 5 is n is 5 7 - 1. So, n - 1 s 1 ^{2} n - 1 s 2 ^{2} + n - 1 s 3 ^{2} divided by how much n - a n - 1 + n - 1 + n - 1 that is n - a n - 3. So, that way we will get MS error similarly, we can find out the MS treatment, MS treatment formula is this the treatment we will find out the treatment mean this is the treatment mean find out what is the grand mean of the grand mean is 6 + 10 + 15 / 3 that will be the grand mean, is not it?

So, from the each mean each treatment mean we will subtract a grand mean for all the levels then divided by the degrees of freedom. So, that way we will get the values of course, the first thing is that we need to know what is the hypothesis we will be able to solve it using the hypothesis testing only is not it the first thing is what is the hypothesis? So, for us the hypothesis is μ 1 = μ 2 = μ 3 that means.

All the populations come from means of all the 3 populations are same that is the null hypothesis. Alternate hypothesis is at least 2 of the means are not equal. So, for that we will have to find out the F ratio. So, if value of f falls in the critical region, then we reject the null hypothesis if the value of L f does not fall into critical reasons, then we cannot reject the null hypothesis.

(Refer Slide Time: 11:29)



So, by doing this calculations, we got MS T is equal to this MS E is equal to this and we got the value of F is this for this this is for the set 1 set 1 we got this value. Similarly, we will do for set 2 also set 2 we got this values F value you see F = 9.53 now, directly from seeing the F value only you can make it out the difference F will remember F distribution is something of this sort and more the value of F means it is going this way more the value of F means F will be this way, and F is 9.53 means f is very much in this range, isn't it?

And as I told you for ANOVA remember it was always a 1 tailed distribution we are only interested in finding out if it is greater than we are not interested in finding out less than. So, because if all the means are not same, that means our MS T value will be quite big compared to the MS E value. So if MS T value will be big that means for more MS T value is more the bigger more the value of F.

So, we are not interested in finding out the lower tail isn't it so lower till when we will get when MS T value is less than only we will get the lower tail and MS T value will be lesser than the MS T value that is very unlikely we are no we are not interested in finding an MS E

MST value will be bigger when the treatment means are different isn't it? So, we are only interested in the upper tail.

So here what is the rejection region if we consider at 5% significance level? So at 5% significance level what is the degrees of freedom here degrees of freedom is for MS T what is the degrees of freedom MS T degrees of freedom is a - 1 that is total level is 3 so degrees of freedom is 2 and for MS E degrees of freedom is n - a and it is 15 total value is 15, 15 - 3 is 12. So, our degrees of freedom is 2 - 12.

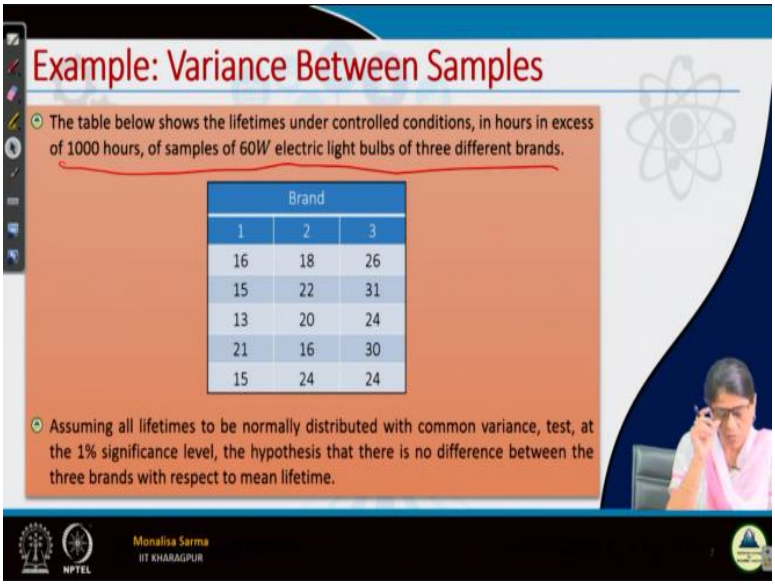
So, if you see the F table for 5% significance level see here we will consider $\alpha = 2.05$ only we will not take $\alpha = 2$ $\alpha / 2$ will not do $\alpha / 2$ because it is single tail it is not 2 tail. So, if we see the table for that we will get the this is develop a test 19.41, 19.41 is the minimum if it is greater than 19.41 then it falls under critical region and see what is the value of F in for the first set F is 406.67. It is much, much greater than 19.4.

And it is very much in the critical region. Even a boxplot also sought from a boxer also we could see that and for the set 2 we got in the acceptance regions 9.53. So though the means of both the sets are same, but still for set 1 we could very conveniently say that all the means are different for the 3 different population null hypothesis is rejected. And for set 2 null hypothesis, we could not reject the null hypothesis.

We do not have enough evidence to reject the null hypothesis. Clear that is how we do the ANOVA testing. So, it appears that there is a stronger evidence of difference among means in set 1 then among means instead to stronger evidence. Since it is very much in the critical region, this confirms the relative magnitude of the 2 variants is the important factor for detecting differences among means.

We have already seen that isn't it, that is why only we have learned ANOVA this was the why we need to do in our just for our justification, we have seen this example.

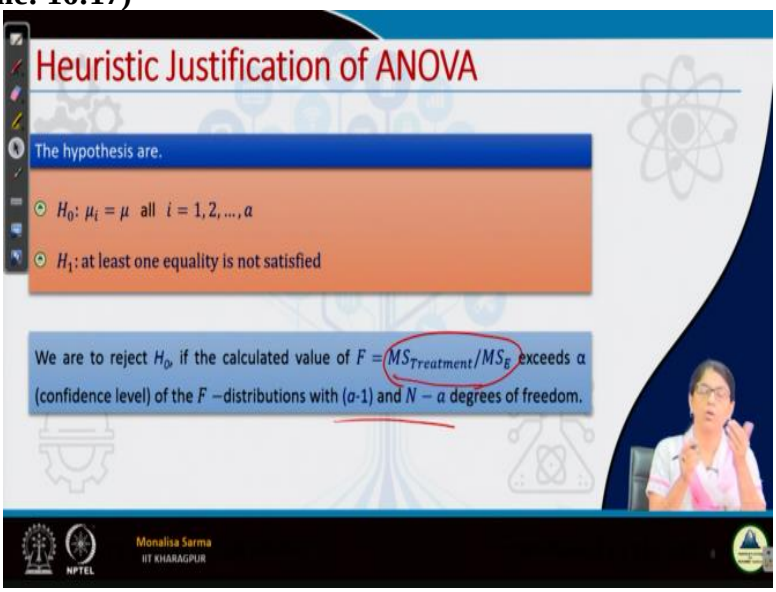
(Refer Slide Time: 15:20)



Now, so, let us take another example, here that table below it shows the lifetime under control condition in hours in excess of 1000 hours of a sample of 60 Watt electric light bulbs of 3 different brands, 3 different brands of lights there it is. So, it gives us the hours and life of the bulb life or the bulb in excess of 1000 hours. So, we need to find out whether the 3 brands are same on the means of 3 brands a lifetime of all the 3 brands are same or not.

So null hypothesis is μ 1 = μ 2 = μ 3 alternate hypothesis is at least 1 of μ 1 μ 2 μ 3 at least 2 are not same so that is the alternate hypothesis. So, we have to test that one person significance level that means 0.01.

(Refer Slide Time: 16:17)



So what are the hypothesis here? Hypothesis is μ is when i for all i = 1 to a. How many here a is equals to just 3. So an alternate hypothesis is at least 1 equality is not satisfied. So we have to reject H 0 if the calculated value of F exceeds A confidence level of the app

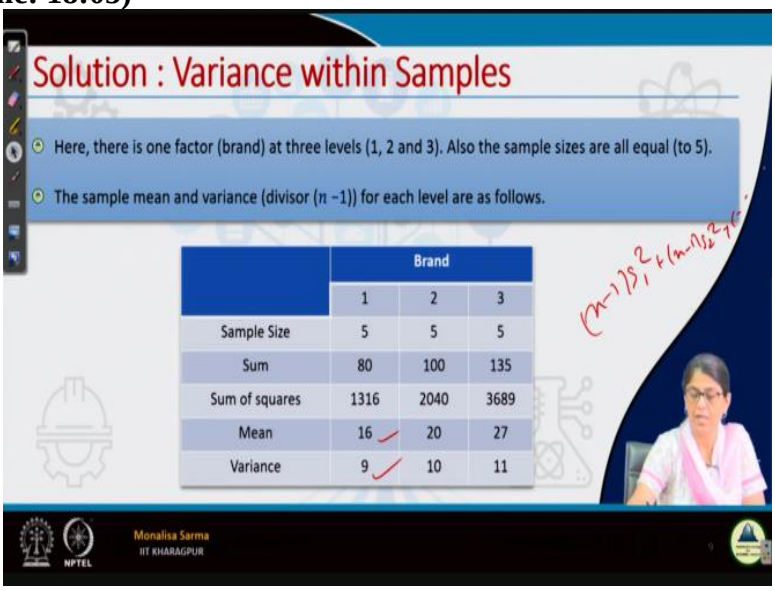
distribution with a - 1 and n - a degrees of freedom, isn't it MS treatment degree of freedom is a - 1 MS E degree of freedom is n - a.

Remember when we have learned F distribution that time it is a general convention in the table it is given that the table follows all the standard textbooks in the table it follows that for the greater the numerator, which is bigger, we keep it in the value of the variance which is bigger we keep it in a numerator that was a while trying to compare population variance of 2 different populations.

But for when we use app distribution for ANOVA, then we do not have to find out which one is bigger and which one will keep in numerator always the numerator will be MS treatment. Because if there is because it does not a difference in the means definitely MS treatment will be more if there is no difference in the mean MS statement will be almost similar to MS E but there of course there will be a slight difference but it will be almost similar to MS E.

So we do not have to bother whether we will have to put the bigger one a smaller one always MS treatment will be MS treatment will always be bigger or may be similar to MS E so an MS treatment is always put in a numerator.

(Refer Slide Time: 18:05)



So now so there is only 1 factor what is the factor that different brands and 3 levels 1, 2, 3 and a sample size is equal to 5 first we will find out the sample this is a good way this actually does I am why have solved with this way any given problem whenever you try to

solve it this way, just draw make a table because then it becomes very easier to calculate make a table first.

And then different levels, put the different levels then sample size for each of course, now we are discussing the sample sizes same for all the samples and of all different populations then find the sum of all the samples values, sum of all the sample values means this well the sum of all these values, sum of all these values, sum of all these values. So, this is the sum then sum of squares, sum of squares means what is the sum of squares.

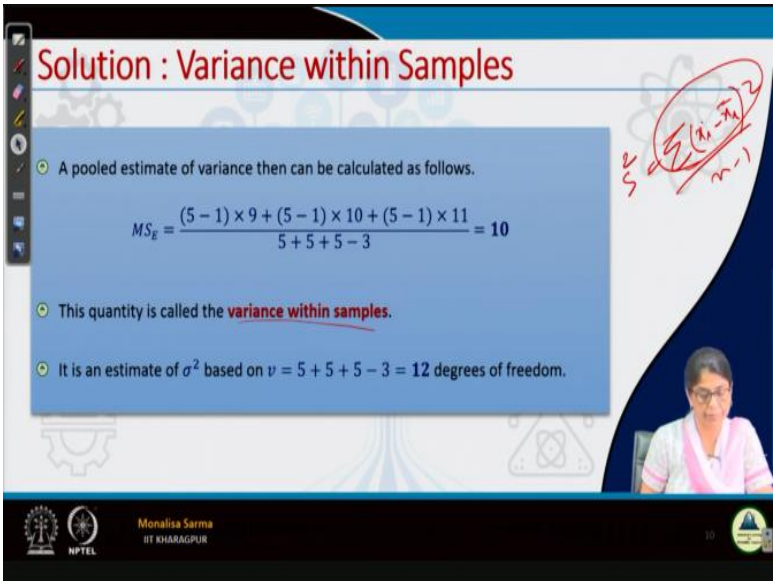
Actually here I did not mention it very properly sum of squares, there are some squares of different types one is the when we are trying to find out the sum of sum of squares between within the variance, isn't it? I mean within one sample, that how do we find out the sum of squares of that, that is, we take each data of a sample subtracted from the mean of the sample mean of the treatment isn't it?

Each data of the treatment substituted from the mean of the treatment and square that is the sum of squares for each treatment. And that is within groups sum of squares basically, and one is between groups sum of squares, how do we find that between groups sum of squares. So, basically, it is better if you have 2 column here, one is within group sum of squares and next is between groups and squares and between groups sum of squares.

How do we find out that is we will subtract them treatment mean from the grand mean and add up all those will suppose we have 3 treatments. Then treatment mean of the sample 1 - grand mean + treatment mean of the sample 2 - grand mean + treatment mean of the sample 3 - grand mean, of course square. So that is sum of squares between treatment. So instead of 1 column here, which I shown as 1 column.

And so it is always better if you do have a 2 column sum of squares between groups and sum of squares between groups then is the mean then you find a variance. Now since once we found a variance, finding out the MS E will be very easy for us, isn't it? What does MS E MS E n - 1 s 1 ^{2} + n - 1 s 2 ^{2} n - 1 s 3 ^{2} how many and divided by n - a. So it will be very easy for us to calculate MS E.

(Refer Slide Time: 20:49)

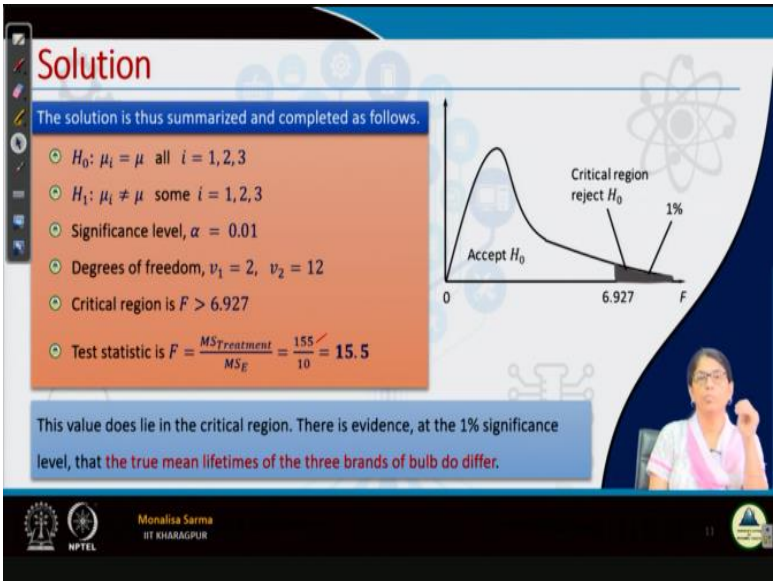


So that is also called pooled variance estimate if you remember. So let us now say it was so easy when we calculate the first we calculate the variance also, if you try to do that, first it at first glance, it may seem it is very complicated, but it is no just find the variance of each sample. Once you find the variance, first find the mean of the sample variance of the sample. And then once you have the mean of the sample, then you find the grand mean.

Grand mean nothing but the mean of all divided by the total number of means, then finding out the sum of squares also becomes very easy sum of squares within sample when you are calculating variance you already calculate that, is not it? What is variance, variance is equals to a sample variance s^2 =, x i - x i sorry x i bar x bar basically whole square / n - 1. So this is variance. So when you calculate variance already, you calculate the sum of squares.

So this is MSE. So, what is MS E? MS E is the variance within samples then as we have already discussed, this is an estimate of is basically the pooled variance estimate, isn't it? Pooled variance estimate, and what is the degrees of freedom of the degrees of freedom is n - a n - number of levels, n is the or sorry, not n, n is the sample size capital and N means sum of all the sample sizes, sample sizes. So, B is degrees of freedom is here total number of item is 15 - 3 is 12.

(Refer Slide Time: 22:32)



So, now, similarly, you find MS treatment also MS treatment, what will be the degrees of freedom for MS statement, the degrees of freedom for a misstatement will be just the level minus one number of level minus one a - 1. So find out the value of MS treatment. So here MS treatment, I did not calculate it and so you can do it very easily so it is 155. So, what is F Statistics we got 15.5.

And what is the significance level is 1% significance level. So, at the 1% significance level from the table, we are using the degrees of freedom 2 and 12. What is 2 is the degrees of freedom of MS statement and 12 is the degrees of freedom of MS E, using this degrees of freedom, if you see the F table, you will see that this value corresponding to this is 6.927. So, any value which is greater than 6.927 it falls in the rejection region or the critical region lesser than 6.927 means is the acceptance region.

So, what we got we got our F Statistics values 15.5. So, it is 15.5 means it falls in the rejection region. So, that means it is false in the rejection region means that means we have a we reject the null hypothesis that means all the 3 brands once we reject the null hypothesis definitely we accept the alternate hypothesis. So, that means the 3 brands or population means the 3 brands are different.

There is evidence at 1% significance level that the true mean lifetime so, the 3 brands of bulb do differ. So, this is a ANOVA, just understanding the ANOVA a bit difficult actually it is not difficult if you try to understand it properly. If you please go once if you cannot understand it once please go through the slides 2, 3 times learn from slide number 1 and

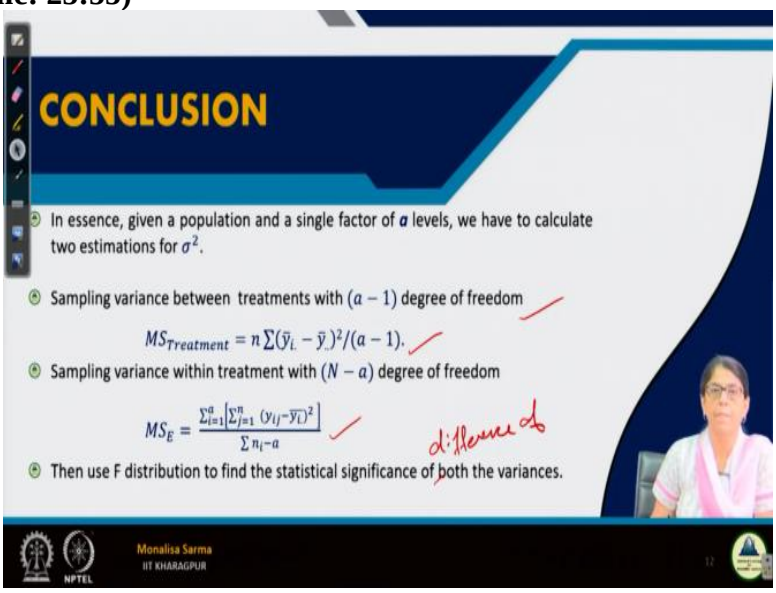
ANOVA I think I have covered in till total 5 lectures and today's the 6th lecture on antibody here we are doing problems despite lecture go through it again and again.

And it will be very easy for you to calculate it is nothing first you should understand how what does actually ANOVA does first you need to understand that basically ANOVA gives us ANOVA entity that ANOVA entity gives us the 2 estimate of the variance. So 1 estimate of the variance we get it from the difference of treatment means, another estimate of the variance we get it from within a sample what is the variance this 2 variants and these are the 2 variance estimate.

So, if the population means are equal then what happened both the variance estimates are same and if the population is not it may not be exactly the same it is similar value and if the population variance or if the population means are not same, then we will get a value which gives them MS treatment basically, that means the variance within that treatment that is a quite a big value and which leads us to F value which falls in the critical region.

And if it falls in a critical reason, that means, we reject the null hypothesis that is null hypothesis is that the all the population means are not same. So, we have enough evidence to say that the all the population means are not see.

(Refer Slide Time: 25:55)



So, now, this to conclude this in essence given a population and a single factor level factor of a levels, we have to calculate 2 estimates of σ^2 sampling variance between treatment with a - 1 degrees of freedom is MS treatment sampling variants within treatment with N - a degrees

of freedom is this then use F distribution to find a statistical significance of both variants statistical difference are both variants basically. So, and in fact, this would be so, difference of both the variances.

(Refer Slide Time: 26:48)

So, these are the references mainly I am covering from design and analysis of experiments. Other books also you can also refer to other books and thank you guys. Thank you.

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

Lecture - 37
ANOVA-VI

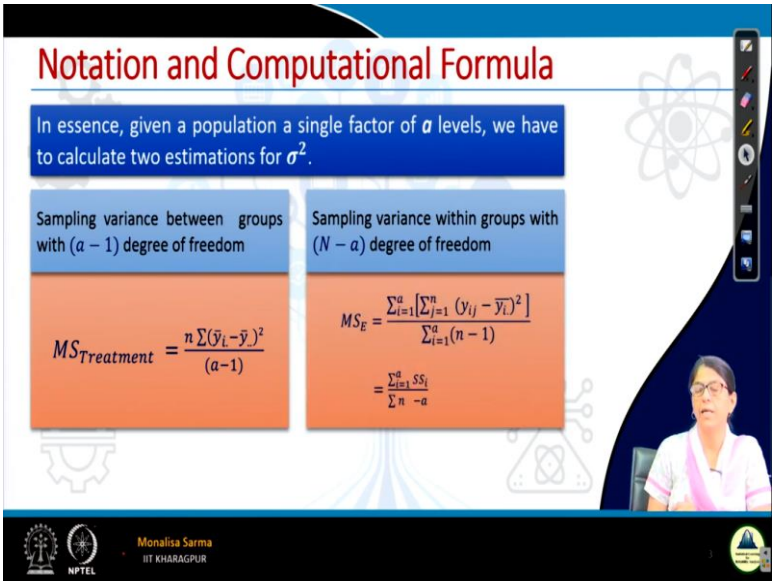
Hi everyone, so, last class we have seen some examples on ANOVA like how we can solve for ANOVA how we can find out whether the means of the different populations are same or they are not same with a particular significance level that is what we have seen.

(Refer Slide Time: 00:45)

And like till last class what we have discussed is that, if you have noticed all the samples that we have considered all the samples are of same size remember all the samples we have taken as n so basically, so, all the samples were of same size some today, we will just take a slight variance of that. So, it is not only possible that always we take a sample of same size, but it is advisable for ANOVA it is really advisable if it is better if we take samples of the same size.

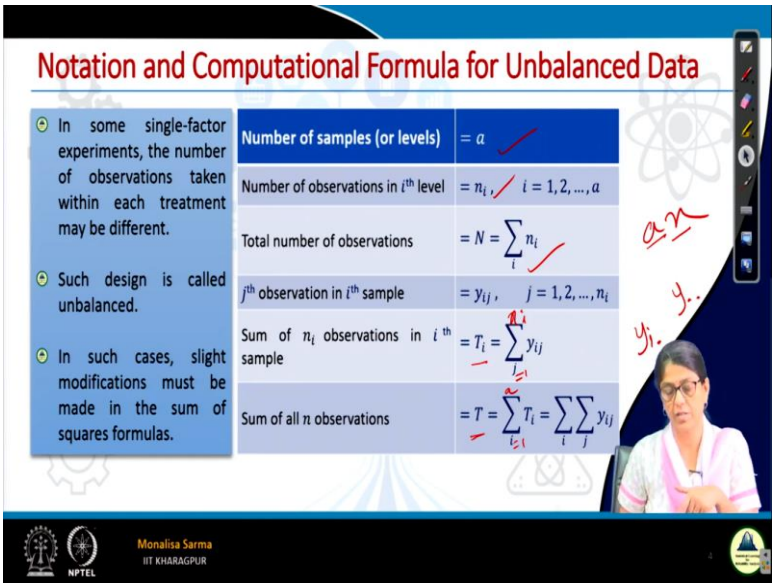
But sometimes it is not possible we have to take samples of different sizes. So, today we will see the computational formula, if we use the samples of different sizes basically, we call that is unbalanced data.

(Refer Slide Time: 01:32)



So, this is what we have already seen in essence given a population a single factor of a levels, we have to calculate 2 estimate of σ^2 sampling variance between groups with a - 1 degree of freedom that is from that we get his MS Treatment. So, then sampling variants within group with N - a degree of freedom that is we call it MS E.

(Refer Slide Time: 01:57)



Than the F is MS treatment / MSE. So, now in some single factor experiment, the number of observation taken within each treatment may be different in some cases, it may not be same due to many other factors of course, our target should always be tried to keep this sample size same, because there are many what to say first now, you may ask why we should keep the same equal sample size first of all see when you have assumed that in remember when we tried solving for 1 way ANOVA we have taken 1 assumption.

We have taken 3 assumption basically out of them 1 assumption is that all the populations variance, we have considered it to be equal that is σ^2 is not it? All the population variance we have considered it to be equal that is σ^2. And for that σ^2, we found out a pooled variance estimate. So, now, if we it is sometimes it is really not possible that all the population variance may be exactly the same.

It is not exactly same, but on when we say same means it is similar and the same range maybe, but sometimes it is that it is not possible to be in a similar range. Sometimes it is a bit from the range. So, if we take the size of the sample as same this slight deviation of variance of this population between different samples, this will not have a very bad effect means this will not make our results incorrect.

This will not make our results this will not make to lose or precision of our results. But if the sample size differs then what happens slight change of a variation of these different populations, it may reflect too much on the our end results. And which makes our result in precise, so, as I told you, it is really not possible at all the variance will be almost same, there will be slight changes. So, that is why it is always better if we take samples of the same size.

The slight variations, if the sample size is not same, it will be reflected in the results and basically we will not get a result which we can be reliability of result will not be high. The result will not be precise result there will be error to it, our type 1 error will increase and type 1 error increase means definitely that is not something which we deserve. That is the first thing. Second thing one more important thing is that if our data is balanced if our sample sizes, say n for all the samples.

Then what happens our power is also maximum that is one way it see if our type 1 error remains low definitely our power will also be more is not it? So when we for unbalanced data what happens our type 1 error increases accordingly our power also reduces. So, it is a direct connection basically anyway I will not show you the total derivation of how the power increases for a stable data and unstable data how a power goes down that I will not deduce and show you but you know how to calculate power we have seen it.

For 1 population we have seen it similar way it can be done actually and I suggest you not to try also those simplification because it is a long computational and there are many software's

to that, we will which will do that when software is available why unnecessarily try doing lots and lots of calculation to find that so, there are many software's by the help of software you can just feed this data and you can see for the thing for the same the significance level for balanced and unbalanced data the same as the power.

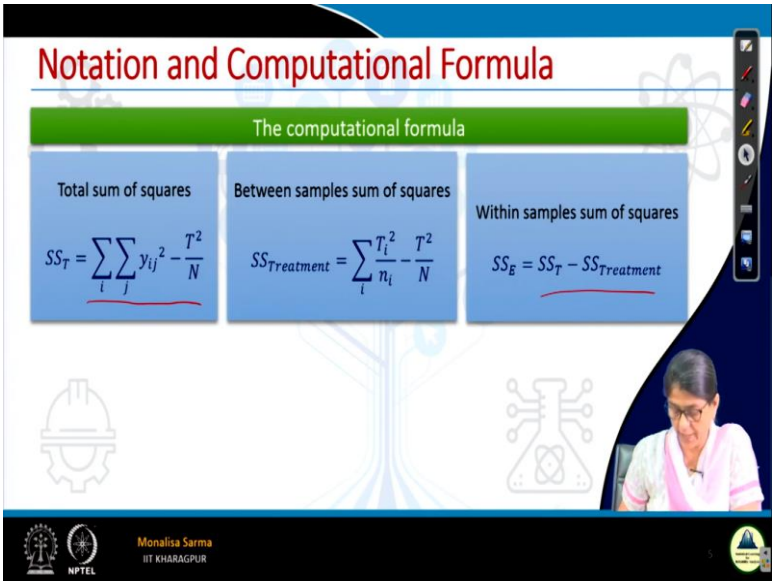
So, now, as I told you this is called unbalanced design in such cases slight modification must be made in this sum of the squares formula when that is cases is when the our design is unbalanced whatever formula for f we have there will be a slight modification in the formula that is obvious right because there in our formulas what we have used? We have directly used n now, here for all the different samples they will be different n it will be n i basically.

So, you see here how is the change in a formula number of samples is a number of observation in the ith level is n i, so, then total number of observation is again N here N is ∑n i i = 1 to a that is number of levels. So, earlier for n is what is N? N directly we could write a into n total number of levels into a number of what to say observation in each level a into n.

So, now n will be this so, jth absorption in ith sample that is same and with j will go from 1 to n i and sum of n i observation ith sample so, sum of n i observation ith sample here, this will be j will go from 1 to n i it will not j it will not from the i goes to 1 to n it will go from 1 to n i for each level, there will be different and maybe same but different for unbalanced data this is the formula instead here I am writing it is at T i actually, if you can remember sum of all the observations in the earlier formula what I was using?

I was using is y i dot remember so actually now the coming formula which I will be using this y i dot this dot dot things it becomes difficult to represent. So I will just use a different nomenclature that is all but it is the same thing, so I am using a different nomenclature that is T i. So do not get confused the same thing, this time using a different variable that is all. So T i is a j = 1 to n i y ij than the sum of all n observations remember it was y double dot I use y double dot in the earlier. So, instead of that I have used a simple T, T is that will be ∑ i = 1 to a T i this i = 1 to a j = 1 to n i y ij.

(Refer Slide Time: 08:42)

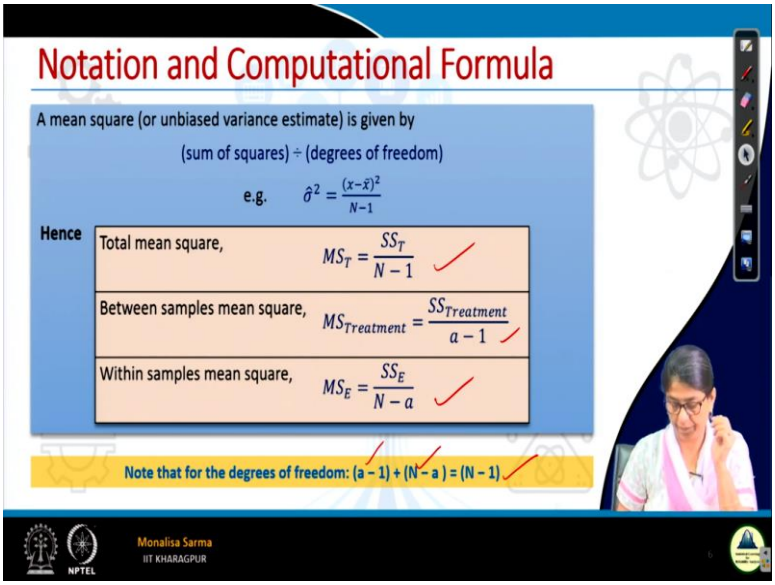


So, this difference will be reflected in their computational formula also, so, how it will be reflected so, SS T is now this, this you can do it by yourself and we will see in SS T earlier what was in SS T in that place now, here you see y ij, y ij will be j will go from j = 1 to n i for each i there will be different n i is not it? n 1 n 2 n 3. So, there will be changed accordingly yourself will be able to solve a simple calculations so, you will get SS T is this.

So, SS Treatment similarly SS Treatment is this I am not doing the derivation you will be able to do it by yourself. So, this is the SS Treatment then and once we know SS T once we know SS Treatment again SS E we do not need to separately find it out SS E is nothing but SS T - SS Treatment what is SS T, SS T is the overall sum of squares. So SS Treatment is the sum of squares between treatment a sum of squares within group I can find out from this formula SS T - SS Treatment, so once I know SS E, once I know SS treatment.

I can find out the MS treatment I can find out the MS E, is not it? What will be MS E? MS E will be SS E - total samples N minus what? N - a, and SS Treatment will be N - 1.

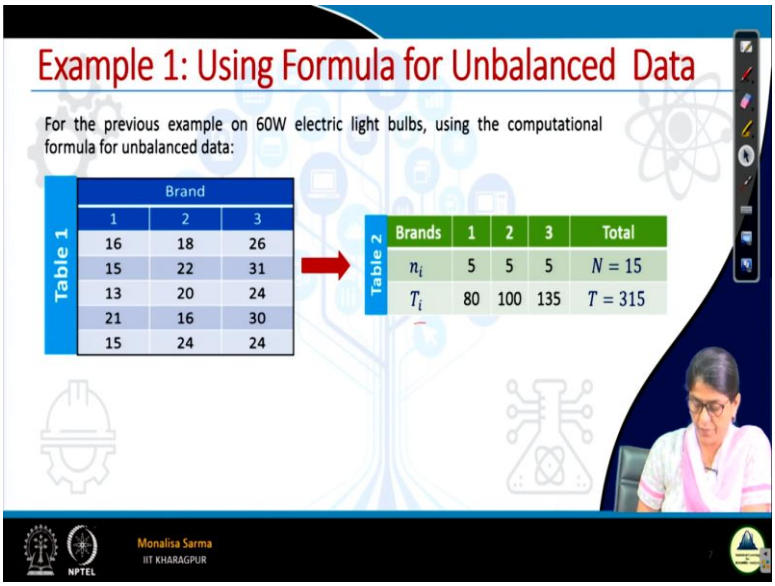
(Refer Slide Time: 10:07)



So, that gives sum of square divided by the degrees of freedom gives us the variance estimate, is not it? So, total mean square is MS T is this SS T / N - 1 MS Treatment is SS Treatment / a - 1 and MS E is SS E / N - a these are same just that there will be a slight difference of this value and this SS T and SS Treatment you can just you take the earlier formula and as earlier formula instead of j = 1 to n there, you will just replace the j = 1 to n i and simplify it you will get it.

So, degrees of freedom is what N - 1 is the degrees of freedom of what, which one, do you remember, N - 1 is the degrees of freedom of the total variability of the data total variability of data degrees of freedom is N - 1, a - 1 is the degrees of freedom of the between treatment variables and N - a is the degrees of freedom within each treatment.

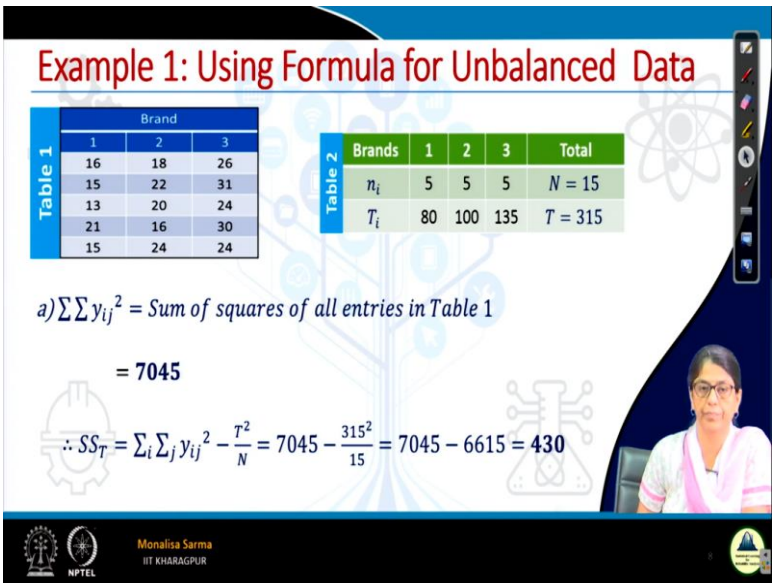
(Refer Slide Time: 11:29)



So, now take an example first, we will see that same example the last class we have seen for that 60 Watt bulb there are 3 different brands. And then we went on to prove that the 3 brands are we have enough what to say sufficient proof that the 3 brands are not same is not it? So, now, though this is balanced, because the sample size is the same of course, when it is a balance or non-balance, we can use the formula for non-balance.

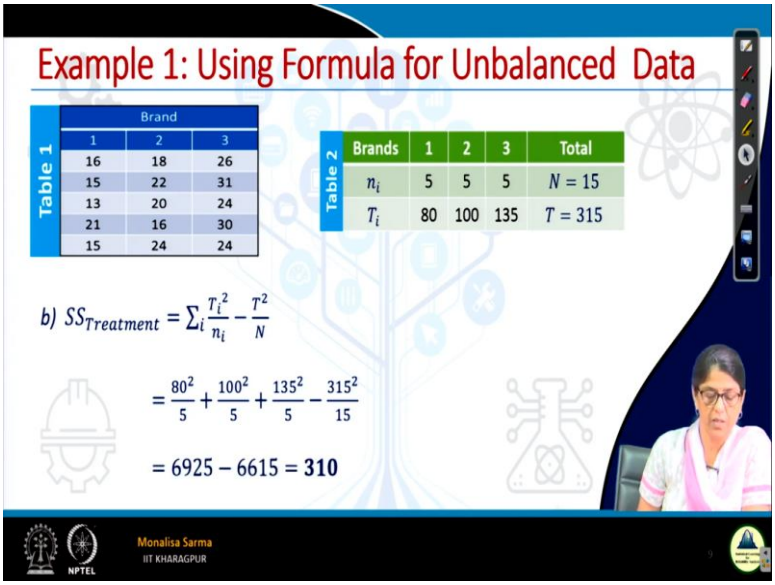
But for non-balance, you cannot use the formula for balance because there are n changes, is not it? So, now, we will use the same example of a 60 Watt electric bulb using we will do it the same problem using the unbalanced data we will see we will get the same results or not. So, here for on further from that we found out what is n i is same for that is 3 and we found out the value of T i, T i is nothing but the summation of all these values remember.

(Refer Slide Time: 12:29)



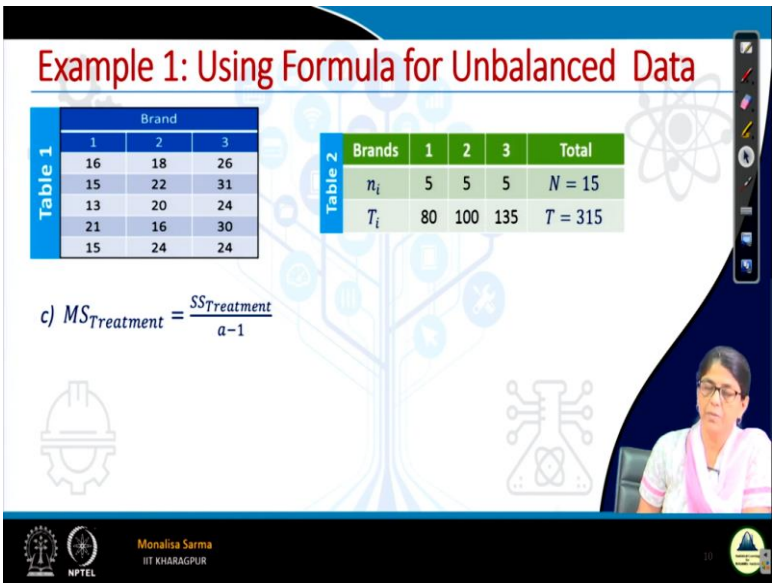
So, now what is this? $\sum y_{ij}^2$ this is the sum of all squares of all the entries in table 1 is not it? y_{ij}^2 then this will give us the value 7045 then what does SS T? SS T is $y_{ij}^2 - T^2 / N$ what is T^2? T is 315 sum of all this T is the grand sum is not it? So 315 so, this is n is 15 we got the SS T we got 430.

(Refer Slide Time: 13:04)



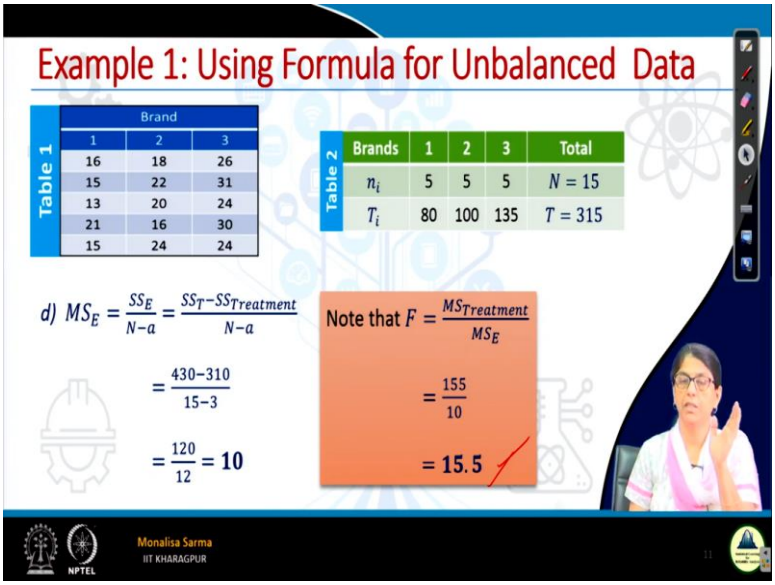
So, similarly, now we will find out SS treatment.SS Treatment this is the formula T i / n i T ^{2} / N. So we have we found out T i what is T? T is the grand total and T i is the total of within each treatment is not it? Earlier for T i we use y i dot again I am repeating for T we use y dot dot. So, similarly we find found out using the putting the value here we found out the value of the SS Treatment once you know SS Treatment and we can find out MS Treatment SS Treatment / a - 1 that is number of treatment - 1.

(Refer Slide Time: 13:47)



So and that is MS treatment, so once we have we found.

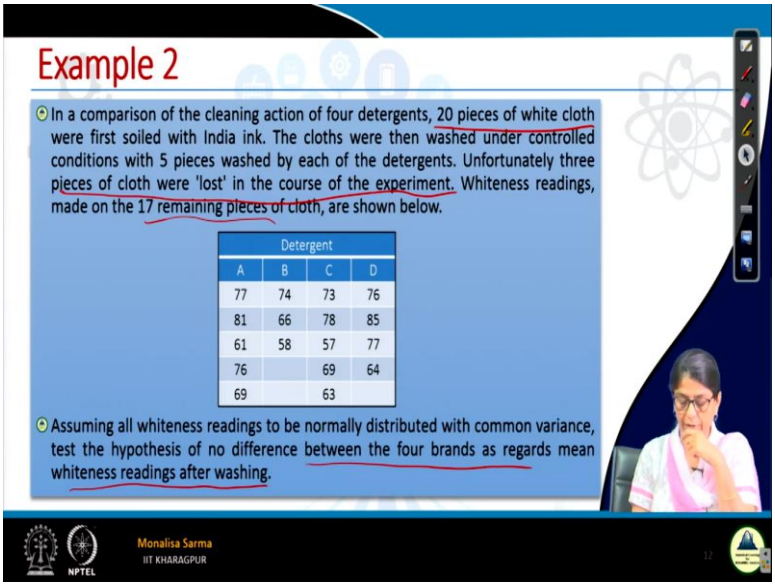
(Refer Slide Time: 13:55)



Now we need to find out MS E for our MS E we can find out SS T - SS Treatment and / the degrees of freedom N – a. This we found 10 and we calculated it and we found 15.5. If you can remember, this the same value we got when in our last class when we have solved for this same problem 15.5 now with the same degree of significance when the significance level at one person.

We will see that this portion this value false in the rejection region. So we got the same values. That means what the formula for unbalanced data we can very well use for the balance data as well. But the formula for balanced data we cannot use for unbalanced data.

(Refer Slide Time: 14:44)



Now we will see a different example. This is an unbalanced example. See in the compression of the cleaning action of 4 detergent 20 pieces of white cloths were first soiled with India ink

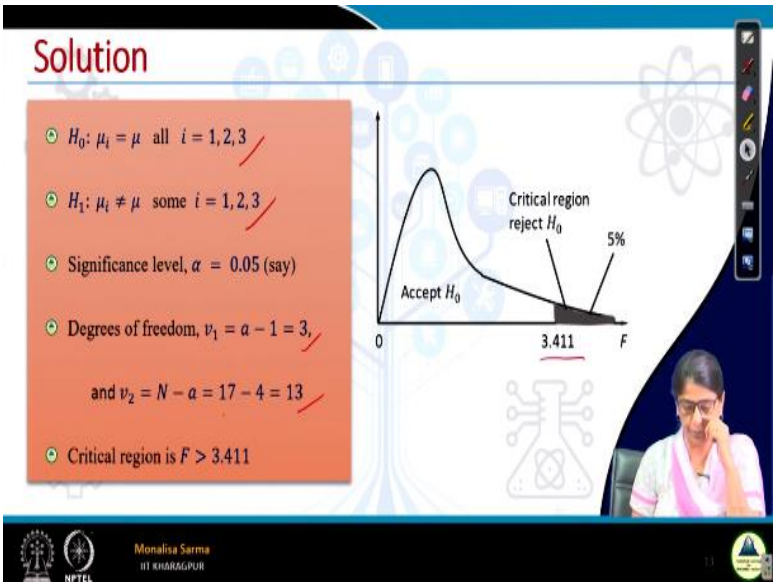
some type of ink. The cloths were then washed out the control condition with 5 pieces wash by each of the detergents there are we have taken 4 different types of detergents and we have soaked total 20 pieces of cloth with some type of ink.

And for each we are trying to we will wash it with 4 different types of detergent to find out there which detergent is best basically. So, unfortunately what happened unfortunately 3 pieces of cloth will last in the course of the experiment, So, definitely our data becomes unstable here, we have tried to keep the stable data because we know ANOVA it is always better if you keep the same number of samples the sample size same.

So, whiteness readings made on the 17 remaining pieces of cloth are shown below. So, for detergent A we got 5 pieces, good for detergent B we got 3 pieces, C we got 5, D we got 4 very much an unbalanced data for detergent type of B we got 3, detergent type of D we got 4 as you mean all whiteness reading to be normally distributed with common variance is the hypothesis of no difference between the 4 brands as regard mean with whiteness reading after washing.

So, we have to test the hypothesis that the mean what to say whiteness readings, whiteness, how white is the cloth after washing is there suppose there is some scale whiteness scale based on that scale that mean reading as is the mean reading of all the detergents same or is it different, that is what we need to find out basically. So, our null hypothesis is mean of all this is same alternate hypothesis at least 2 are not same.

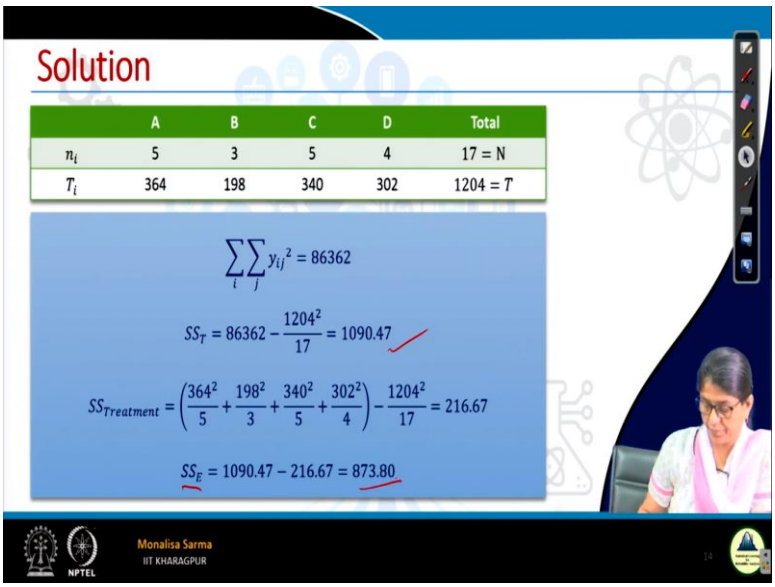
(Refer Slide Time: 16:40)



So, now, this is the null hypothesis, this is the alternate hypothesis, this is the significance level consider a significance level of 5% that is 0.05 then degrees of freedom of course, v 1 is 3 number of treatment is 3 - 1 sorry, it will be 4 different brands of detergent 4 - 1 is 3 and v 2 also 17 - 4 is 13. So, we found a critical region with for 5% significance level and with degrees of freedom 3 and 13.

5% significance level with degrees of freedom 3 and 13 the value of F is 3.411 So, now, we will this is from the table we got this is the value this is the critical value of F. So, any value greater than that will fall in the critical region, rejection region so, now from this data, we will find out the what value of F we get? We will find out MS Treatment we will find out MS E, NS Treatment / MS E will give me the F value.

(Refer Slide Time: 17:47)

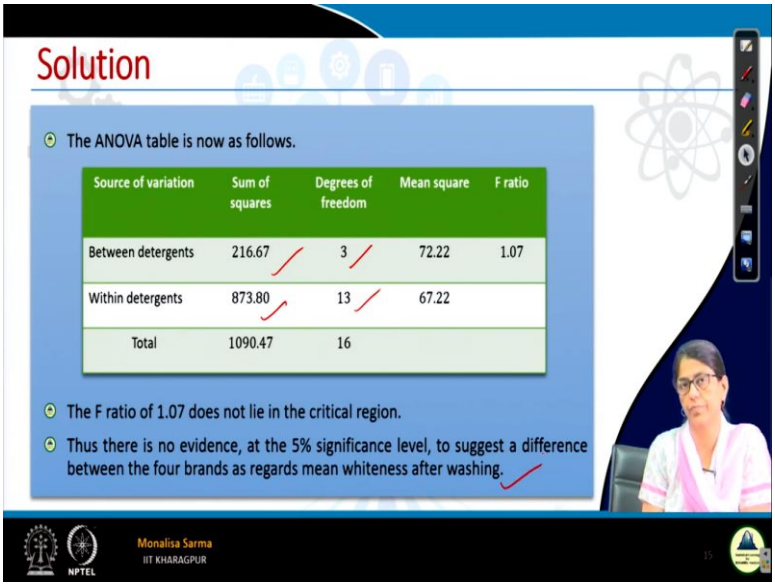


So, we found out first y ij ^{2} we found out this as the n I, n i is the sample size for these 3 samples then we found out T i sum of all the samples of sum of all the observation of first detergent some of the all the 3 observation of second detergent this is the T i value and this total is the T some of all this is T so, for our SS T is y ij ^{2} and this T ^{2} / degrees of freedom.

So, this is what our SS T value is not it? Our degrees of freedom are for degrees of freedom SS T what is the degrees of freedom? Overall variability in the data overall variability in the data what is the degrees of freedom? Degrees of freedom is N and it is whole total number of observations of all the brands together minus the number of total levels is not it? So here what is the total number of observations?

Total number of observation is 17, and then how many brands are there 4 brands are there is not it? So then we found out the SS Treatment we found out similarly we found out SS E. What is SS E? SS T - SS Treatment will give us SS E.

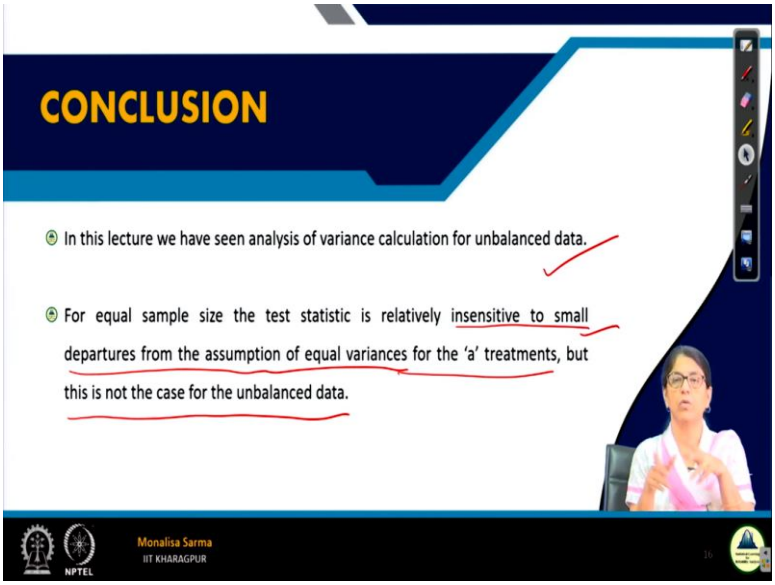
(Refer Slide Time: 19:11)



So between the detergents we found this is the sum of squares this degrees of freedom is 3 to 4 different samples. Within detergents this variance we found is 873. This is the value Within detergents is this one, this one also this is between treatment. This is the within treatment 873, 873 within the treatment 873 than this degree of freedom is 3 this degrees of freedom is 13. Then we found the F value is 1.07, 1.07 is the F value.

And what is our critical reason, our critical reason is 3.411 then 1.001 is something here in this range. So, it very much lies in the acceptance ranges that means our null hypothesis is not rejected. So F ratio does not lie in the critical region does, there is no evidence at the 5% significance level to suggest a difference between the 4 brands as a result, mean whiteness after washing. So, there is no evidence to suggest that there is a difference, difference of whiteness in the different 4 brands.

(Refer Slide Time: 20:40)



So that is what we have what we have seen for ANOVA I do not think now, I do not think initially I started with ANOVA discussion telling that it is a bit complicated, but now, I think you will not tell that this is complicated, it is very easy just find out the 2 variance estimate. One thing is completed this is to understand how to find out the 2 variance estimate. If you just try to do it; using the formula that will be very easy, but why do not; do that, why we are using this formula that you should understand.

That is what we have tried to explain in all the lectures why we are using the how this formula has come when we are trying to find out the variance within treatment, when you are trying to find out the variance what to say between treatment and why we are trying to find out this 2 estimate of σ squared and why we have used F distribution, why does it mean greater will give us a getting a value of F which is better than the critical region why does it is this what to say the null hypothesis is rejected all these things you should understand.

Just calculating it is very easy, and this is one thing this is what we have done this is for basically when you try to compare the means of the different population. Now, similarly, when you try to compare the variance of the different population, there is also another technique that is called MANOVA. So that we will not be discussing this is beyond the scope of this course. So, in this lecture what we have seen we have seen analyse variance and variance of pool unbalanced data.

And see as I have mentioned, for equal sample size, the test statistic is relatively insensitive to small departure from the assumption of equal variance. I have mentioned this is not it? For

equal sample size the test statistics whatever test statistics means here, the test statistic means that F value, when he call it a test statistics, it is when I am doing chi square distribution, my test statistic is the chi square value, when I am using F distribution, my test statistic is the F value t distribution my test statistic is the t value.

So here for equal sample size the test statistic is relatively insensitive to small departure from the assumption of equal variance for the a treatment for the a treatment, we have assumed that a variants are equal, but small departure from the assumption that it is equal it is if we take equal sample size, it is insensitive to it; it will not have what to say it will not have a results which will give us a which will inflate our type 1 error.

So, but this is not the case, but unbalanced data unbalanced data or type 1 error gets inflated if there is a departure from the assumption of equal variance. So, that makes the test if a slight variation if it is very sensitive to slight variance that means what? That means the test is not robust. If it is insensitive to small variance that means the test is robust. So, when our test is robust when we take the sample is of equal size.

So that is why it is always advisable to take for ANOVA takes sample size, equal sample size, but always it is not possible as we have seen in the example.

(Refer Slide Time: 23:54)

So these are the references and thank you guys. So we have completed an hour today. So we will be starting a new topic in my next lecture.

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

Lecture – 38
Correlation Analysis (Part - I)

Hello guys. So, today we will start a new topic while before starting the topic, let us discuss few things like AMC annual maintenance contract I think all of you know about this right what is this? We have heard of this term annual maintenance contract for a water filter and annual maintenance contract for the washing machine for AC. So, it is something all of us we use it right in our day to day life we have to deal with is AMC. What is this?

This is basically a maintenance strategy and we do this we use this annual maintenance contract so, as to first so, that our system does not fail our machine does not fail. So, now, when the system becomes more and more complex, this annual maintenance contract it becomes very difficult as well as very expensive and a new asset a new engineering discipline as EMRs like it is called prognostic and help management.

To deal with such complex machine to different sort of maintenance strategy by which we can like find out we can like improve the state of the health of the system, we can monitor the health of the system and by that we can improve the health of the system one technique of PHM is called remaining useful life estimation. So, it is called RUL estimation remaining useful life estimation. So, it is RUL is gaining wide acceptance in most complex system and many of the complex system in RUL basically, it is totally data driven methodology.

What we do here is that we use different data. Now, this data is what data does is condition monitoring data. This condition we take this data continuously we take the continuous monitoring data and how do we collect it there are different ways of collecting this data maybe we can collect this data by if we have some what to say sensors added in included in different equipment's in a complex system we have sensors in different, different components different sub components level.

And then we collect this data and from this data we try to estimate the remaining useful life. Now, how do we do that basically, we have lots and lots of data now, okay sensor data let us dig into sensor data, we have many sensors for a complex system say around 20 sensors 50 sensors and we have data from all the sensor this data may be ranging of different pressure variable temperature variable different capacitance value different resistance values different power converters, different power convert as well as a different rate how it degrades.

So, essentially what we need to do is that now, when we have this data, we need to analyze this data analyzing means there can be different things, we may want to find out is there any relationship between 2 different data 2 different variables of data like so, we may be interested in finding given a data can we predict some other data so in today's topic and income in fact, I will from this today's class onwards, I will be mainly dealing with I will be mainly talking with techniques for this data analysis.

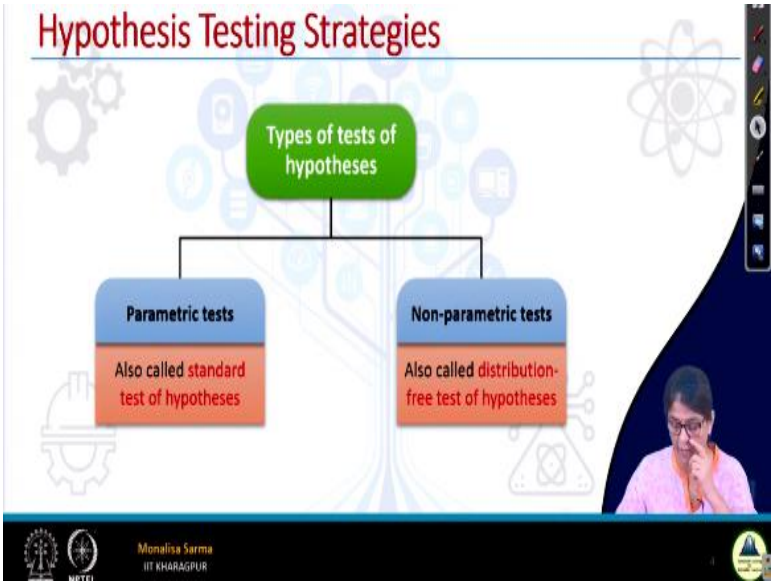
So, this is what I the example what I have given is totally from the reliability perspective. So, now this data analyses technique, which I will be discussing, it is not only for not only applicable for what to say reliability is the studies it is equally applicable for any data analyses strategy, data analysis, wherever data analysis is necessary, it is this methods are applicable to all such situations.

(Refer Slide Time: 03:59)

So, now, coming to our discussion on data analysis move towards thing which we will learn here today is relational analysis, given the 2 variables of a data like when I talk about variables of data, I think, you know, because what are the variables of data we talked in my in my first lecture, first lecture series of this course. So, when if there are 2 variables of a data lake say temperature and pressure, so, now then, is there any relationship between these 2 data?

So, that is what we need to find out. Now, this relationship when we talk about this relationship, is there any way to measure this relationship? So, how weak the relationship safe is how strong the relationship is there some way to measure this relationship? If yes, then how that we will be discussing in this lecture. Now next, one of the methods of relationship analysis is correlation analysis. So under the banner of correlation analysis the different techniques one such technique is Pearson correlation analysis and today's lecture will be discussing Pearson correlation analysis.

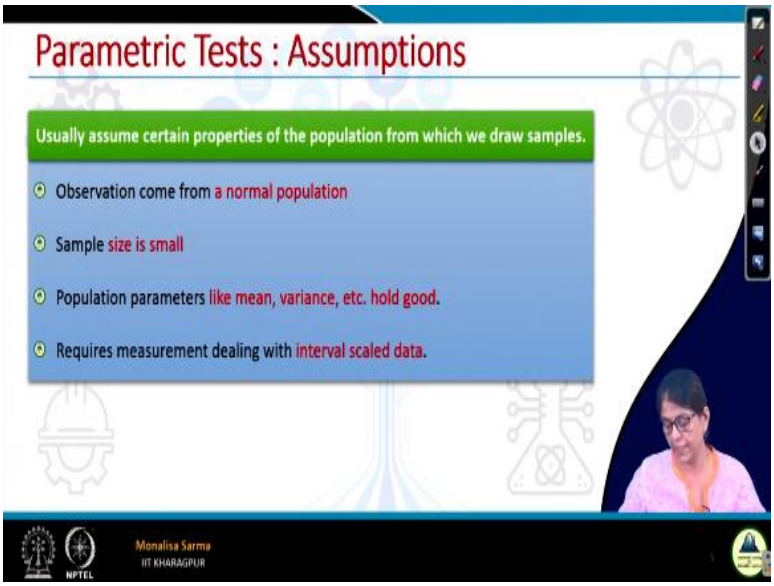
(Refer Slide Time: 05:11)



So, now before going to relation analysis first take a quick recap of hypothesis test which we have done in a last couple of classes. So, the hypothesis test which we have done it is for that whatever we have assumed populations we have taken the populations from the we have taken a sample from the sample we have tried to find out the infer something about the population. And though for all these cases what we have talked about that is always we have assumed something about our population.

This type of hypothesis test is called parametric tests, so, it is also called standard test of hypothesis. And there's another one test which is called nonparametric tests. This is called the distribution free test of hypothesis we do not assume anything about the parent population here.

(Refer Slide Time: 06:02)



Now, so, this parametric test what we have used in the hypothesis testing till now, what we have done is parametric test parametric hypothesis testing that means, here we have certain assumptions about the populations what that we have considered like the populations has to be it has to come from a normal populations, I mean the parent population has to be normal populations, and the sample size is reasonable it is not very big populations, it is a reasonable size.

In fact, for a small sample if the population is totally normal, then small sample size will also so do and in the parametric cases, what we usually try to do is that we usually try to infer about the population mean and variance from the sample mean and variance from the sample statistics, we try to infer the populations mean and variance. So, population parameters mean and variance this sort of things holds good when we are talking about the parametric test.

Then the fourth thing is that it requires measurement dealing with interval scale data interval scale data actually, this is I did not mention it in my previous classes, as you mean that you guys already know about this. So, at this point, I think I will talk about it if you know it is very good,

it will be a quick repair recap if you do not know that it is better than better means now you can learn actually, so, when I talk often variable and the first class have discussed variables there are 2 types of variables remember.

One is quantitative variable and one is qualitative variable quantitative variables basically is the numeric variable and another is qualitative variable. So, numeric variables basically there are 2 types of numeric data one is we call interval scale data and ratio data basically numeric variables only. So, an interval scale data and ratio scale data there are 2 types of numeric variables. So, there are only one difference is that difference is that an interval scale data 0 has meaning and a ratio scale data 0 has no value.

An interval scale data basically we call it interval scale data when we can measure the data in continuum within an interval, like set temperature ranging from - 10 degrees centigrade to + 10 degrees centigrade that is one interval another interval may be from 10 degrees centigrade to 30 degrees centigrade another may be 30 to 50 degrees centigrade. So, here when I have taken the range from - 10 degrees centigrade to + 10 degree here in between 0 is also there 0 has a value here. So, this is interval scale data.

But if at the same time if I measure the data in Kelvin instead of measuring in Celsius or Fahrenheit if I measure the data and Kelvin, so, I cannot take this range from - 10 to + 10 because 0 degree Kelvin does no value. Similarly, if you consider height, zero height has no value has no meaning. So, height cannot be interval scaled data it will be a ratio scale. So, like weight it will be a ratio scale so, this is both are numeric variables only with this slight difference on meaning of 0 in interval scale data there it is mean 0 has its own significance.

In ratio scale data 0 does not have any significance and this is for quantitative variable that is numeric variables. Now, for qualitative variables we also call this categorical variable I think you remember that so, categorical variable there are 2 types one is ordinal and another is nominal. So, what is nominal data nominal variables are those variables which we can divide into different groups and there is no intrinsic ordering among those.

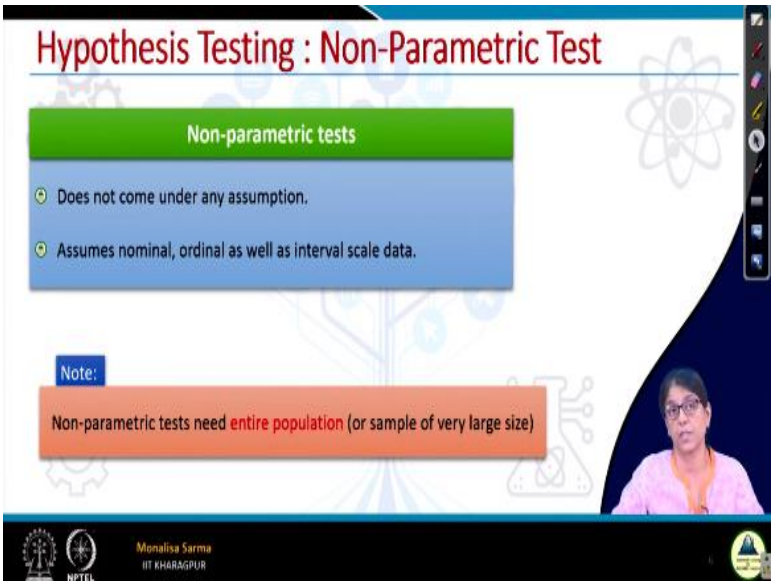
There is no order we can divide this into different categories, but then there is no ordering among these groups like you can take an example like PIN code PIN code of different places in a state. So, PIN code if we consider in a district maybe the PIN code of different places in a district if we take this different PIN codes different PIN codes indicates different places, but is there is no ordering among them. So, that is a nominal data.

Like again the nominal data one example I suddenly remembered. So now suppose I divide the population into different groups one group who is COVID vaccinated another group who is not COVID vaccinated. So, they are 2 different groups. So, they are nominal groups they are nominal data since there is no ordering I cannot say this is this is better than this is based on something that nothing not like that I cannot order them this is the has the highest order this is the lowest order nothing like that.

So, now, this nominal data if it has only 2 groups if it has a 2 group we; call it as dichotomous variable. So, you have had 2 groups let us suppose we consider gender if we consider gender has this to group male and female then it is a dichotomous variable dichotomous type of variable, it is basically a nominal variable but we can since it is this 2 group we can call it as a dichotomous variable. Similarly, now come to ordinal data ordinal data similarly, if we can divide this data into different groups, but there is intrinsic ordering among them.

Like the size of a garment size of a garment is very much an ordinal data size of a garment is triple x double x x medium small this is the ordering among them. So, then again policies of the government policies of the government may be very good, good, bad, worse. So, there is an ordering among them. So, these are called ordinal data. So, when we talk about parameter test so, parameter test it deals with an interval scale data.

(Refer Slide Time: 11:20)



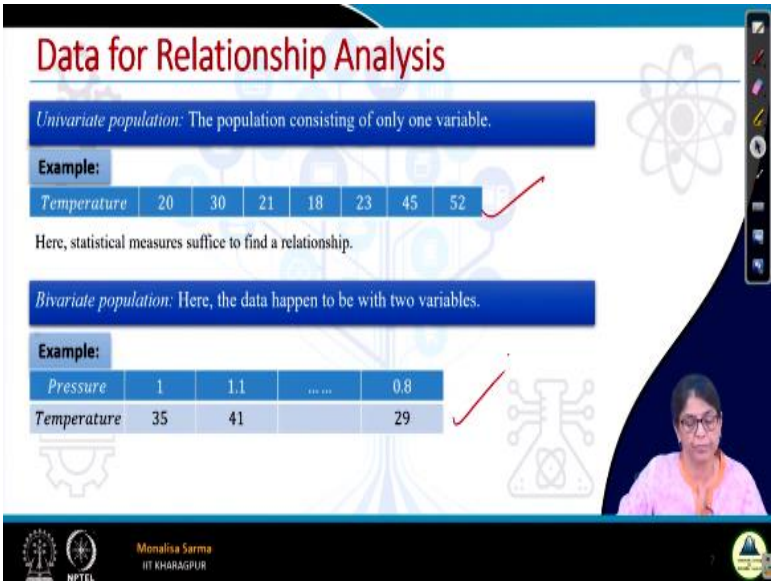
So, now nonparametric test, nonparametric test is does not come under any assumption and means parent population can be anything we do not have to assume anything about parent populations and it assumes it works with nominal data ordinal data interval scale data ratio data everything. And but the con the con of nonparametric is that it needs the entire population or a sample of a very large size.

So, now, the question comes is that at least is when I started with relational analysis, I wanted to talk about relational analysis why suddenly I started talking about parametric nonparametric and all this stuff? Basically to highlight the fact that till now, what we have discussed the different methods that have discussed is totally related to nonparametric tests all the methods we have for statistical inferences, what we have discussed all our nonparametric types of tests.

Now, from now onwards like when I will be discussing today, correlation analysis, there are some methods are parametric some methods are nonparametric. So, that is why I thought of bringing this into the picture. Now, when I will be discussing a parametric method to let you know that this is parametric and when this was nonparametric, I will let you know this is nonparametric. Anyway, I do not need to say also it will be obvious you will be able to find out okay this is a parameter method this is a non parameter method you also will be able to say.

When because you know what is the requirement for the parametric method what are the requirements for nonparametric method mainly from the data perspective when we talk of nonparametric method we can assume nominal data ordinal data any type of data.

(Refer Slide Time: 12:58)

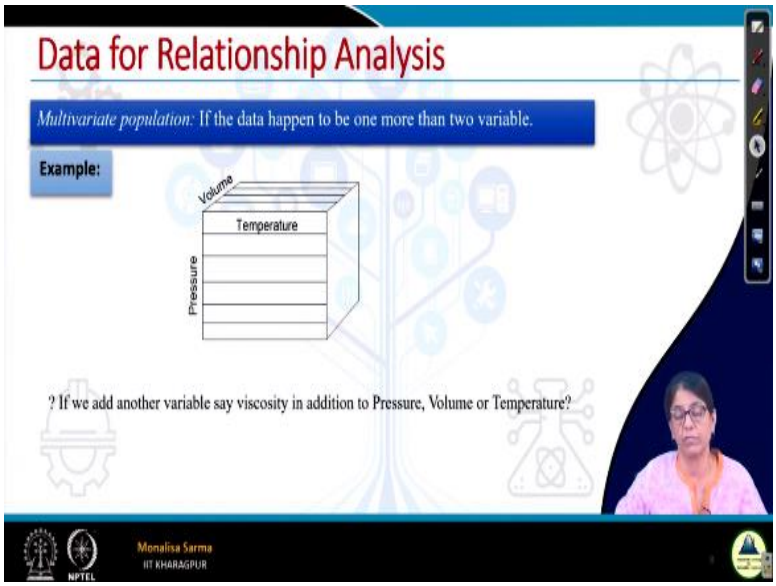


So, now, come to again coming back to the relationship analysis when we talk about relations, usually we talk of relations between 2 or more variables, we do not talk about relations with oneself it a single variable, suppose any populations there is only one parameter of interest, then we do not have any relations, there is only one parameter of interest, we are interested on only one term then maybe other parameter but we are interested in only one parameter then we do not talk a relation in such case.

Example, suppose you take the first example where it is what it is given here. So, this example, so, maybe this we can see that maybe these are the maximum temperature that is recorded in the 7 days in a week and the maximum recorded temperature recorded at MV. So, this is just only one variable. So, we are not interested in relation of such data and this is such thing, data is a call such population are called univariate populations.

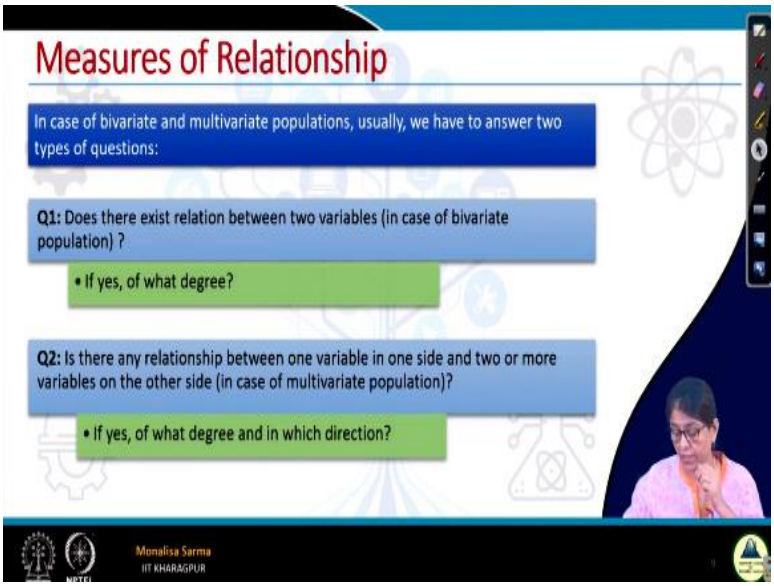
So, then bivariate population may be in a system, consider a system we are interested in finding a pressure and we are interested in data pressure data and temperature data. So then it is called a bivariate population.

(Refer Slide Time: 14:09)



Now, in for the same system, suppose we are interested in the volume as well. And suppose we add one more test is viscosity, then what happens it becomes a multivariate population. So when we talk about relation analyses, we talk about bivariate or multivariate, we do not talk about relation unless it is in case of univariate population.

(Refer Slide Time: 14:27)

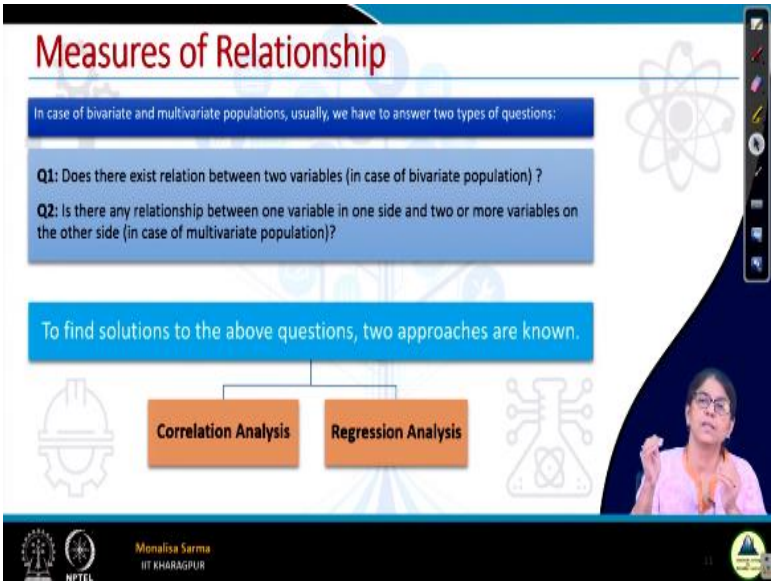


So, now, coming to this relationship, this relationship it seems very much like a categorical variable very much a qualitative variables. Now this qualitative variable, but really, is it a qualitative variable or we can measure it, if we can measure it, how we can measure it? So there are some questions which might come to our mind; like does there exists relationship between 2

variables in case of bivariate population, if yes, then of one degree, then when I talk about one degree that there comes some quantitative measure.

Similarly, is there any relationship between one variable in one side or 2 or more variables in the other side that is maybe I am talking about multivariate populations. So, if yes then what degree and in which directions.

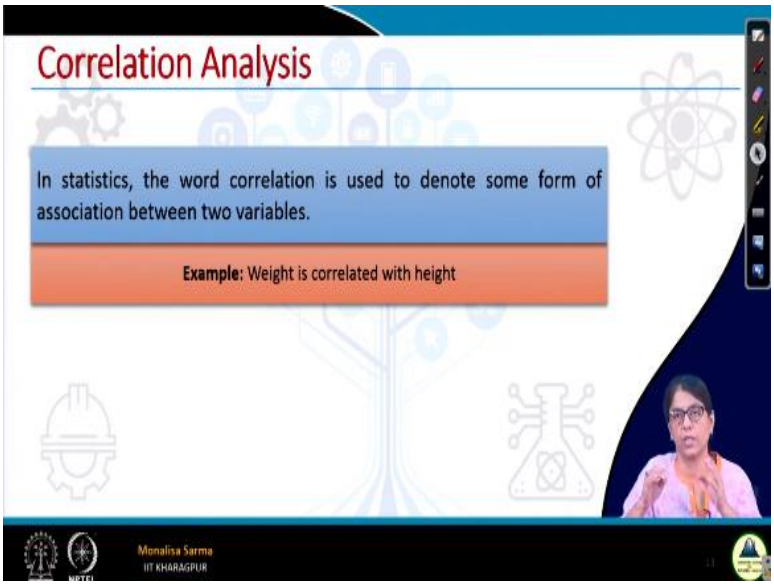
(Refer Slide Time: 15:12)



So, these are the problems we need a solution for this how we measure the relationship first is general relationship and then how do we measure this relationship. So, to find the solutions to there, the 2 approaches are known basically one is correlation analysis, one is regression analysis, correlation analysis and regression analysis both objectives are quite different. I will not talk regression at this stage I will talk and after 2 3 lectures maybe.

So, now, when I am interested in finding out the strength of the relationship, whether there is any correlation ship between the 2 variables and what is the strength of the correlation what is the strength of association is there any association between 2 relationship is any relationship between 2 variables and what is the strength of this association that is basically we will be learning that is the correlation analysis regression analysis is different.

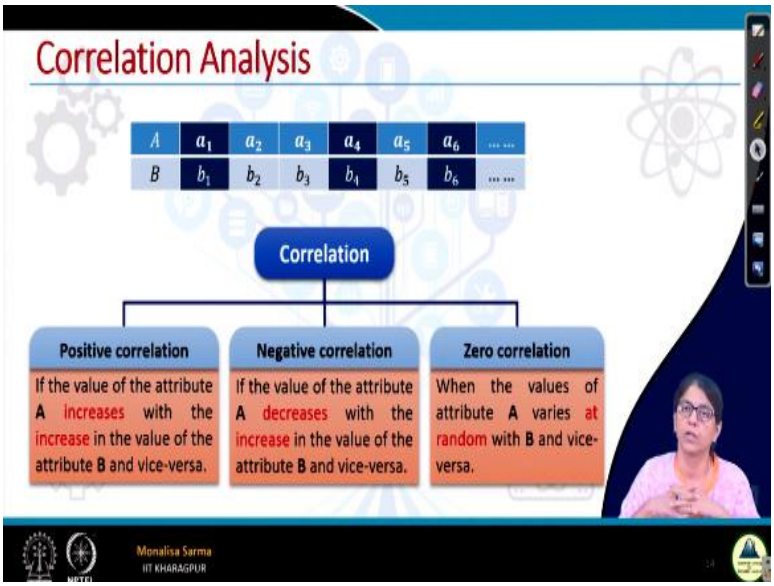
(Refer Slide Time: 16:08)



So, there in correlation in statistics the work correlation is used to denote some form of association between 2 variables, maybe weight is correlated with height, when height increases weight also increases maybe, or maybe absence from the class and score marks score in the test. So, when absence increases score decreases, the earlier the first example whatever even when height increases weight also increases maybe that is a positively positive correlation.

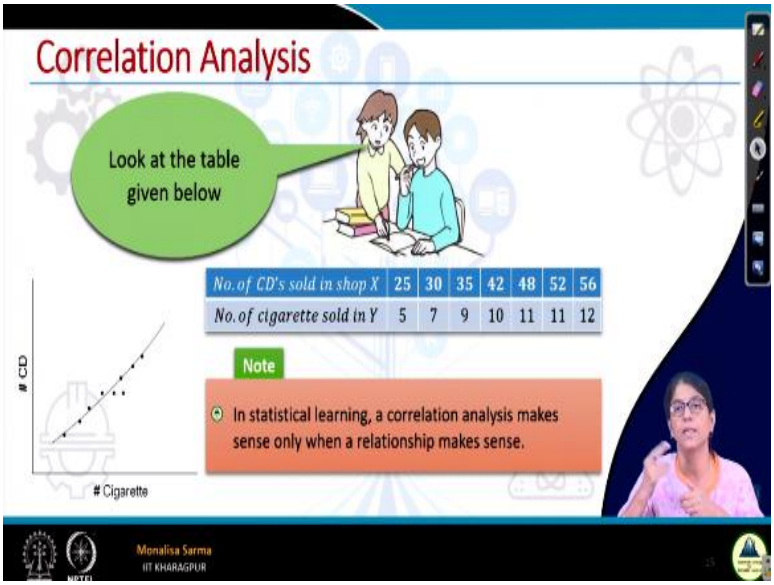
Based on when the next one what have told when I mean if you are absent from the class your test score decreases. So, one is increasing the other is decreasing this may be a negative correlation again maybe there are some factors where there is no correlation at all. So, we call it at zero correlations.

(Refer Slide Time: 16:54)



So, this is positive correlations if the value of the attribute A increases with the increase in the value of the attribute B and vice versa. Similarly, negative the value of A decreases with the increase of B and vice versa that is negative one is increasing other is decreasing other is decreasing one is next is increasing does this negative correlations and when the values of attribute A varies at random with B and vice versa, what is it? A is varying randomly it is not at all dependent on B when that is zero correlation.

(Refer Slide Time: 17:26)



So, now, see here one example is given like we can see there are 2 columns basically I mean sorry, 2 rows one is showing the number of CDs sold in one shop X another is number of cigarettes sold in another shop Y. So, if we just see this data, one is talking about the CD selling

of CD, another is talking about selling up cigarettes. So, we tried to find out the correlations and we found some correlation is there and we found this is how what is the strength of the correlation that also we found? But does it carry any meaning?

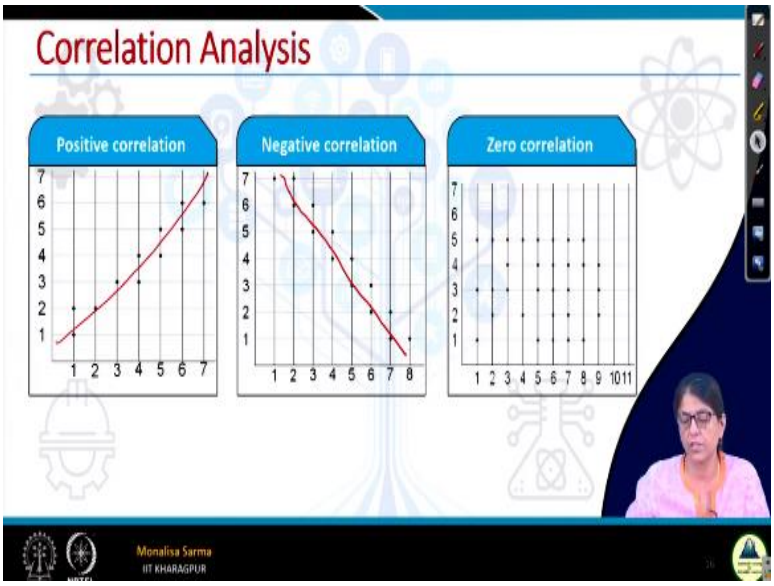
So, do we really need to find a correlation of distance. Now, there's the in statistical learning a correlation analysis makes sense when a relationship makes sense actually. So, just like that, we do not go on finding the correlation unless the relationship so make sense, then we go and find a correlation analysis. Now when I see this CD's sold in one shop and cigarette sold in others, maybe in different places that carries no relations actually, that carries no meaning no sense.

But if at the same time if we see it in this way, if we watch lots of CD, that means you may smoke also more so if buying the CD and cigarette in a similar in a nearby vicinity, then maybe this has a correlation, so you are buying lots of CD means you will be spending too much time in front of your TV that means you may smoke also more so that way it has a relation. Now, one thing very important to note here is that when we were working on correlations analysis do not get caught confused with correlation with causality.

Causality means is cause and effect went because of something because of X Y is the effect cause because of because X is happening that is why Y is happening there is a direct cause and effect cause and effect is a different association altogether is a different association, it is called causality. So, we are not looking on this cause and effect analysis, we are just looking on this the correlation analysis cause and effect is just because of this; this is happening.

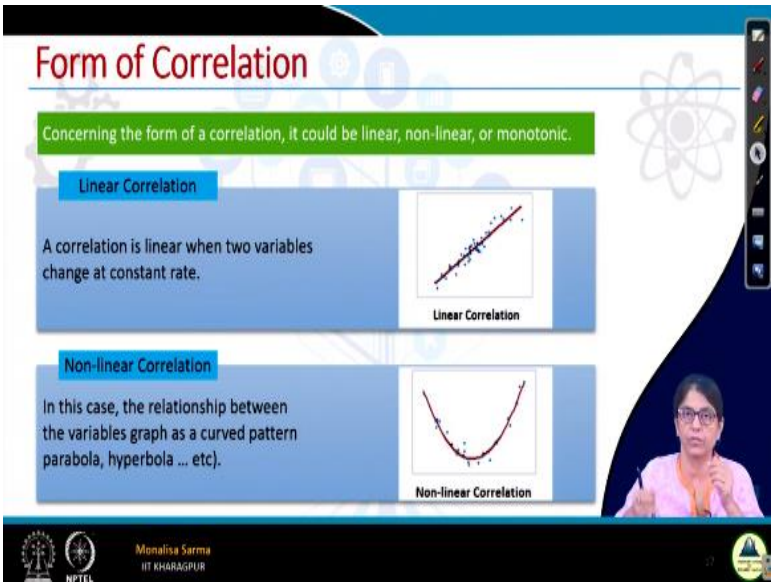
Now because of CD cigarette sell, because that definitely this is not the cause and effect of this causality analyses a different design of excellent techniques. We'll have to look into do that along with discussing in this lecture.

(Refer Slide Time: 19:45)



So now we can see the example of positive correlation here. So you see positive correlation. If I just draw a line here, then you will see the increase in one increases or the other So, similarly, this is negative correlation increasing one is decreases the other value if I take x and y coordinates, so x is one variable y as another variable okay. So, if y increases x decreases; if so, this is a negative correlation. Similarly, if this is the third figure there is no correlation at all the randomly it is ascending.

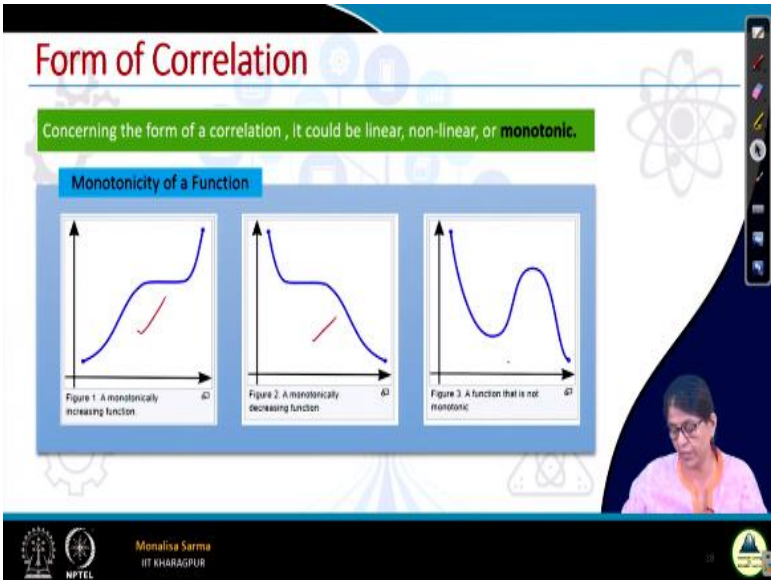
(Refer Slide Time: 20:24)



So, concerning the form of the correlations there can be different type of correlation it can be linear colinear or monotonic. So, if the relationship is linear between 2 variables then basically we can graph a straight line we can graph a straight line through the points then we call it as a

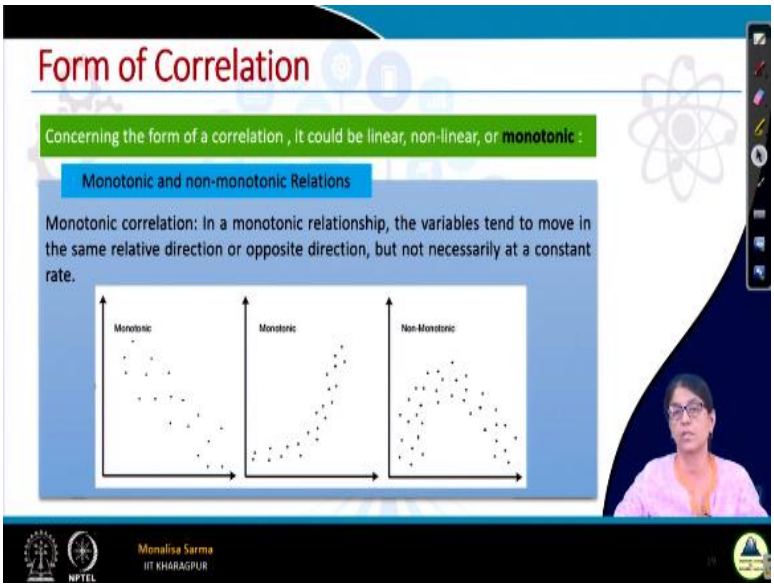
linear when it is when the 2 variables change at a constant rate, then nonlinear correlations are less obvious and there is in the relationship between the variables it graphs as a cloud pattern like a parabola hyperbola etc. If we plot the dot it will graph as a curve pattern so, that is a nonlinear correlation.

(Refer Slide Time: 21:09)



And we have one more form of correlation and that is monotonic, monotonic also so, this is a monotonic function first one is a monotonically increasing this is monotonically decreasing and this is a non monotonic.

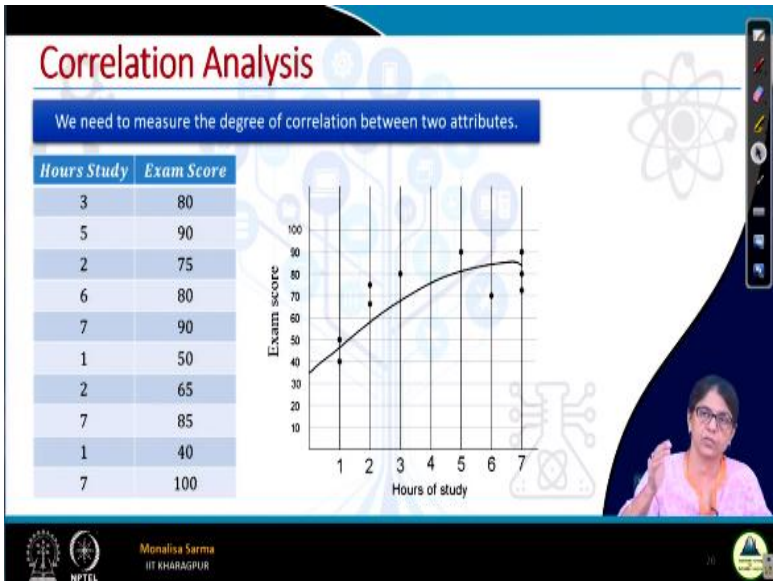
(Refer Slide Time: 21:20)



So, what is monotonic basically monotonic relationship the variables tend to move in the same relative direction or opposite direction like to linear relation which tend to move in the same direction or opposite direction similarly, in monotonic also same direction or opposite direction, but not necessarily at a constant rate and linear it moves at a constant rate either increasing or decreasing moves in a constant, but in case of monotonic it is not necessarily it moves in a constant rate.

So, a monotonic relation can be a really linear relationship, but a linear relationship will not be a monotonic relationship.

(Refer Slide Time: 21:52)

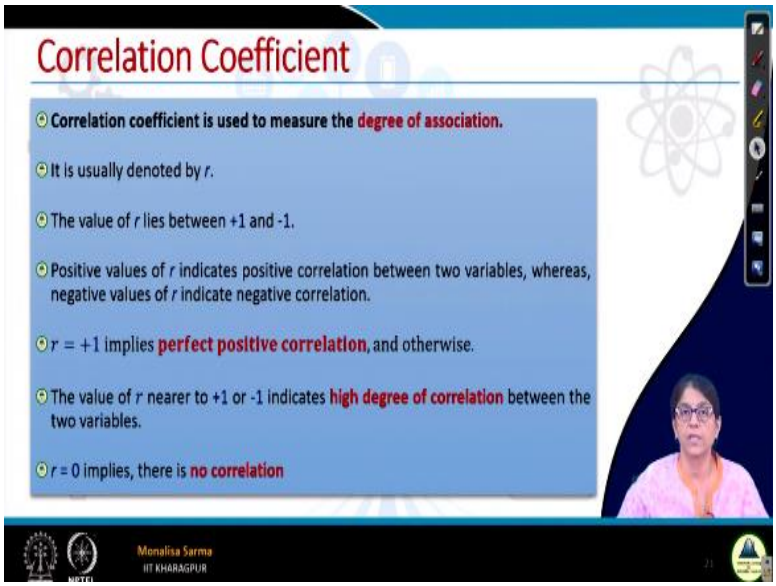


So, now, suppose we want to measure the relationship association degree of correlation between 2 attributes. Suppose the example given we have a data for few students, where students has put how many hours of study and then what is the exam score, so, is there any relationship between the hours of study they have put as well as the score they have received after putting this much hours of study. So, first, if we draw the graph first if we draw a scatterplot for this, for the data given we see that it is in positive correlations.

As one increases the other increases when we are putting more number of hours of study the other score is also increasing, because of course, all the points are not in the straight line, but

definitely it is very near to the straight line some underline and some are just quite near to the straight line.

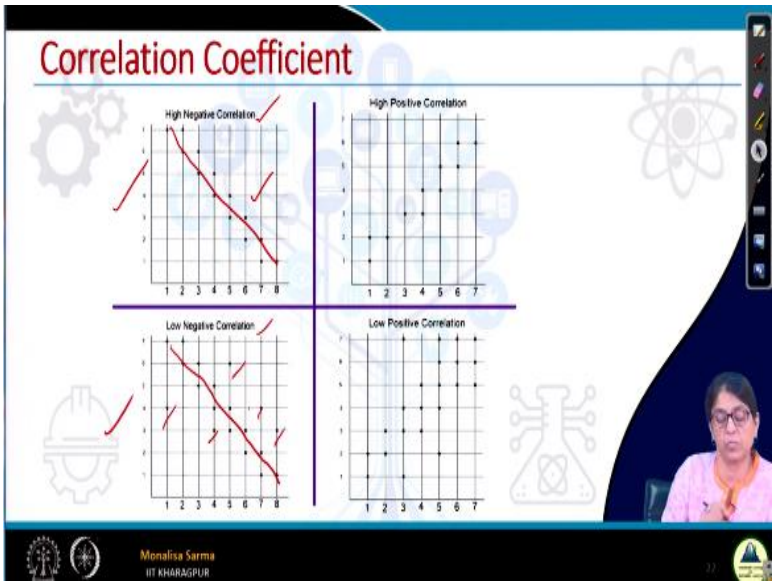
(Refer Slide Time: 22:43)



So, now, how to find the strength of this correlation, how much this association is a strong association or it is a very weak association or there is no association at all. So, correlation coefficient is used to measure the degree of association. So, to find a degree of association we what we use is a? We use this correlation coefficient. So it is usually denoted by r correlation coefficient is denoted by r. So value of r lies between -1 to +1 so, if it is +1 then it is very highly correlated or near to +1 it is very highly correlated, 0 it is not there is no correlation.

And -1 also it is very highly correlated, but it is negatively correlated one increases the other decreases. So, the r = +1 implies perfect positive correlations. And the values of r near to +1 and -1 indicates high degree of correlation between the 2 variables, and r = 0 implies no correlation.

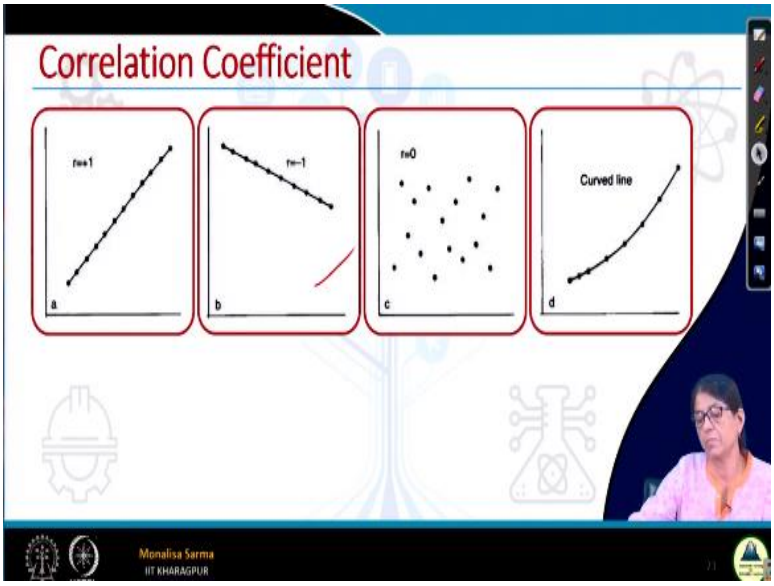
(Refer Slide Time: 23:42)



So now you see in this figure, the first one, this one, you see this figure and this figure, both have negative correlation. If I draw a line here, see, line here that runs basically closer to the points here also, I am drawing a line basically, which is closer to the point here, I am saying it is high negative correlation here, I am saying this low negative correlations, because you see the points here and discuss the points are dispersed. I am getting one point here, here, here, the points are dispersed.

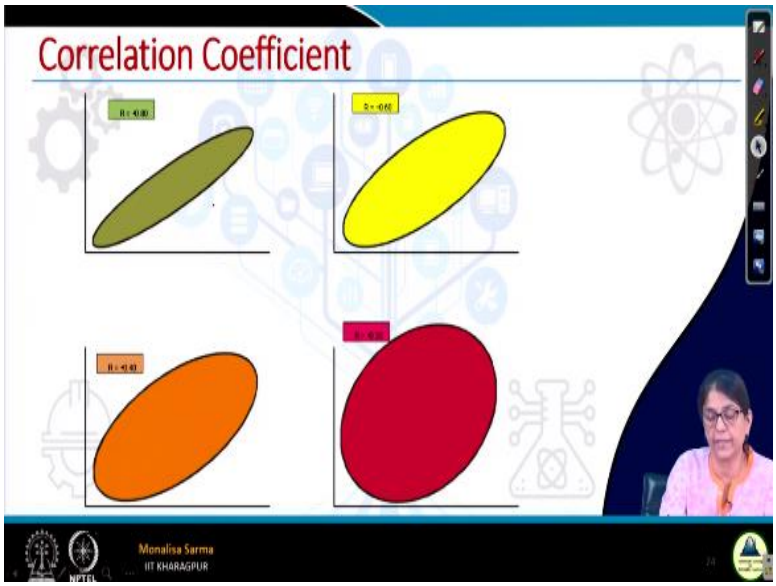
It is the points are not all the points are not near to the straight line that I have drawn. But here you see all the points are nearer. So when I am trying to find the association, maybe I can maybe I am graphing a straight line to it because it is so the linear changes, but here the points are not very not very much disperse your points are very much dispersed. So it is a low negative correlation. Similarly for this example, high positive correlation low positive correlations.

(Refer Slide Time: 24:35)



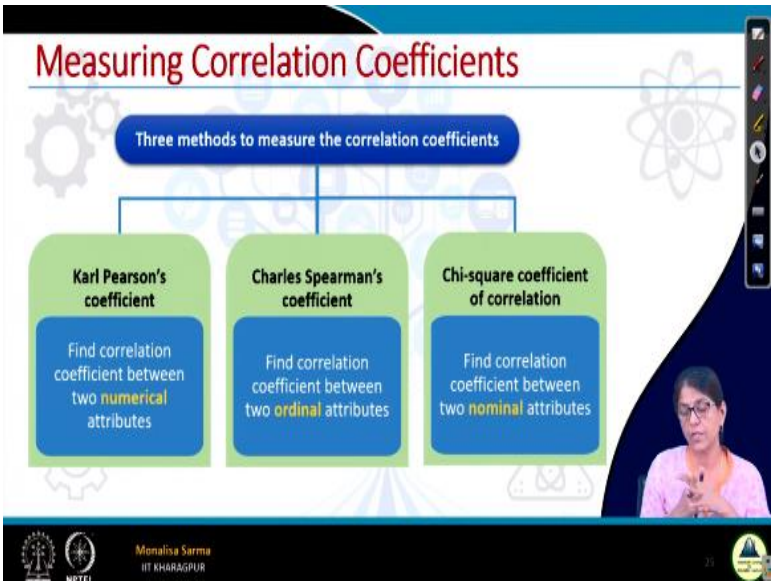
So, r is equals to + 1 something like that is also - 1 something like this figure. Then r is goes to 0. And this is a problem this may be monotonic fine.

(Refer Slide Time: 24:50)



So, you see, have a figure for that. It says this figure we are I got this was 0.80 where the dispersion is there but not very high dispersion this person means for variance. But here in our last point we do a very high dispersion. Similarly is equals to 0.60 better, dispersion is greater than this but less than this is 0.30 again, this person more than more than 0.40 as well.

(Refer Slide Time: 25:20)



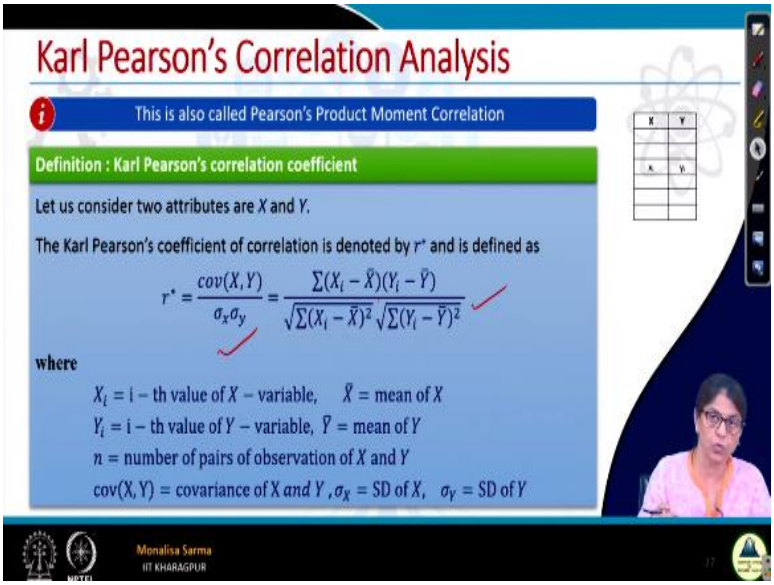
So, now, there are 3 different methods of there are not 3 different methods, there are many methods to measure the correlation coefficients, we will be discussing these 3 methods in this lecture not in this lecture in this course basically, today I will just discuss only Pearson's coefficients so, there we will be discussing three that is Pearson coefficients, Charles spearman's coefficient and chi square coefficient of correlation.

So, Pearson coefficient it is used for numerical attributes Spearman we use for ordinal attributes remember ordinal attributes means data's which are grouped into categories and which can be ranked. There is an intrinsic ordering among the categories and chi square provision its use for nominal attributes, there is no ranking among the data.

(Refer Slide Time: 26:04)

So, in this lecture today, I will just discuss Pearson correlation analysis. So, Pearson correlation analysis it is as it is already mentioned here, that is between numerical attributes, we find try to try to find a correlation coefficient between numerical attributes. So, Pearson correlation coefficient hold is an analysis it is used to find is there any linear association is there any linear association between 2 variables.

(Refer Slide Time: 26:30)



So, the data is it a data is numeric data then we use Pearson correlation analysis. So, now, how do we find out? How do we know can we use Pearson first when the set of data is given, we will try to first plot it maybe the scatterplot from the scatterplot. I will just get an idea. It is SOS is linear, then maybe I can use Pearson correlation analysis. Pearson correlation coefficient, as I

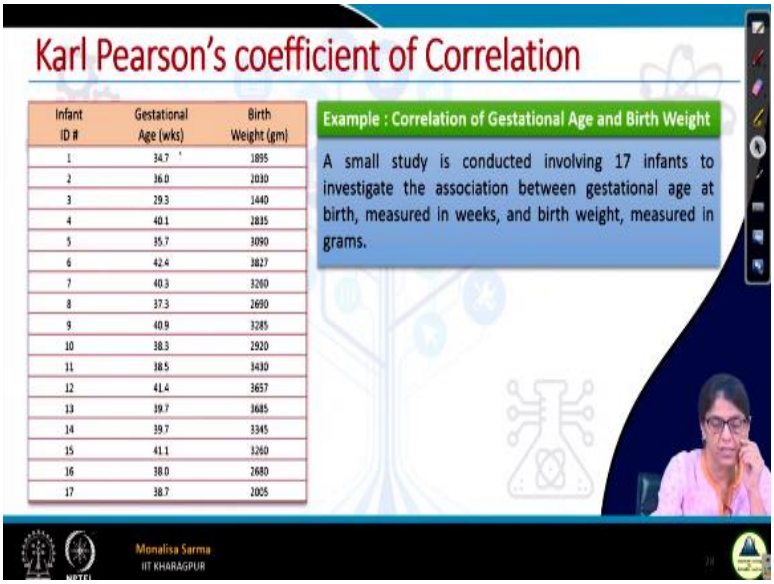
told you, it tries to measure the correlation coefficient measure try to measure the strength of the correlation for 2 numeric variables.

And if the relationship is that linear, if there is a linear relationship maybe positive, maybe negative, but a relationship is linear than Pearson's correlation coefficient is applicable. So what is the formula for that we try and the formula for Karl Pearson's coefficient is this. This is the formula covariance of X and Y divided by standard deviation of x and standard deviation of y. Now, when I find out the covariance of X and Y my degrees when I will find a covariance of X and Y, what is the formula for covariance of X and Y?

This is the formula ∑ x i - x bar x y - y bar / degrees of freedom what will be my degrees of freedom total data and - 1 total data that is the n - 1. So, similarly, when I do σ x and σ y, so, σ x and σ y here there will be √ n - 1 n - 1 so, there will be one n - 1 to n - 1 n - 1 get cut. So, this is the formula for Pearson correlation coefficient essentially what I need to know X i is the ith value of this there are 2 variables X variable and Y variable.

So, different X i is the ith variable where i value goes from 1 to n and X bar is the mean of the all the X variables Y bar is the mean of the Y variables. So, putting these values together we will be able to find the correlation coefficient.

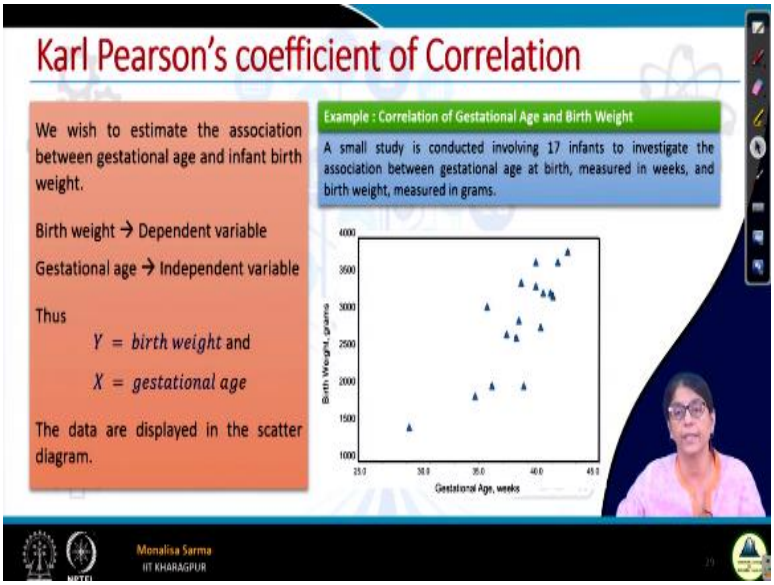
(Refer Slide Time: 28:19)



Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

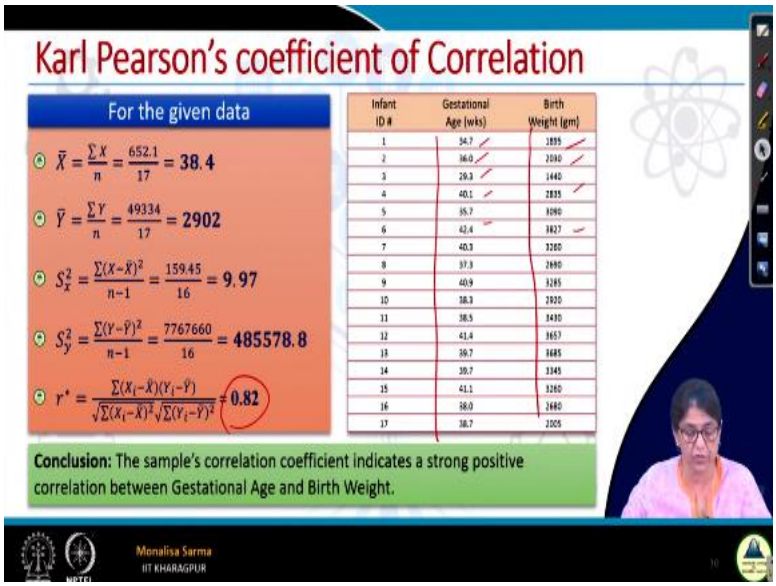
So, here given an example suppose we want to conduct a small study involving 17 infants to investigate is there an association between the gestational age at birth measured in weeks and the birth weight. So, we want to find is any association between gestational age and what is the weight of the baby at the birth is there any association between that if we want to find out the ones and these are the values given so, this these are the values of gestational age which is given in weeks how many weeks and this is the birth weight is in grams.

(Refer Slide Time: 29:02)



So, we have done first we have done is scatterplot this is the one we got. So, we could see there is a linear association. So, let us see not that we will use Pearson correlation coefficient to find out what is the strength of this association.

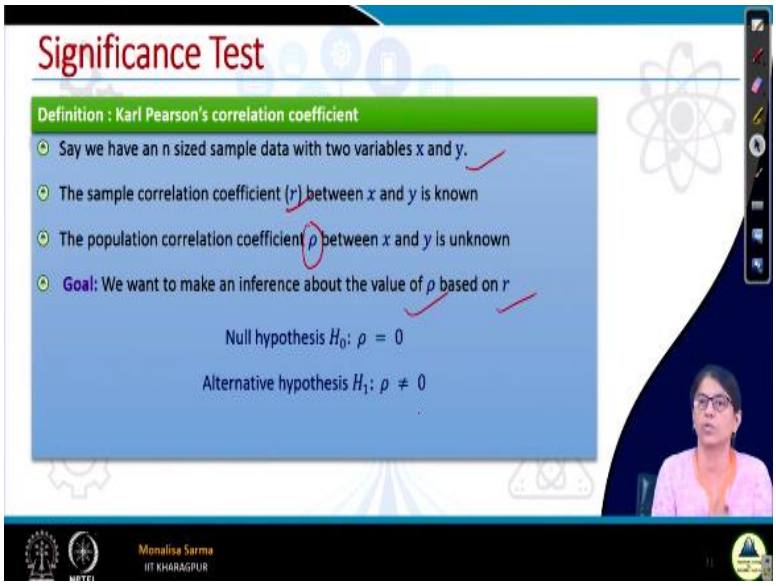
(Refer Slide Time: 29:17)



So, from this whatever values we will require X bar will require Y bar then for all X we will have to find out the this is simple just simple use of the formula just formula putting nothing else the values are given these are all X i values these are X 1 X 2 X 3 or X i values for till 17 these are all Y i values Y 1 Y 2 Y 3 and now we will have to find a mean of these we will have to find a mean of this mean of this will be X bar mean of this will be Y bar, just put it in a formula.

And at last what we got we got a value of 0.82. 0.82 is that means it is always highly correlated that means gestational age and birth weight is highly correlated. So, the sample correlation coefficient indicates a strong positive correlation between gestational age and birth weight because 0.84 is it is very much close to 1 it is it is a strong correlation strong positive correlation.

(Refer Slide Time: 30:14)



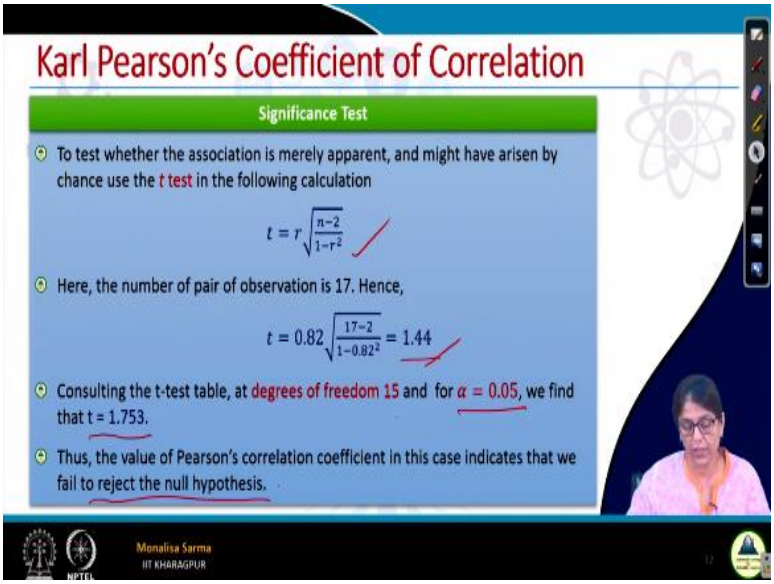
Now, the thing is that sometimes what happens because this we are doing bring it from a sample whatever it is, we are doing it from a sample we are not doing it for the whole data. So, when we are doing it from a sample maybe it happens that the sample that we have collected is by chance this coordination is shown here we got a very high positive correlation the sample that we have collected by chance by in simple Mayor by some coincidence, maybe we got this correlation, a high positive correlation.

Maybe it is not reflective of the actual populations, that is what we have done in hypothesis testing when we have tried to enforce those statistical inference similarly, for correlation analysis also we will have to find a significant stays will have to form the hypothesis will have to have the significance level same method everything same. So, here what is say we have n sized sample data with 2 variables x and y.

The sample correlation coefficient between x and y is r the sample correlation coefficient r that is this last value, what we got this 0.82 this is the r value from the sample we got. So, the population correlation coefficient is ρ between x and y is unknown population correlation coefficient is not known to us we know sample correlation coefficient. So, we want to make an inference about the value of ρ based on r like when we are doing hypothesis testing different way.

We wanted to infer the population mean from sample mean population variance from sample variance. Similarly, here we want to make an inference of the value of ρ based on r. So, my null hypothesis is ρ = 0 there is no correlation, alternate hypothesis is ρ ≠ 0, there is some correlation. There is a significant correlation.

(Refer Slide Time: 32:05)



So, how do I do that to significant test is basically, in this case, we will do use the t test whatever t test formula we have used while doing statistical inference for mean of the populations. So, here also we are using t test but the representation of this t whatever representation we are using we are using a different representation, but actually we can from that we can get this from this we can get that. So, when it is not you do not take this why there we have used a different t when we use the t test we have used a different expression here we have using a different formula.

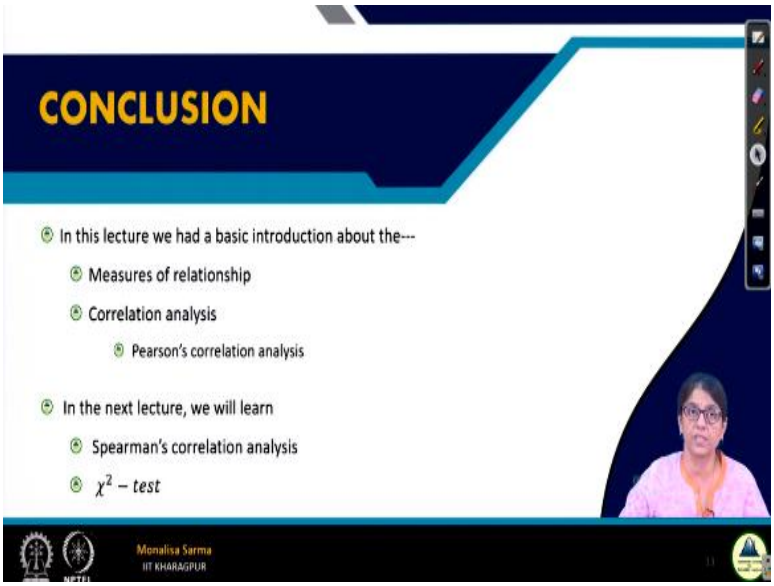
No, it is just a different representation. So, here we will be using this representation. What is that? Because here, we; need the t in terms of r that is the correlation coefficient. And there we need a t in terms of mean and standard deviation, the standard deviation of the sample. So, this is our formula for t. So, we will put everything here we know what is r what we have calculated n is the total sample size. So putting here so what we got we got a value of 1.44, suppose we have as we made significant level of 0.05.

So, for a significance level of point 05 degrees of freedom is 15 and - 2 because we are considering to between 2 variables within 2 sets of data each has n data, so n - 1 - 1. So, it is n - 2. So, for degrees of freedom 15 and alpha equals to 0.05 if we see the t table, t table, we have already seen lots of t table I think you already know that if we say the t table will get the value t = 1.753. That means our critical reason starts from 1.753 and a value greater than 1.753 it falls in a critical region it falls in rejection region.

So if it falls in the rejection region, and what happens then we reject the null hypothesis, is not it? We reject the null hypothesis and we accept the alternate hypothesis. Now, what we got we got a value of 1.44, 1.44 is very much in the acceptance region. So does the value of Pearson's coefficient in this case indicate that we fail to reject the null hypothesis. See, here, we got a strong correlation coefficient, but actually in the population there is no correlation.

So that the data that we have got is totally purely by chance, so in this indicates that we fail to reject the null hypothesis and what is the null hypothesis? Null hypothesis is that there is no relation between the 2.

(Refer Slide Time: 34:37)



So, we end this lecture here. So in this lecture, what we have learned? We will learned how to measure the relationship and we have also seen Pearson's correlation analysis. And in my next lecture I will discuss Spearman's correlation analysis and chi square test.

(Refer Slide Time: 34:53)

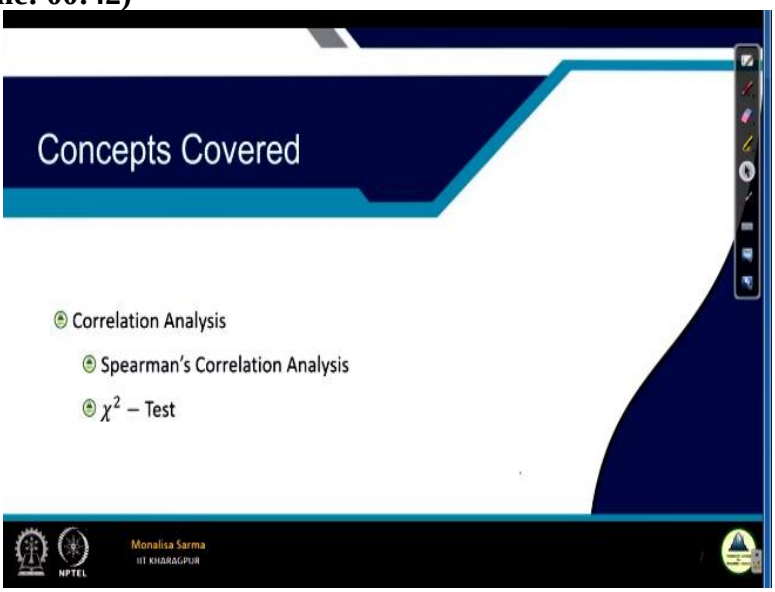
So this is the reference. Thank you.

Statistical Learning for Reliability Analysis
Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 39
Correlation Analysis (Part – II)

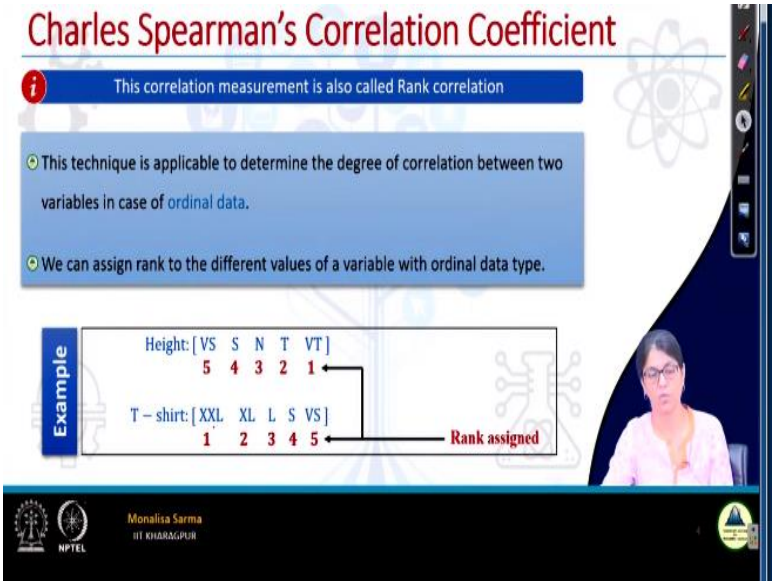
Hello guys. So, in continuation of our discussion on relation analysis, we will be the last class basically we have discuss Pearson correlation analysis and we will start a class from that point onwards.

(Refer Slide Time: 00:42)



So, in this class we will be discussing Spearman correlation analysis and chi square test, these are the 2 different correlation analysis where Spearman correlation analysis we have already seen in the last slide that we use for ordinal data and chi square test we use for nominal data. The Spearman correlation analysis is also called rank correlation coefficient, because why because Spearman correlation less is as I told you, it is applicable for ordinal data ordinal data means there is a ranking among the data. So, that is why it is also called rank correlation coefficient.

(Refer Slide Time: 01:16)



So, now, if we can like see this example, the height of a person and the size of the garment so, if you are interested in finding out the relation between is any association between height of a person and the size of a garment is there any association between these 2 and if it is there, what sort of association, what is the strength of the association? So, as already mentioned, here this technique is applicable to determine the degree of correlation between 2 variables in case of ordinal data.

Now height if we can consider it as a interval I mean we can consider the ratio data also now see the ratio data we can convert a ratio data to ordinal data some likewise, I want to give you one more example here likewise the size of the house. The size of the houses is very much a numeric variable that is 1200 ^{2} feet 1400 ^{2} feet 2000 ^{2} feet 3000 ^{2} very much a numeric variable. This again, we can convert it to ordinal data how.

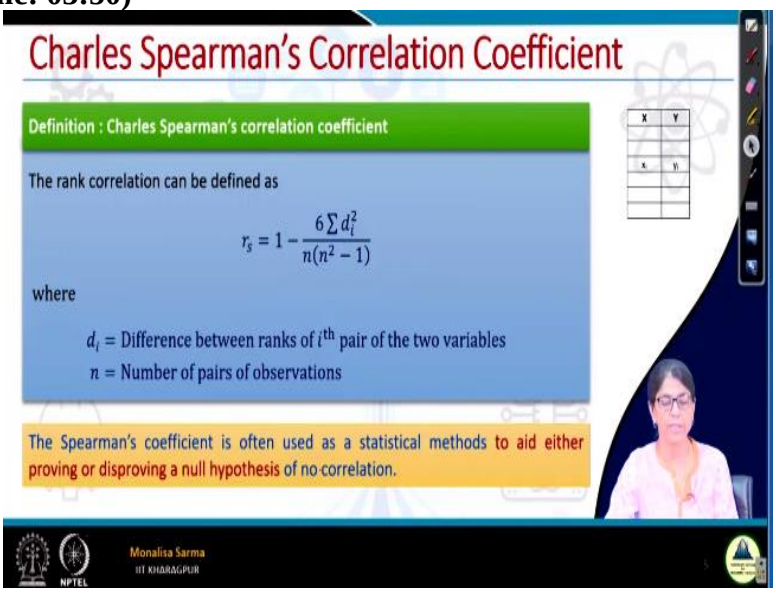
If we can convert the size into different categories, which is width the interesting ranking among them. So, that was similarly now height I have converted to an ordinal data like how this is so starting from very tall, tall, normal height, short, very short so, it is there is an ordering among this data. Similarly, this size of the garments are very small, small, large, extra large extra extra large.

So, now, when we design this way, these are the 2 data these are 2 variables height and size garment size, we want to find out the association between these 2, in such case we will be using Spearman correlation coefficient. So, here one thing to be noted here, see if when we have to assign so first we will have to assign the rank to these different data different

categories. So, for the height we have different categories very short, normal tall, very tall, we will assign the rank of each category.

Similarly, for T- shirt also we will assign the rank to each category. So, now, how this rank is assigned to the highest value we assigned a rank 1 like in a class who scores the higher we give it a rank 1. So, similar the highest value we assigned it rank 1 so here are very tall is rank 1 tall is 2 similarly normal is 3 and very smallest 5. Similarly, for garment size, XXL is the highest so we give it rank 1 and very small, we give it rank 5.

(Refer Slide Time: 03:50)



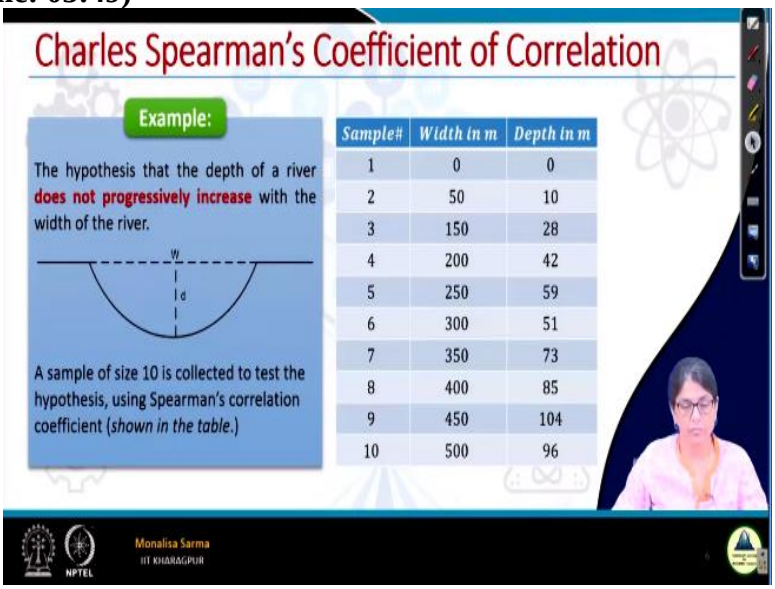
Now here to find out the correlation coefficient between 2 ordinal data type, this is a we have to find out basically the value of r s that is called as Spearman's correlation coefficient r s is a Spearman correlation coefficient. So, it has how the Spearman correlation coefficient what is that expression for this? This is the expression for this. So, what it is given we have a term d i and we have a term here n, n we know is the size of the sample and what is d i? d i is the difference between rank of the ith pair of the 2 variables.

So X has different values x 1 to x n Y has different values y 1 to y n. So d i is the difference between the ith pair difference between x i and y i we have assigned rank to each variable. So difference in rank of the ith pair difference in rank to the first pair means what is the difference in rank of y 1 and x 1 difference in rank of the second pair means difference in rank of x 2 and y 2. So d i is the difference rank between ranks of ith pair of the 2 variables. So n is equals to number of pairs of observation.

And like Pearson correlation also the Spearman coefficient is often used as a statistical method to aid either proving or disproving a null hypothesis of no relation here also we will consider a null hypothesis that there is no correlation. Now, Pearson correlation analysis what we have seen that is very much a parametric analysis. The data that we have considered has to come from a normal populations. So, that is a Spearman correlation is a parametric analysis.

Now, Spearman you see, when the data is or ordinal from that only it makes it clear that it cannot be parameterised and of course, it is not parametric here we are not bothered about the populations parents, the data has come from what sort of populations it is a very robust is we are not at all bothered about the type of the population from which where we got the data there is a nonparametric analysis.

(Refer Slide Time: 05:49)



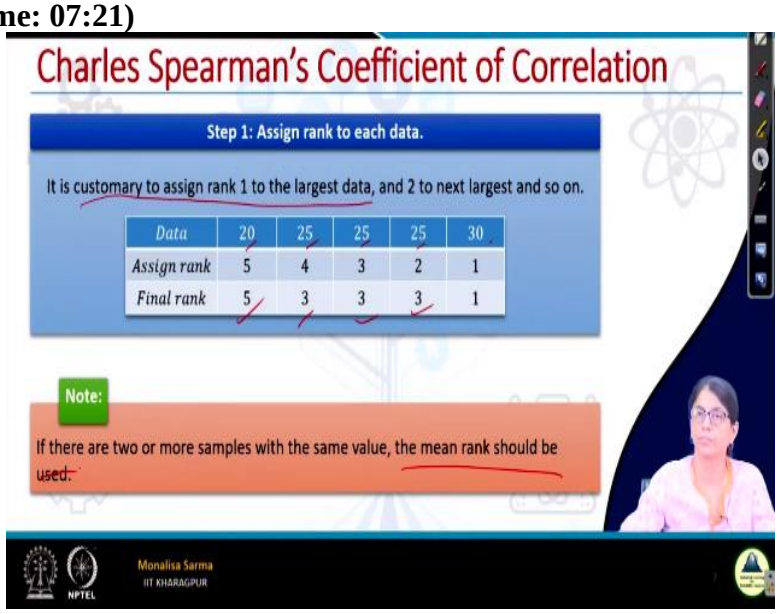
So, now, let us discuss this Spearman correlation coefficient using an example. So, what example is given here the hypothesis that a depth of a river does not progressively increase with the width of the river. So, I am hypothesizing that depth of the river does not progressively increase the width of the river means as the width increases height does not increase there is no association between the height and width of the river, it is not mandatory that mean width increases height increases, there is no association that is my hypothesis.

But I will have to prove that is it correct that means my null hypothesis is there is no correlation between height and width of the river. My alternate hypothesis is there is correlation means as width increases height increases. So, to prove this what I have taken? I

have taken a sample of size 10 to test the hypothesis using Spearman correlation coefficient so, these are the different data that we have taken.

Width in meter width in first data we have taken 0 0 height 0 width is not it is 0 actually because 0 does not carry any meaning here, how what can be 0 return 0 that basically by 0 what I meant to say is a very negligible width and very negligible depth, maybe you can consider cannel so very negligible width very negligible depth, which I am considering for simplicity is a let me consider it as 0. So these are the different width of the river, these are the different depth of the river.

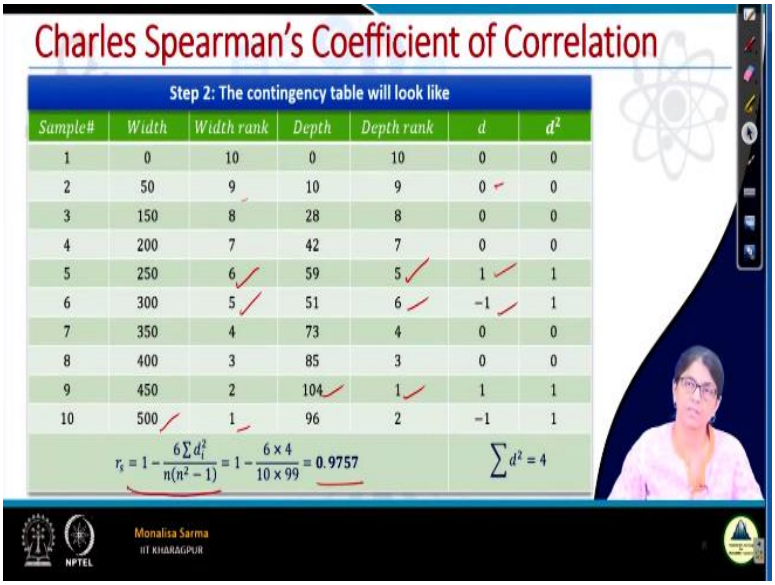
(Refer Slide Time: 07:21)



Now I will have to rank it, how will it rank? Height we will get the now not only that, there is 1 more thing to be here. So, of course, it is customary to assign rank on to the largest data that we have already discussed 2 to the next largest and so on. Now, what happens if there are same type of data like you see here the example here, so, it starts from the 30 and we have 25, 25, 25, 3 25 and then 20. So, what the same the definitely 30 will get rank 1 then what about this 3 25 it will get rank 2 all will get rank 2 or will get 1 will get 2, 3 and 4.

So what happens is that first we assigned a rank in like this first 25 I have got 2 second 25 I have got 3 third 25 I have got 4 and then we find a mean of this. So mean of this 3 then rank 3 assigned to all this data and 20 is 5, if there are 2 or more symbols with the same value mean ranks should be used.

(Refer Slide Time: 08:15)

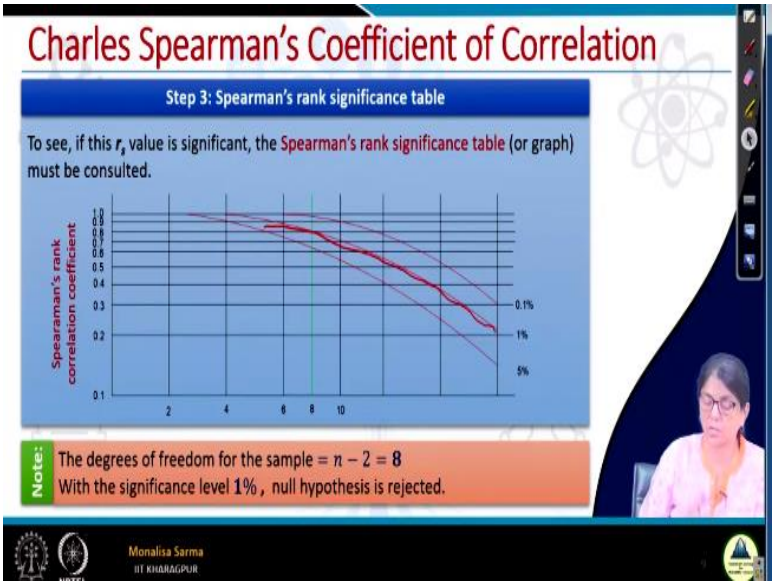


So, now for this example the height and width of the river. So, that accordingly I will rank the highest value here you see highest value is 500. So I have rank 1. Similarly, you see in the depth the highest value is 104. So 104 is rank 1. So, accordingly I have rank all the data width I have some rank depth I have given some rank, when the width is 0 depth is 0 and this total there are 10 data and this is the 10 rank. So when width 50, depth is 10 this rank is 9 and this rank is 9. So, what will be d 2? d 2 will be x 2 - y 2, so 9 - 9 is 0.

So, similarly, I found a d i value for all this data, like see here example here, for 250 it is the width rank is 6 and it is depth rank is 5. So, what will be the d i? d i will be 1, 6 - 5 = 1. Similarly, here for 300 this is 5 this is 6, so, it will be -1. So, this is my d i data I found out the d i data then I found this square then this is my formula which we have already seen. I put the value and I got this 0.9757. So, 0.9757 is my r s value.

Now this 0.9757 is it really significant value that we will have to find out. Again similarly, the way we have done significant test for the Pearson similarly here we also will do the significant test for Pearson.

(Refer Slide Time: 09:44)

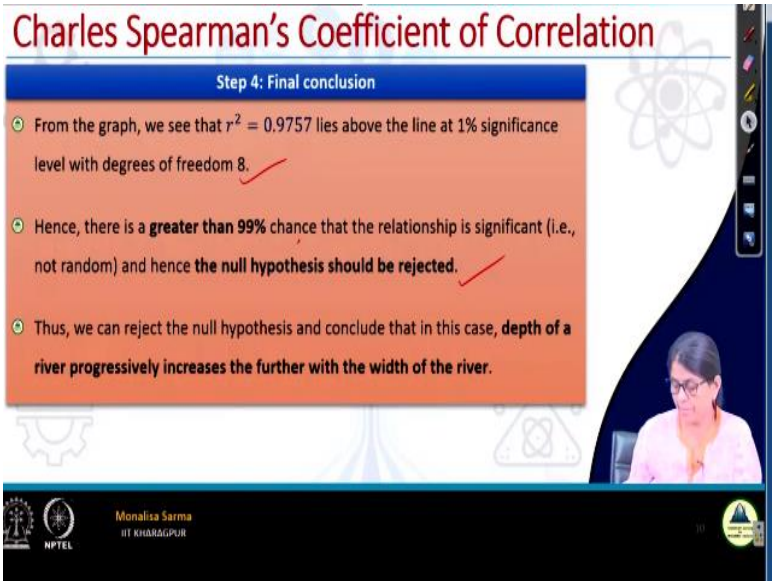


So 0.9757 likes for Spearman we have in the any standard textbook you will find Spearman and graph in this sort of a there is a Spearman graph basically and for other T table normal table we get a table of values. Spearman we get this sort of graph it will in any standard statistical book. So, here you see if now, for this example how many data's are there? There are total 10 data, 10 data means what is the degrees of freedom 10 - 2 8. So, 8 degrees of freedom.

Suppose, we are interested in finding at 1% significance level that means, 1% significance level means 0.1 you see the 1% significance level is that this can you see this is the graph this is the 1% significance level. So, it my value lies above this then it is in a critical reason if my values lies below this graph then it is an acceptance region. Now, what is my degrees of freedom? Degrees of freedom is 8. So, 8 this is my 8 and what is my value? My value is 0.9757 so, for 8 0.9757 will be somewhere here it is very much above the graph.

So, it is very much above the graph for 1% that means, it falls in a critical region it falls in a rejection region. If it falls in the rejection region that means, what my null hypothesis is rejected that means, there is what is say correlation there is an association between height and depth of the river.

(Refer Slide Time: 11:08)



So, far we see that lies above the line at 1% significance level with the degrees of freedom 8 hence, there is a greater than 99% chance that a relationship is significant that means it is not just randomly we got this data and the hence the null hypothesis should be rejected. Thus we can reject the null hypothesis and conclude that in this case depth of a river progressively increases the further with the width of the river.

So, before starting with the chi square correlation analysis. So, I want to mention one more thing here. As I told you our Pearson correlation analysis is very much a parametric method Spearman correlation analysis is very much a nonparametric method. So, when we use Pearson? Pearson we use when the data is linear. Spearman we use when the data is monotonic we have seen in the last class remember this.

When it is linear if we can graph a line monotonic we have seen it may increase in a relatively maybe increase in the same direction or in the opposite direction, but may not be in a constant rate as what we can see in linear. So, basically we may get some sort of slight curve we may get it may increase 1 increases other increases, but it may not be at a constant rate then we call it a monotonic relation, so Spearman correlation analysis we use when the relationship is monotonic.

Now, the question is how do we know that whether it is linear or monotonic? Suppose, we have plotted from given the data we have plotted the scatter plot it. Now, from the scatter plot if we can very well understand what does this scatter plot indicates it does it indicates a linear

or at negative monotonic if we are not very sure whether it is monotonic or linear, we are not very sure of it, then it is better we go for the Spearman correlation analysis.

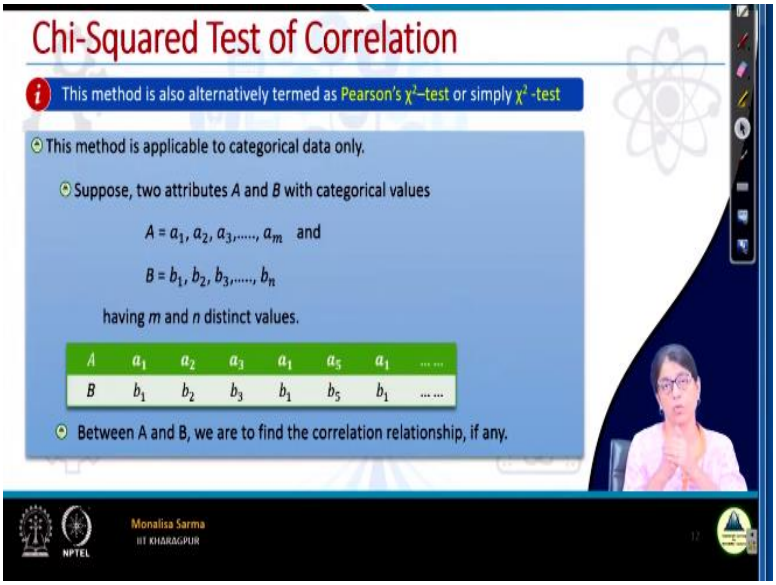
Definitely, if it is linear, then we use Pearson will give us more results, because when we are in a numeric data when we are using Spearman that means what we have to be this numeric data we have to convert it to ordinal data, when we are converting to ordinal data, we are losing some information. So, that way so, in a linear relationship, if using Spearman, then of course, we lose some information that way our results will not be very accurate, but still we will get a result which is closer to Pearson only.

But if it is not linear and we try using linear we will lose lots of information we will not get the correct information at all. So, from the graph if we see that we are not very sure it is linear or monotonic, better to go for Spearman, result may not be as precise as Spearman, but we are not taking the risk because if it is monotonic and we are using Pearson we will not get a good result. That means what if it is Spearman can be used for linear relationship also.

It is mainly used for monotonic relationship it can be used for linear relationship also, that means, when the data it very much violates the linear relationship we go for Spearman relationship, but it can also be used for linear relationship with precision becomes lesser. So now next we will discuss is the chi square correlation analysis. This is also a very much nonparametric analysis very robust way we are not considered about the distribution of the parent populations.

We do not have to bother about the variance of the populations how it is whether the variance of the different population is same or not nothing we not bothered anything. So, as I told you chi square correlation analysis we use for nominal data where we have different categories of data.

(Refer Slide Time: 14:49)



Like take an example here suppose there are 2 attributes A and B with categorical values. Let me give an example what sort of example? Suppose in a population there are different categories, we can divide the whole people in a particular population, you can divide into 3 categories one, the person who had severe COVID, one group of persons who had mild COVID and another group would not have COVID. So help if I consider this in my best variable as health status of the people of the population.

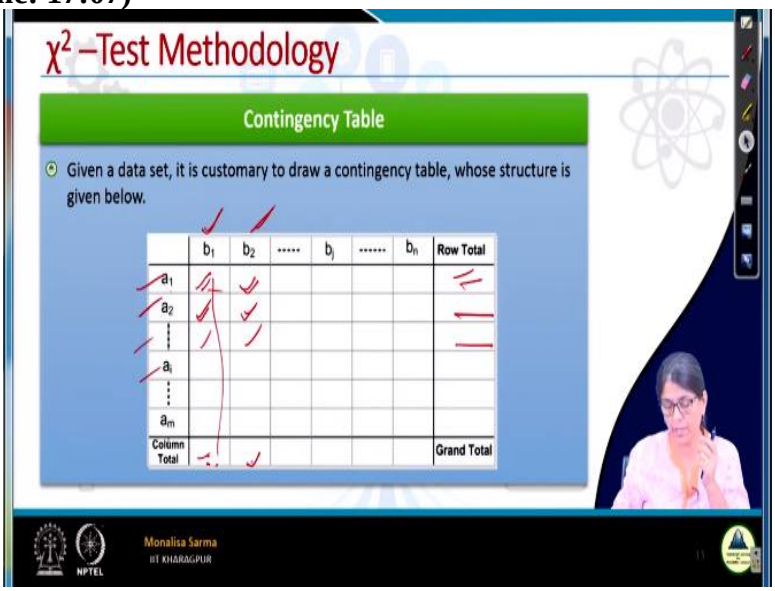
So I will categorize it into 3 different parts. So, let me say this is a 1, a 2, a 3, what is a 1? a 1 is the people who had severe COVID, a 2 is the people who had mild COVID, a 3 is the people did not have COVID. So A is the health status of the people and the populations. So this attribute I have 3 different categories. Similarly, b may be again, I am classifying this population B into 2 different categories based on the vaccination status.

So my attribute is the name maybe vaccination status, so 1 is fully vaccinated. Another maybe partially vaccinated another maybe not vaccinated at all or maybe 2 categories vaccinated, not vaccinated. So, the b 1, b 2 vaccinated, not vaccinated if A is 3 categories is by fully vaccinated, partially vaccinated, not vaccinated b 1, b 2, b 3. So, this way if we have in a population if we have 2 or more attributes, which have different categorical values, then we can use chi square test of correlation.

So, how we do that? So, basically here we need to find out the correlations between A and B. So, correlations between A and B means in the example, whatever given again, my correlation is that vaccination did not have any effect on the COVID health status of the

people. So there is no correlation vaccination did not have any health status other people this is my null hypothesis and alternative hypothesis vaccination have effect on the COVID status. So that means between A and B, I tried to find a correlation is there any correlationship among that?

(Refer Slide Time: 17:07)



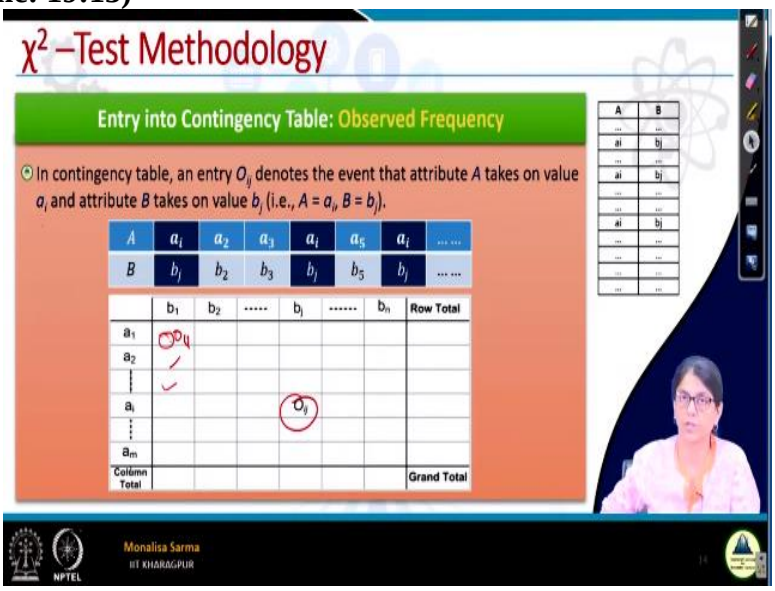
So, how do I do that first and for chi square test? I draw a contingency table. What is the contingency table? And here in the rows, I have all the A values, all the categories, all the groups of the A attribute. So what was my attribute? My A attribute maybe had severe COVID, mild COVID, had no COVID. So, these are my different groups. So this I have the different rows for that. And similarly, my columns are the vaccination status, not vaccinated, vaccinated. So b 1, b 2, maybe here we have m b 1 to b n similarly, a 1 to a m.

So, I will draw the contingency table this way that is my first step. Second step is then I will find a row total and I will find a column total. What does row total indicates? Row total indicates all the sum of all this. What does this indicate? This indicates this is suppose b 1 is vaccinated, vaccinated people how many vaccinated people got severe vaccine COVID, minor COVID and no COVID this total will give me my column total.

Now, this column total non vaccinated people had severe COVID, mild COVID, no COVID so, this will give me my column total this is also called residual column this is column residual. Similarly, I can find out row total what is row total? People who have severe COVID how many of them are vaccinated, how many of them are non vaccinated, this total as I got is row total. Row total is also called row residual.

Similarly, here I will find what to say vaccinated people having mild COVID non vaccinated people having mild COVID some of this I will get it here that is row total, similarly and get a row total here. So that is how I first draw my contingency table? I have drawn the contingency table. Now, I have to draw the contingency before filling out the row table, I will have to fill these values, I will have to fill as well as within this column, in this continuance is it.

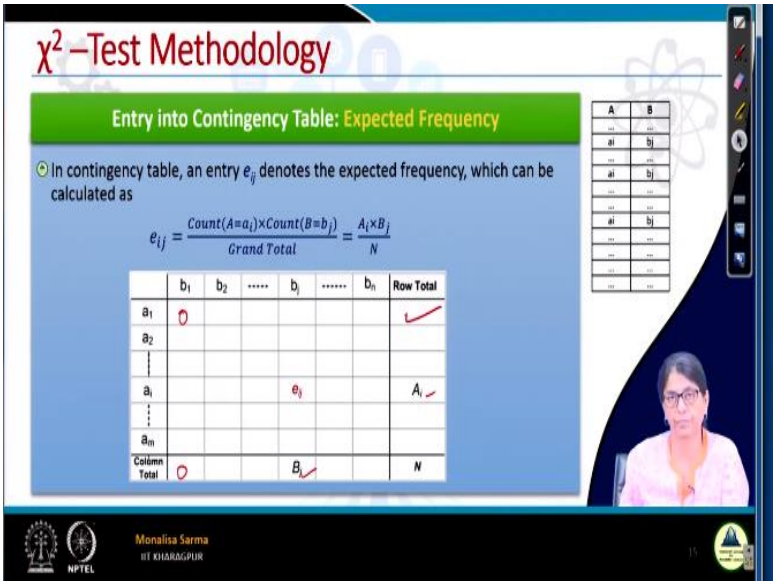
(Refer Slide Time: 19:13)



How to fill this table before going to the row total first so to fill this total filling this total is any entry O ij what does this indicate? O ij indicate this A = A i, B = B j. So, in my example, what does a 1, b 1 indicates a this is O 11 this would be what this O 11 indicates here O 11, O 11 indicates people who are vaccinated and got severe COVID so, what is this? This is O 21. What is O 21 indicates? O 21 indicates people who were vaccinated and get got mild COVID.

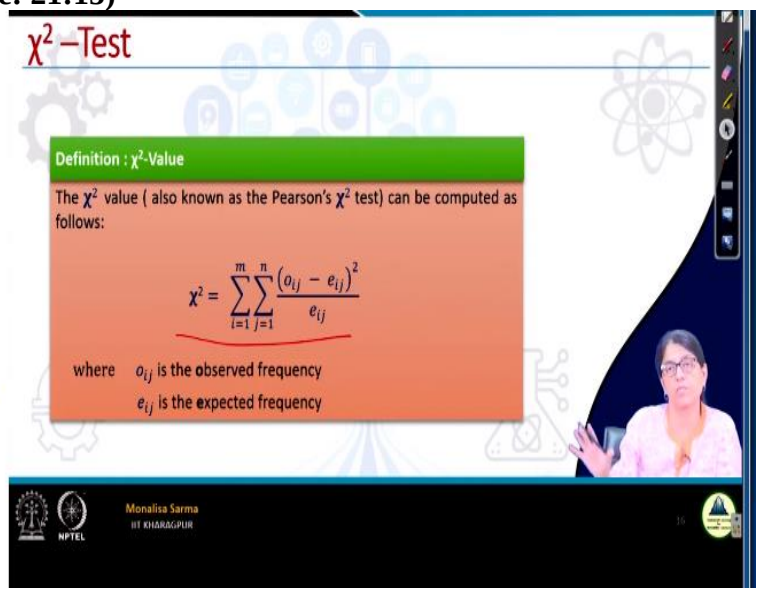
This is people who are vaccinated and got severe COVID. So, this way I fill this data. So, this is how I fill the whole table and then I find out the row total and column total. So, in their contingency table and entry O ij denotes the event that attributes A takes on value A i and attribute B takes in a value of B j.

(Refer Slide Time: 20:29)



So, now, next is after finding the O ij and after filling up the whole contents table than finding out the row residual and a column residual next step is that I will find the expected frequency. How do I find the expected frequency? Expected frequency is this basically the row residual and column residuals divided by the total number of values that is the expected frequency. So, this is A i x B j / N for each value. If I am interested in finding this, so, it will be this value what is a 1 x b 1 / N that is how I will find out the e ij? e ij is called the expected frequency.

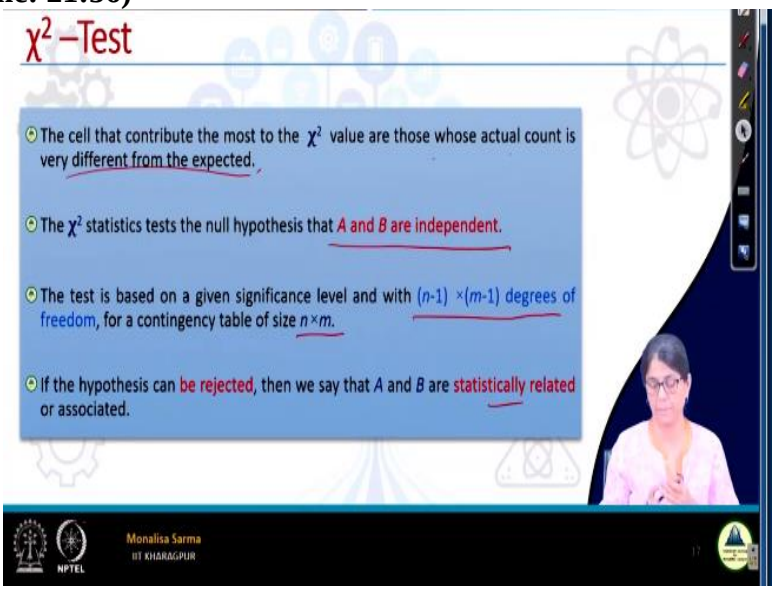
(Refer Slide Time: 21:15)



So, once I have the O ij once I can have the e ij then I can find out the chi square value. So, how is the chi square value? This is the formula for finding out the chi square value. So, for each table how do I get the chi square value? I have the O ij value for each column cell I have the e ij value for each cell. So, how do I find out a chi square value? O ij value for each O ij - this formula (O ij – e ij / e ij) ^{2} that is the chi square value for 1 cell. So, when I am interested

in finding out the chi square value of whole cell so, some of all cell value will give me my chi square.

(Refer Slide Time: 21:56)

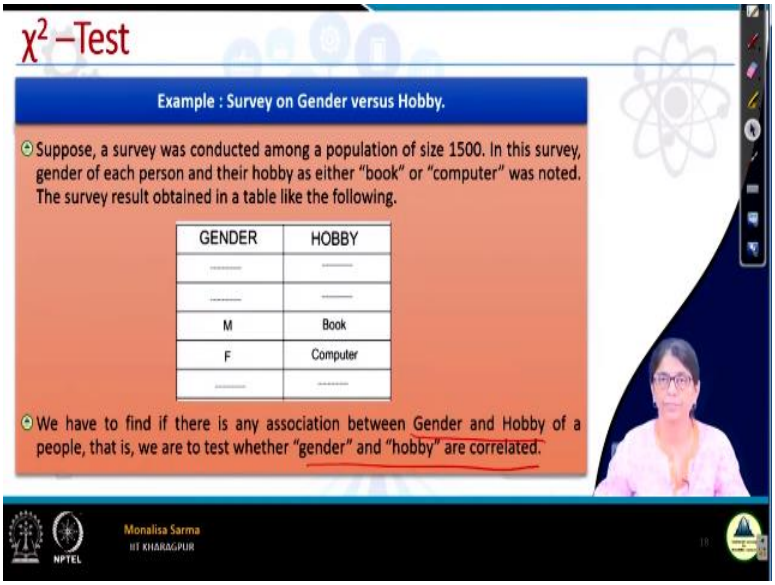


So, then again we will have to do find us what to say significance test whether the whatever value I got whatever value for chi square test I got, whether it is really significant whether the null hypothesis is accepted or null hypothesis is rejected. So, similarly, I will like in the previous method I will be doing a significance test what is the significance test? First in the cell that contributes the most to the chi square value or those whose actual count is very different from the expected when I get the chi square value for each cell.

If you take an example and if you calculate the chi square value for each cell will get different chi square value maybe same may be different for each other cell in the contingency table the cell which has the highest chi square value it is because it is actual count is very different from expected what is O ij? O ij is the actual count actual even so, and e ij is the expected value when the actual count is very much different from the expected value then we get a high chi square value.

So, chi square statistics test the null hypothesis that A and B are independent, this test is based on a given significant level with n - 1 x m - 1 degrees of freedom for a contingency table of size n cross m. If the hypothesis can be rejected, then we say that A and B are statistically related or associated. If we can reject the hypothesis, if it falls in the rejection region, then we can say A and B are statistically related they are not independent.

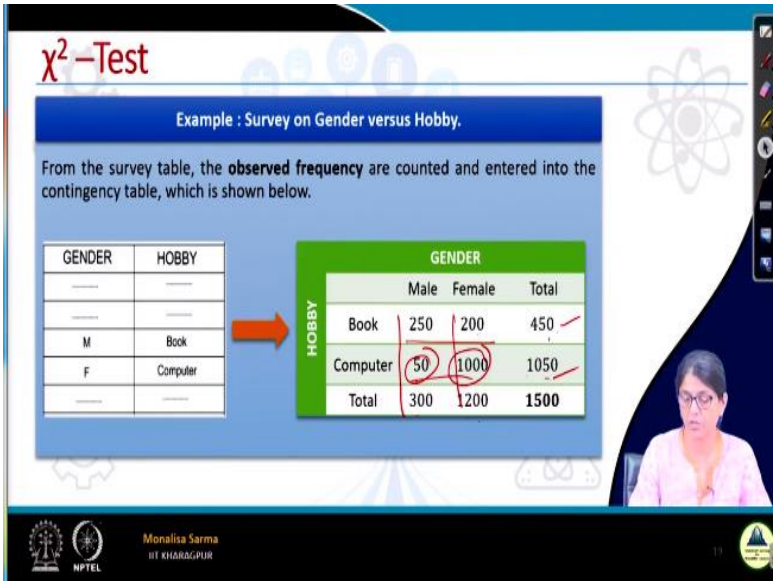
(Refer Slide Time: 23:33)



So, we will see with the help of an example a very simple example will also do some example in tutorials. So, do not worry, this is a bit critical. So, suppose a survey was conducted among a population of size 1500 total population sizes 1500 and this survey is gender of each person and their hobby as either books or computer was noted. So, what is the survey? Survey the gender of each person is noted and a hobby is either a book or computer we are making the problem very simple.

There we are just considering 2 gender there is male and female and we are considering just 2 hobbies book and computer there may be many other hobbies but we as a sponsor as if people hobbies and reading books and walking on computer these are the 2 hobbies to make the problem simpler we have just considered that. We have to find if there is an association between gender and hobby of a people that is we are to test whether gender and hobby are correlated.

(Refer Slide Time: 24:29)

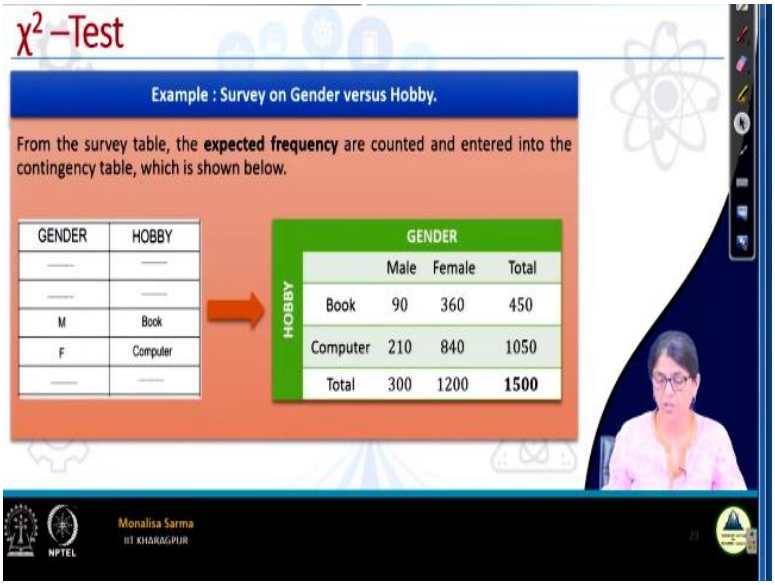


So from the survey table, the observed frequency are counted and entered into the contingency table. From the table we will have to now first develop that contingency table. So this is the contingency table we have developed. So total 250 male has given reading books as a hobby and 200 female has given reading books as a hobby. Similarly, 50 male has given working on computer the hobby 1000 female has given working on computer the hobby so, this is the observed frequency.

This is what we have observed from the survey this is called observed frequency that means filling up the O ij values now that they O ij values. So, now we will find a row total and a column total that is row residual and a column residual, what is the row residual for this is addition of this 450. Similarly, this is 1050 and column residuals, adding this and adding this. Now, next what we will find? We will find the expected frequency. How do I find the expected frequency?

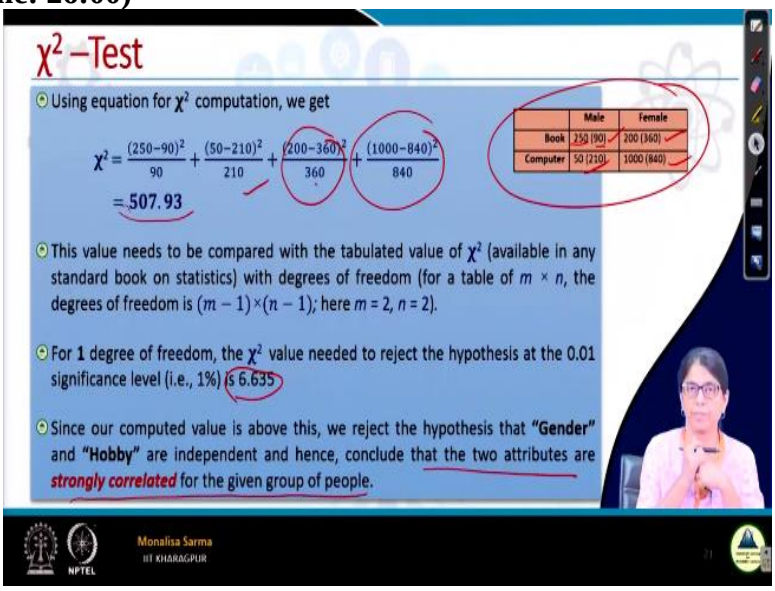
Expected frequency for this is 450 x 300 / 1500 for this is 450 x 1200 / 1500 for this value, for this value is 1050 x 300 / 1500 for this value is 1050 x 1200 / 1500. That is how I will find out the expected frequency.

(Refer Slide Time: 25:53)



So, this is the expected frequency value I got from the observed frequency.

(Refer Slide Time: 26:00)



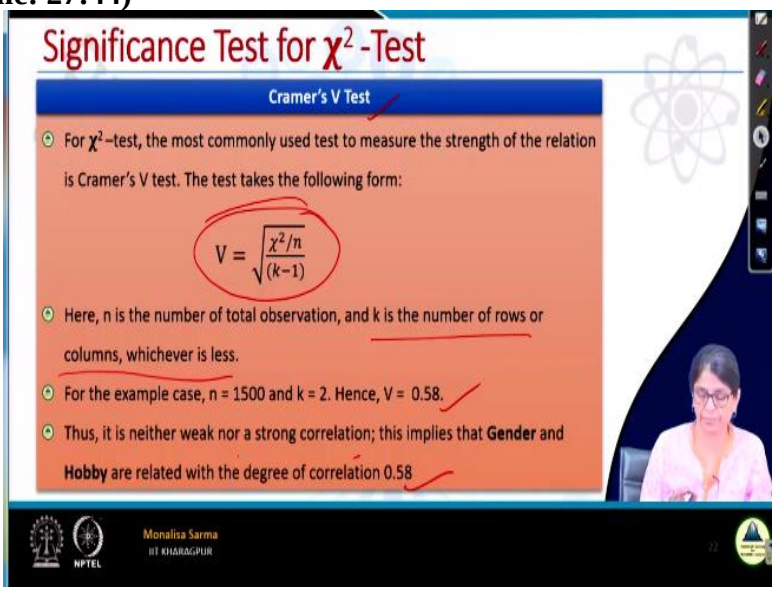
So, here in this table I am showing this is the observed frequency this is the expected frequency. Now, I have the observed frequency at the expected frequency then I can find out the chi square value, chi square value of each component of each cell, if I am interested in finding out the chi square value of this cell what will be its $(250 - 90)^2 / 90$.So, this is the chi square value for this. This is the chi square value for this, this 1.

This is the chi square value this 1 is the chi square value for this 1 this cell this value is the chi square value for this cell adding all together I got the chi square value. So, I got a chi square value of 507.93. Now, what is my degrees of freedom, what is my m and n? Both m and n is 2. So, n - 1 x m - 1 so, it is 1 + 1 is my degrees of freedom is 1. So, I will see their consult the chi square table where degrees of freedom 1 and significance levels 1%.

If I considered a chi square table for degrees of freedom 1 and significance level of 1% I will get the value 6.635 you can see from the chi square table. Now, what I value I got 507.93 is very bigger. Of course, this is a very toy example that is why I have got this different result. So, 507.93 is very much bigger than this. That means it falls in the rejection region. That means hobby and gender is there is a correlation between hobby and gender.

Since our computer value is above this, we reject the hypothesis that gender and hobby are independent and hence conclude that the 2 attributes are strongly correlated for a given group of people so it is strongly correlated.

(Refer Slide Time: 27:44)

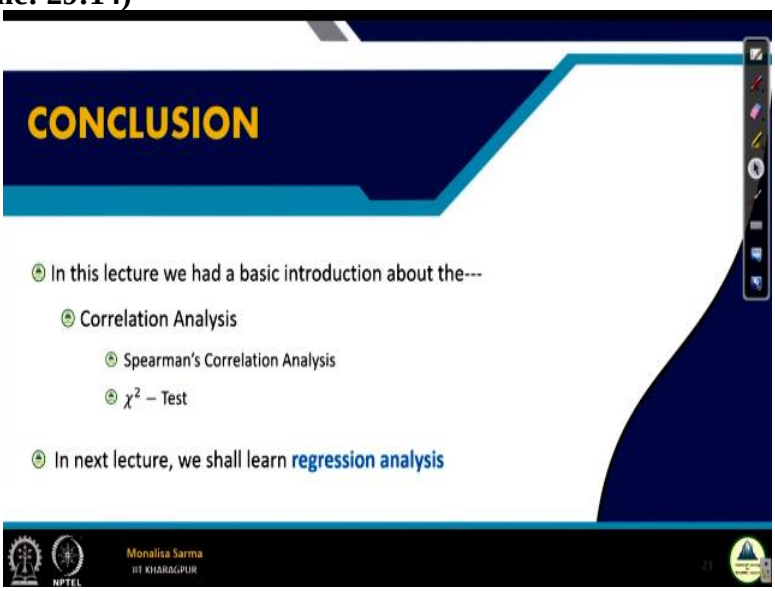


Now how to find out the strength of the correlation? We found it is correlated. Now how to find the strength of the correlation? So to find strength a correlation for chi square test 1 very famous test is Cramer's V test, we call it Cramer's V test. So, for chi square is the most commonly used test to measure the strength of the relation is Cramer's V test. So, the test takes this form I have the chi square value I have n and what is this k? k we can consider it a number of rows or number of column whichever is less.

What is the number of rows and number column different groups of the attributes? So, in this case, our number of rows is also 2 number of columns is also 2 and example what I have given regarding COVID my number of rows is 3 number column is 2. So then I will consider k = 2. So putting the value in this expression, I got the V value. So what is V? My V is 0.58 it is neither weak nor a strong correlation.

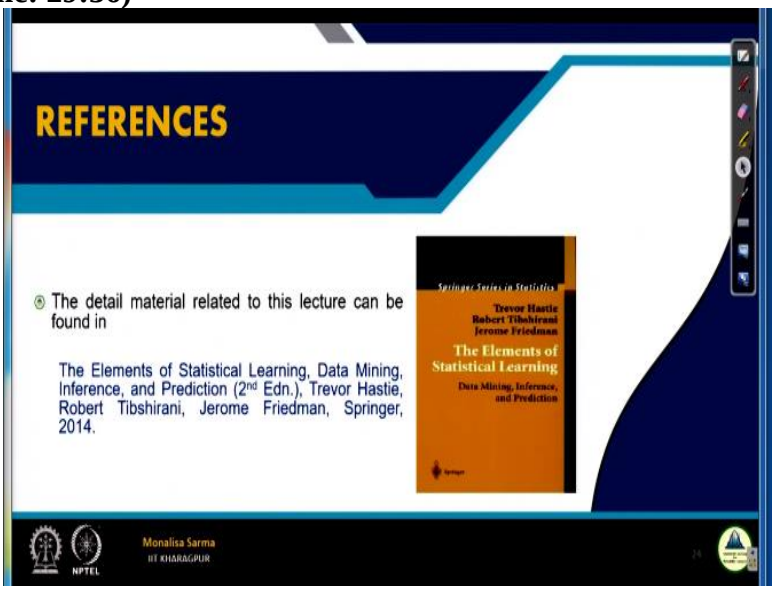
This implies the gender and hobby related with a degree of correlation 0.58 it is not a very strong correlation, not a very weak correlation or something very near to 0, I will call it weak correlation. So, it is not a very weak correlation also. So, we hobby and gender is correlated with a degree of correlation 0.58.

(Refer Slide Time: 29:14)



So, we have discussed as Spearman correlation analysis, we are discussed chi square test and that is all for correlation analysis. In the next lecture, we will discuss regression analysis. Remember I have already told in relational analysis has either correlation analysis or regression analysis. So, in my next lecture, I will be discussing regression analysis.

(Refer Slide Time: 29:36)



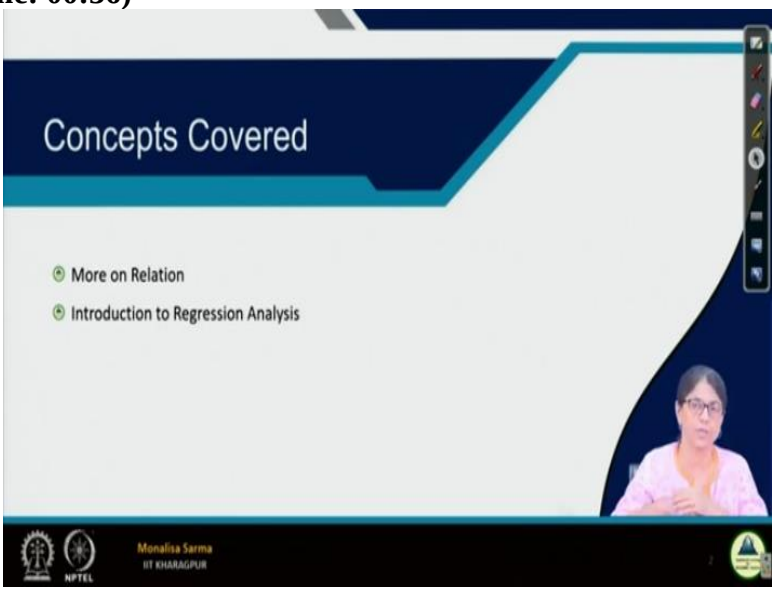
So, this is the reference and thank you.

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

Lecture - 40
Regression Analysis (Part -1)

Hello guys; so, in continuation of our discussion relation analysis today is lecture we will talk of regression analysis.

(Refer Slide Time: 00:36)



Last 2 lecture what we have seen between 2 variables or more than 2 variables if there is any relationship if there is an association and if there is an association that how strong is the association how strongly the associated we have considered that in my last 2 lectures. So, now, we will consider now in this lecture will consider more on relation, more on relation means, what I want to say like.

Suppose like if I want to if I know one value I know that there is a relation between 2 variable now, if I know 1 variable can I predict some other variable from that variable. Let me give you an example like if I know that expenses have made on advertising a product from that can I estimate the revenue of the product or if I know the expenses I have done on a producer has done on promoting a movie.

Can he predict the box office performance from the expenses done on the movie or rather another example maybe if I can I want to see if I know the house is in a particular area knowing the area can I predict the price of the house our a particular size of course. So, this if

I want to know this sort of from a relation if I want to know this sort of thing, this is basically we can do this using regression analysis as I told you when discussing correlation is both correlation and regression these are different objective and correlation analysis I tried to find out whether there is any relation between these 2 variables and how strongly they are related what is the strength of the relation.

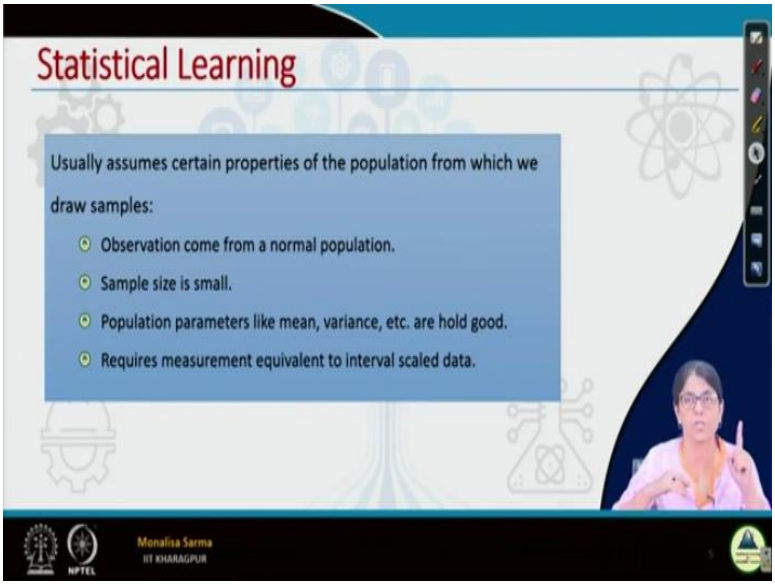
So, now, in regression analysis basically, if I know one variable from that variable can I predict the value of the other variable. So, this is basically the regression analysis. So, we in this lecture it will be the introductory class for regression analysis and the next coming 2, 3 lectures will complete regression analysis.

(Refer Slide Time: 02:35)

So, now, before talking of regression analysis here I would like to bring to your notice about learning strategies learning there are 2 types of learning concept one is statistical learning another is machine learning. So, you might think why suddenly I bought machine learning because we are talking about statistical learning, statistical learning what is this we have already seen we have done lots of thing methods and the statistical learning.

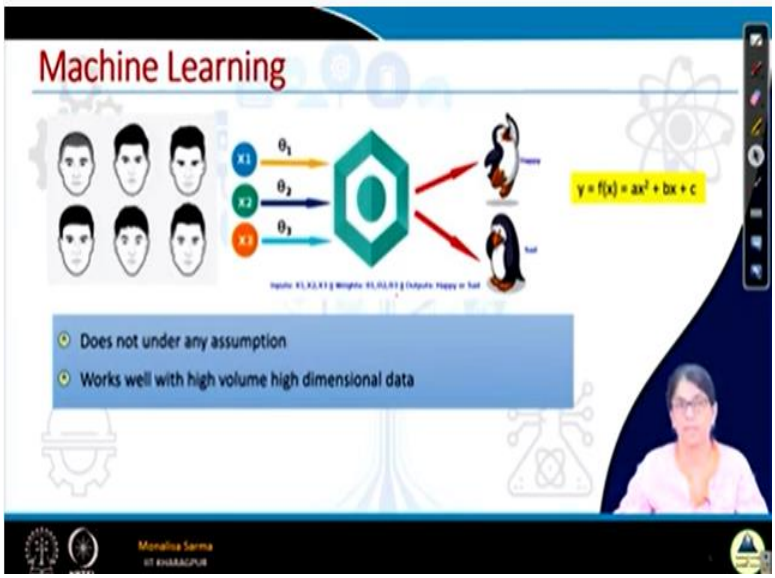
Basically we try to learn different population parameters and even the correlations also we try to learn from a population remember we have done population parameter rule we are trying to find it from the sample parameter R. So, these are all these are called statistical learning. Now, this something or some other type of learning that is called machine learning I will come to that later.

(Refer Slide Time: 03:25)



First, let us see when we talk of statistical learning always we see there are some assumption I am repeating those assumptions and whatever assumption their population has to be normal which means parent population has to be normal sample size is small, we do not need to consider a huge sample size population parameters like mean variance are whole growth and of course it is applicable for interval scaled data that is when we talk of statistical learning.

(Refer Slide Time: 03:48)



Now, when we talk of machine learning machine learning we do not need any such assumption that is what is machine learning? Basically, first let us come to the see the figure here suppose I want to find out the emotion of this people. So what my interest is what to say, predict the emotion of a people by looking at the face of a person, my interest is my objective is to predict the emotion of a people by looking at the face of a person that is what I already given.

The example by giving the amount of money spend on promotion of a movie can I predict my box office performance knowing. What I want to predict rather similarly here, knowing the facial expression can I predict the emotion of the person. So basically, what I predicting means what I need to do is that basically I need to find out the function where y = fx where x is my facial expression, y is my emotion.

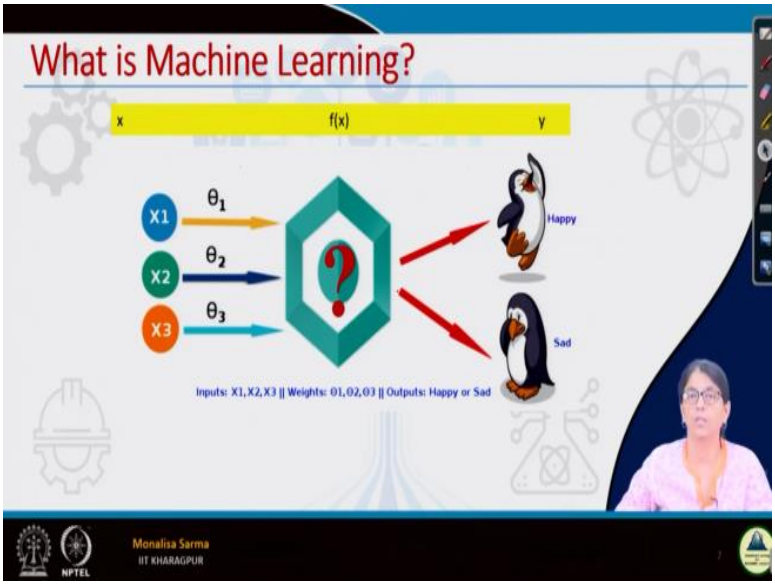
So it has nothing just to find this for me is just finding the function y = fx, same thing we do in the statistical learning also, let me just try to find out why so as to effect so, in machine learning. What we do that we try to find out the function y = fx where maybe f x is ax square + bx + c maybe our fx is ax square + bx + c. So, basically here if we our input will be this X1 X2 X3 our attributes, different attributes of this person, maybe physical attributes maybe the different pixel information of the person.

Or whatever form May be so, X1 X2 X3 are the attribute and then this theta 1 theta 2 theta 3 are the different parameters which machine we will find out basically machine learning will find our model which fits this facial expression to the emotion. So, this model maybe y = fx so machine will find a function it will find a parameter once the function is known once the parameters are known.

Then definitely we given a input of a person with a particular facial attribute will be able to tell whether a person is happy whatever the person is sad or whatever it is. So, one advantage of this machine learning is that it does not come under any assumption it works well for high volume and high dimensional data. So, when we have high volume data as well as high dimensional data, so, we were considering till now only 2 dimensional or 3 dimensional and when we have high dimensional data, there can be different categories of the data.

So and that case is this machine learning work is very well. Important point is this learning strategy needs a very large sample data that is the concept we need a very large sample data. So find a model which fits the x to y this y = fx.

(Refer Slide Time: 06:34)



So basically to summarize this given x we have to find out the fx we have to find out all the parameters. So, now the thing is that why did I talk of learning strategies why did it talk up statistical learning ready to talk on machine learning So, as I have already mentioned this lecture will be on regression analysis. So, regression analysis is very much a statistical analysis tool, we do regression analysis from population by taking a sample it is very much a statistical learning concept.

But at the same time regression and when we learn machine learning, the machine learning usually starts I can very conveniently say machine learning starts with regression analysis. So, regression analysis is like, it is statistical it is a part of statistical learning also is a part of machine learning also. So, that is the reason why I bought this here. So, now, as you could see from the constraint what I have given for both the 2 types.

Regression analysis is possible when I am confined with a very low dimensional data. When there is a high dimensional data definitely statistical learning will be definitely will have to go for machine learning. So, now, we will focus on statistical learning only we will not go for high dimensional data. So, we can very well do it by statistical learning only, as the things are seen just that we will have to if it is high dimensional data instead of statistical learning.

It was very difficult very computationally intensive, so, definitely will have to go for machine learning.

(Refer Slide Time: 08:11)

So, now coming to a main focus of this relation of this lecture that is relation analysis, So take an example suppose there is an example we have large data regarding the wages of a group of employees of a particular regions eastern region of India and we have a large data wages of the different peoples are given and this data set as a different attributes. So, what are the different attributes employee's age, employee's wage, employee's education level.

Calendar year by calendar year means at which each year how much salary he got 2022 how much he got 2021, 20, 19, 18 likewise. So, these are the different attributes in this data what we have with these are the attributes of the data.

(Refer Slide Time: 08:50)

So, now we may be interested in finding out so, how wages vary with ages. Given there are 2 things how wages vary between ages maybe if I am interested in finding out is there any correlation between is any association between wages and age it is a correlation analysis our

if I am interested in finding out knowing the age can I predict the wage then it becomes a regression analysis.

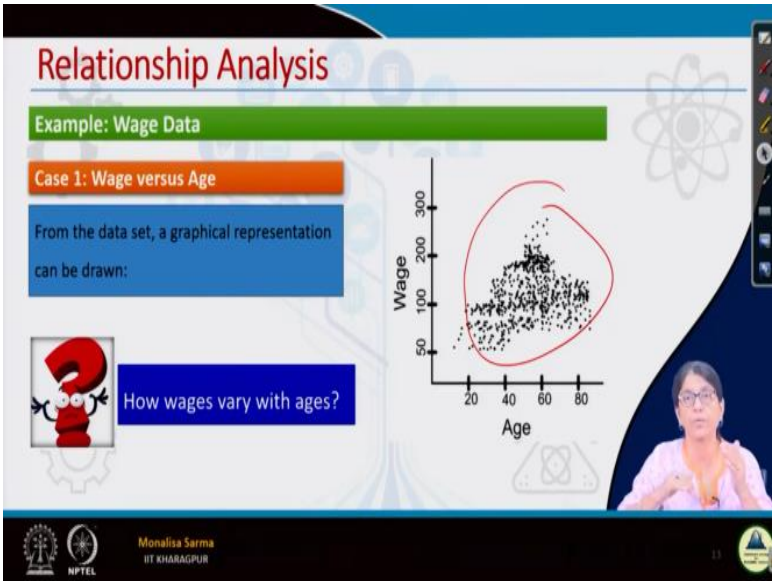
(Refer Slide Time: 09:21)

The second thing which may be of my interest is how wages vary with time, we can think in terms of regression analysis is given the wage value given a time can I predict the wage.

(Refer Slide Time: 09:35)

So, whether wages are anyway related with employee's education level.

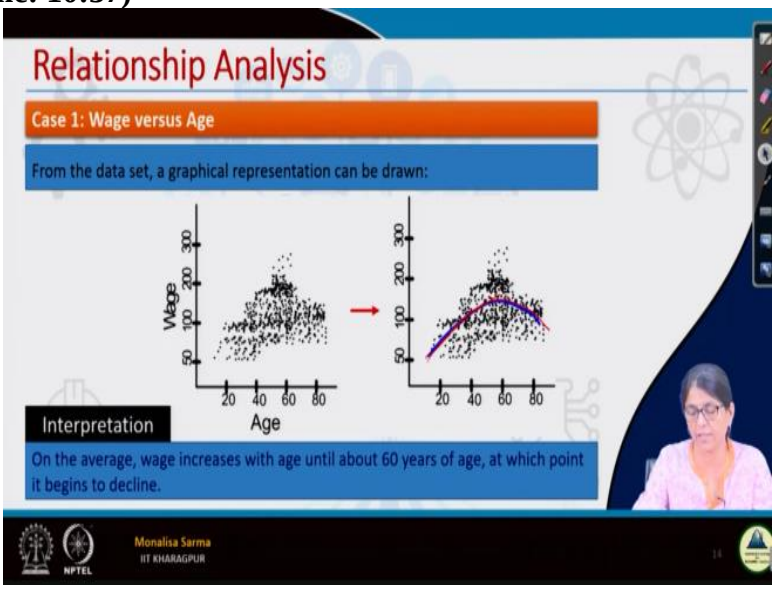
(Refer Slide Time: 09:40)



These are different on different relations which I may be interested in finding out. Now suppose I am interested in finding out how wages vary with ages. So given is can I predict wage so first what I do is from the data set, I have done a graphical representation, now see it is graphical representation. This graphical representation is very confusing like, so, is it linear is it monotonic.

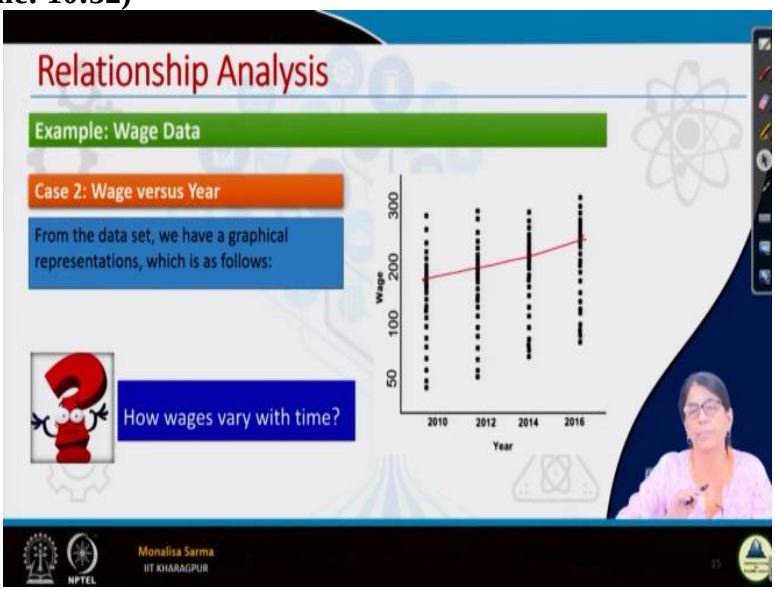
Or what it is nonlinear or what this is what can sometimes come from the graphical it is very difficult to find out what relation does it hold. So, if in the; simplest case when we try to find out the relation between 2 variables. The simplest case is that we try to graph a straight line that is called a linear relationship that is also done if we from one if we try to predict the other than it is a linear regression.

(Refer Slide Time: 10:37)



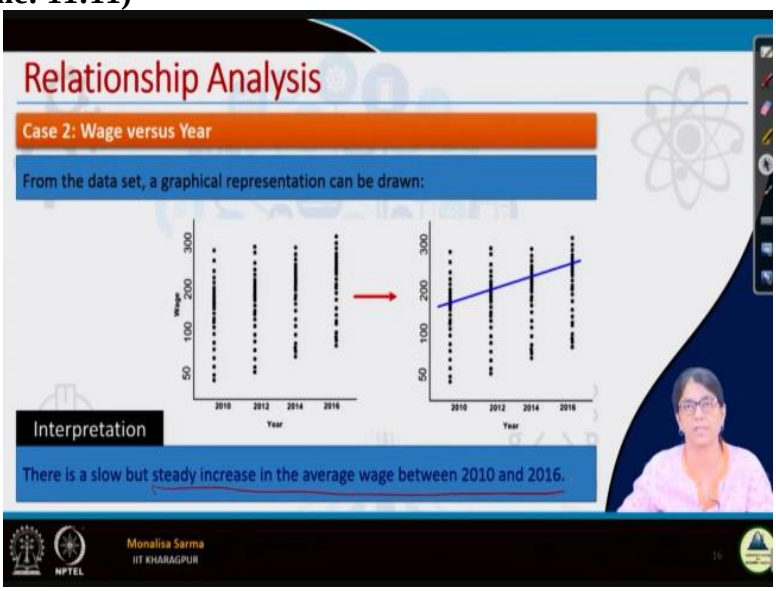
So, now, in this case this does not fit a straight line it basically you can see it is fitting a curve line which at first it is increases then again it is going down then that means it is nonlinear, we cannot fit a straight line here.

(Refer Slide Time: 10:52)



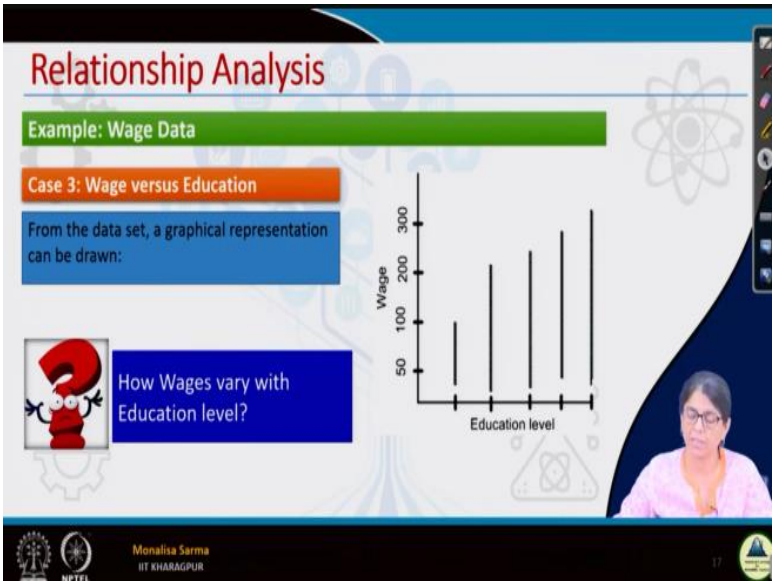
Similarly, wage verses year so, if you see wage versus time, so, if we plot it this way, so, what we get. So, maybe this if I consider this portion, so, this is there is an increasing trend, it is very much a linear relationship it is very much showing linearly it is increasing.

(Refer Slide Time: 11:11)



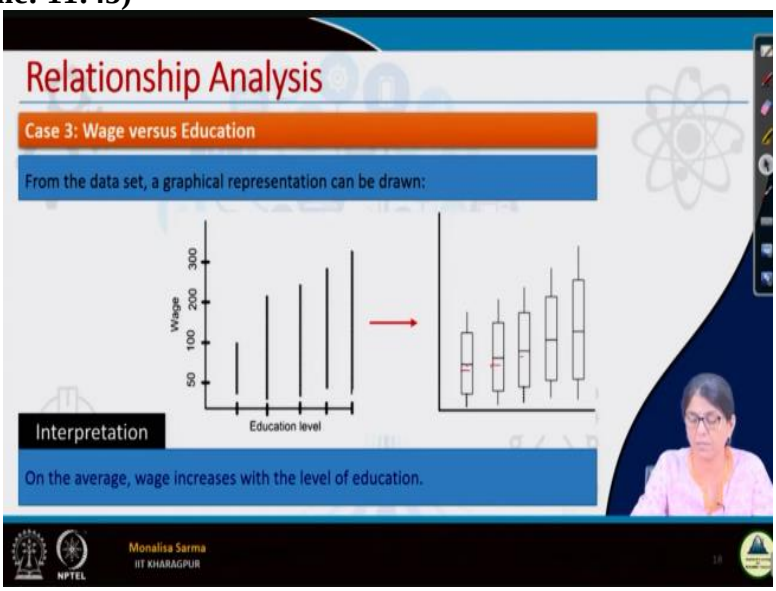
There is a slow but steady increase in the average wage between 2010 and 16.

(Refer Slide Time: 11:19)



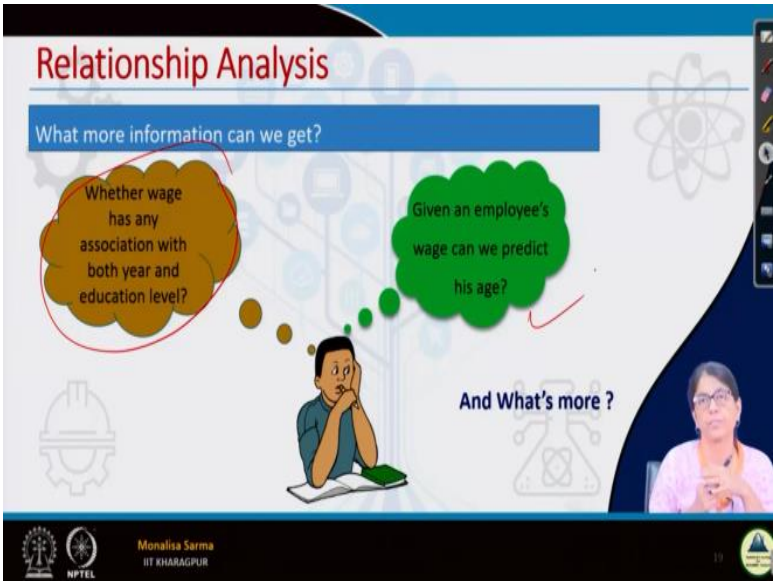
Now, again does wage vary with education level, if I just plotted it this way for different education level, if I plotted the wages for different people, I am finding it hard to find out how whether it is depicting as increasing relation or decreasing definitely not decreasing it is increasing relationally but if I find it hard to predict this, I may change this figure and I may draw a boxplot remember a boxplot how we have done.

(Refer Slide Time: 11:45)



So, if I am drawing a box plot and this in the boxplot this middle one is the mean. So, if I can see a mean increase in a steady increase in mean as the education level increases.

(Refer Slide Time: 11:57)



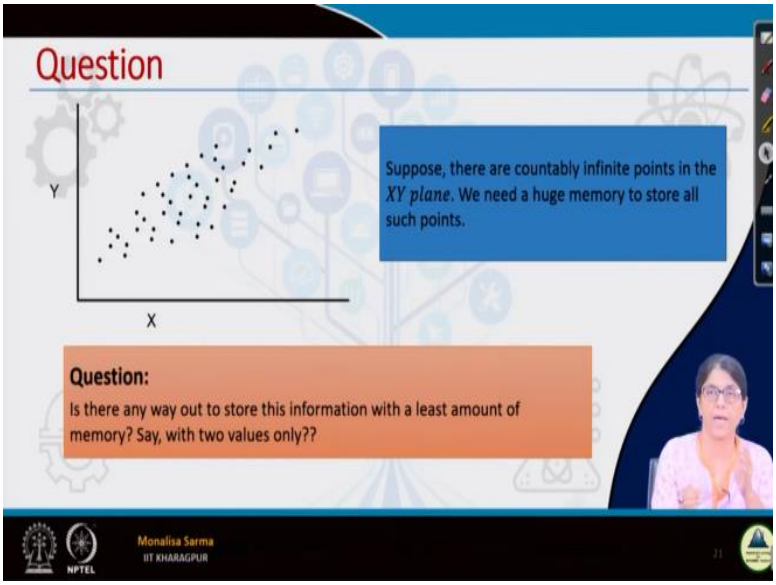
So, likewise, from a relationship, we can get different information. So whether wages is any association or both year and education level with the wages it has an association with year and education level that means what? I am trying to find out the association is there any association between these 2 that means I look correlation analysis or I may be given an employee's wage can we predict his age.

That is a different thing I have not done till now, I have done this, this one I have done but this I have not done this sort of things if I want to answer so basically, I will have to do the relation analysis is to find a relationship between the variables, if I can find the relationship between the variables and given one I can predict the other understood. So I will have to do relation analysis to find a relationship in the relationship is given.

That means what is the relationship is given means the function is given y = fx the function is given in then I can find out given x, I can find out a y, here what is y what is x y is the independent variable, x is the dependent variable, this independent variable sorry, x is the independent variable independent variable is also called regression, x is the independent variable right y = fx, x can change independently.

So, x is the independent variable that is also called regression. And y is the dependent variable that is also called response, please remember y is called the response variable or dependent variable, x is called the independent variable or regression variable in y = fx.

(Refer Slide Time: 13:33)

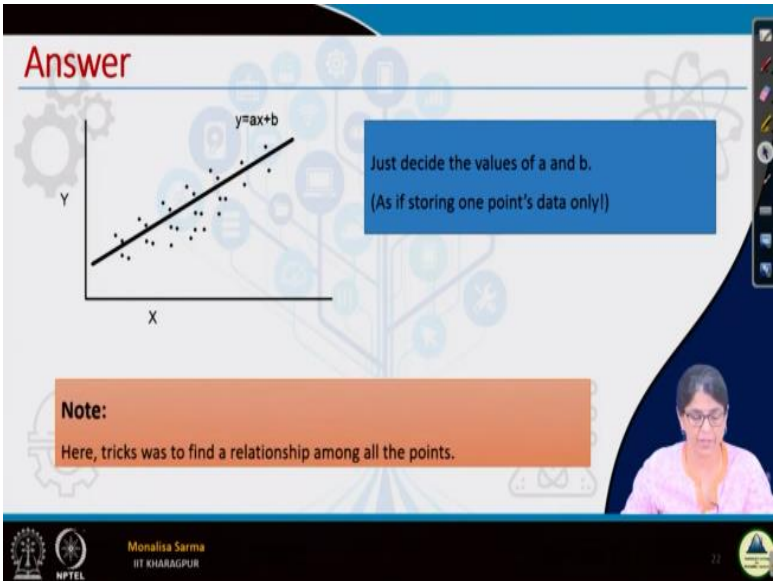


So, to how to find out the; relational analysis, how to do the relation analysis is to find the relationship let us see a simple example. Suppose here, there are countable infinite points in the XY plane, we need to huge memory to store all such points. Suppose we have 5000 points in an xy plane, which 2 attributes x and y, we have total 5000 points. So, 5000 points means we will have to store 5000 x and y in a array we will have to have an array of size 5000 x 2 so this is a huge size.

Now I do not have such much memory, so much memory, is there any way out to store this information with the least amount of memory say with 2 values only. So can I do not have such high huge memory 5000, 10,000 or maybe 50,000 data points, I do not have such huge memory to store this data. So I do not have so much huge memory to store this data. So, can I somehow store this data in a very less memory yes that is possible how this is possible.

If somehow I can find a relationship between these 2 points that means if somehow you can find a function that gives y from x, if I have the function then is the stored a function storing the function means just storing the parameters of the function.

(Refer Slide Time: 14:53)

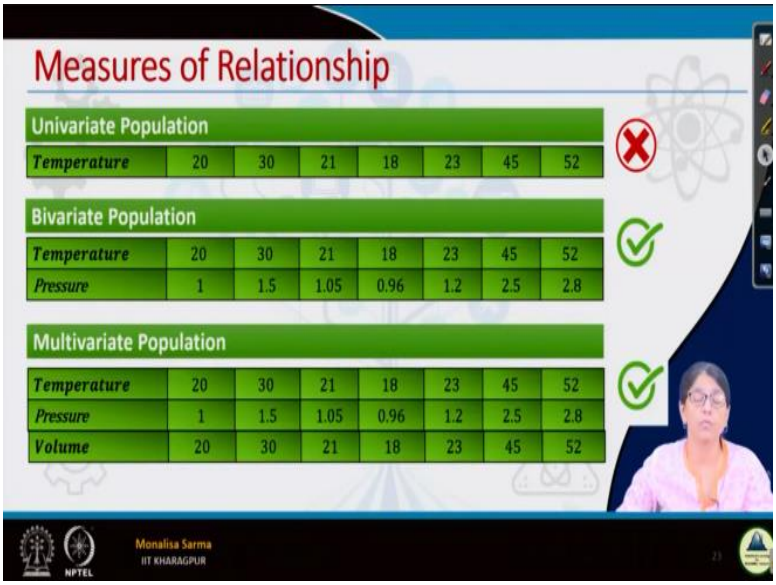


Suppose this values what I have given suppose these values fit straight line. Of course, all the points will not fall in a straight line with a slight error, we will be there that we will be discussing later. What is this error how to deal with this we will be discussing later. So if we can somehow find it relationship, suppose here the relationship or model, whatever you call, it is a straight line.

So if I can find out the straight line, finding out a straight line means I have to find out the intercept, I have to find out a slope, so y = ax + b means where b is the intercept and a is the slope. If I can find out this relationship, I can find out the intercept, I can find out the slope, then what happens I need only to store the intercept and a slope and I am done. Given any value of x.

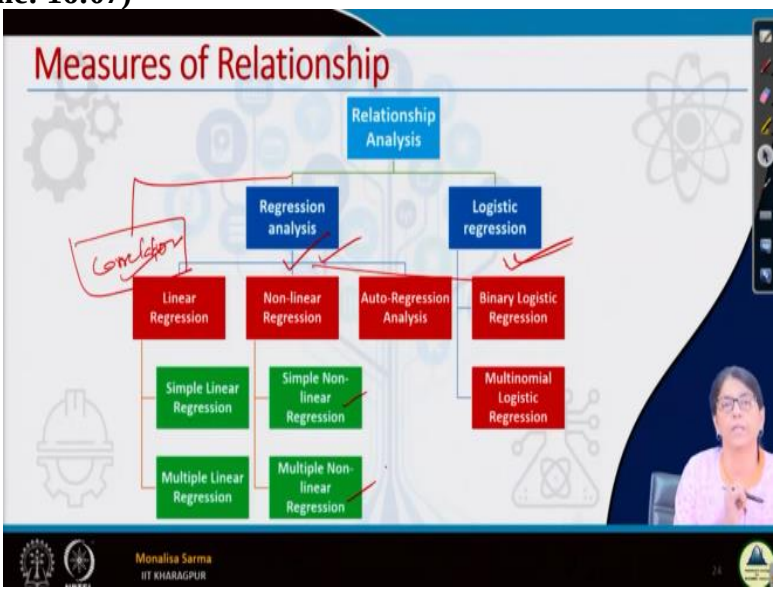
I can find out the value of y, I do not need to store all the values of x and y whenever I need, I just give the value of x, I will get the value of y this is just one example of using the relational analysis to find the relationship.

(Refer Slide Time: 15:54)



Now, when we talk of relationship as I have already mentioned before, we will talk often we will work on bivariate population or multivariate population we will never work on univariate population.

(Refer Slide Time: 16:07)



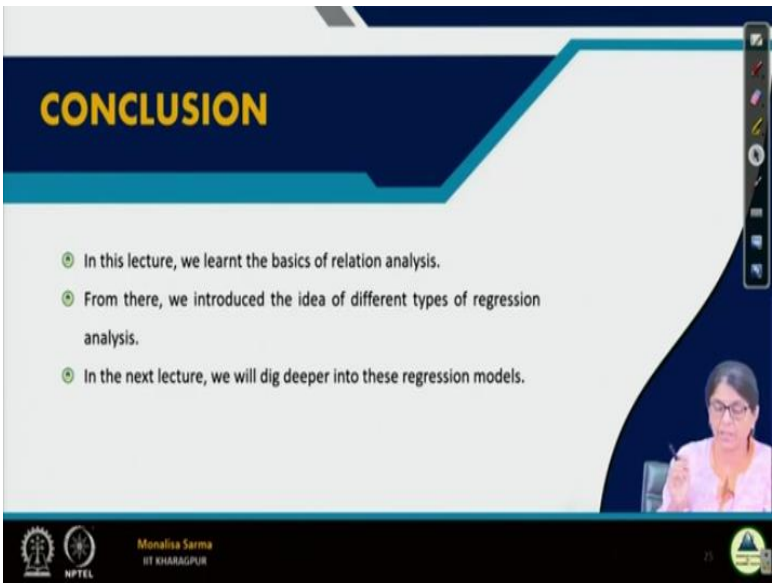
Now talking a relation analysis, last class I told relational assessments there are 2 parts, 2 ways of relational categories of relational analysis that is correlation analysis and regression analysis that is also do not get confused. Let me draw 1 more line here that means I have correlation analysis. So, there is basically regression analysis I am dividing into 2 different categories.

Because this is regression analysis and logistic regressions are 2 different types, where I will not discuss logistic regression now, I will discuss in after 2, 3 lectures maybe. So, relational analysis as I can talk regression analysis and logistic regression with the objective of

whatever objective I told of regression analysis with that objective I can have 2 different regression and logistic regression because correlation is totally different.

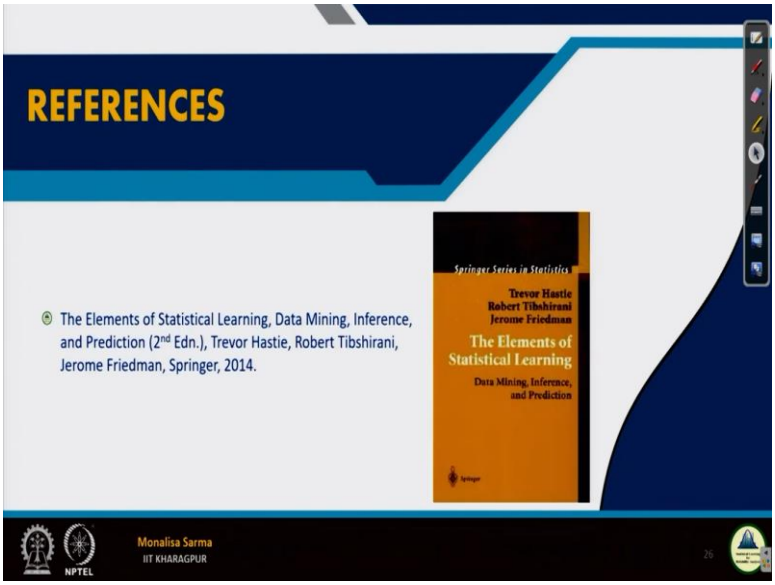
So, now, when we talk of regression analysis, there may be linear regression nonlinear regression and auto regression analysis, similarly, for logistic regression, there might be binary logistic regression, multinomial logistic regression. Again for linear regression or nonlinear regression, it might be simple linear, multiple linear regression simple nonlinear and multiple nonlinear equations. So, we will be discussing all this in the coming few lectures.

(Refer Slide Time: 17:34)



So, basically in this lecture, we learned a basic of relational analysis how we can analysis the relations to find the relationships. From there we introduced the idea of different types of regression analysis. In the next lecture, we will dig deeper into the regression models.

(Refer Slide Time: 17:53)



So this is the reference and thank you.

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

Lecture-41
Regression Analysis (Part-II)

Hello guys. So, in continuation of our lecture on relation analysis, today we will be studying regression analysis. Regression analysis we have already done one lecture and regression analysis will be doing a couple of other more lectures as well.

(Refer Slide Time: 00:42)

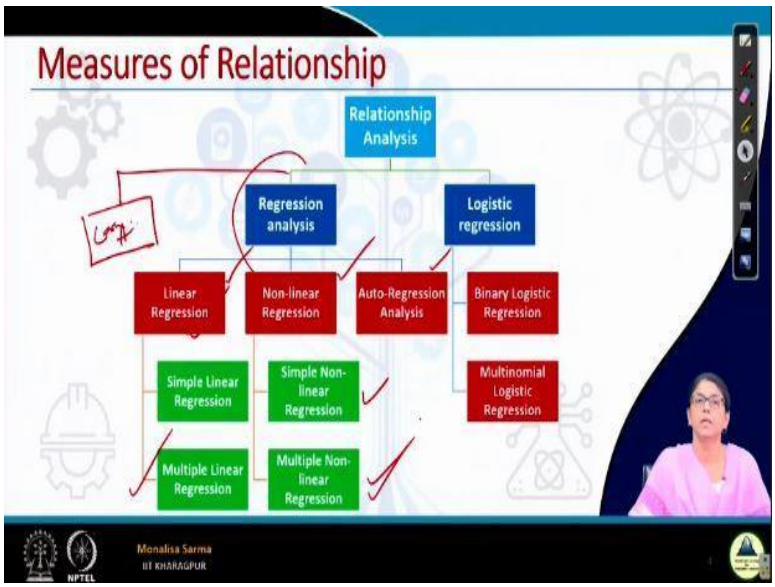
So, in today's lecture, we will basically learn simple linear regression. Simple linear regression as I already mentioned in the last lecture it is like, when there is one independent variable and one dependent variable. That means, the dependent variable we call it the response, independent variable we call it a regressor. So, when one dependent variable, when one response is dependent on the basically the regressor.

And they are basically linearly related and we can find a function that the relation basically, if we want to find out the how these two variables are related they say dependent variable and the dependent variable, how they are related if you can find the function to it. A linear functions basically, which graphs a straight line. So, that those will be discussed in details and that is in simple linear regression will be discussing those.

And now the function that we will find out, for simple linear regression, which will give an independent variable to find out the value of the dependent variable, the function that will do that. So, now this function how good this function is, how the means is this function is properly fitting the relations. Now, there are different ways to check that way. So, now one such measure is called quality of fit.

So, we will be discussing one such measure of finding out the how good the function is the function that we have found out for the simple linear regression. This two will be discussing in this lecture.

(Refer Slide Time: 02:14)



So, now coming to regression analysis well, as we have already mentioned in my last class there is a relation analysis, there are basically two type regression analysis and logistic regression. Well, I have remembered in one of my lecture I told you relation analysis. They are basically two categories, one is correlation analysis and another is regression analysis. So, here basically I can put one more thing, this is correlation analysis, regression analysis and logistic regression.

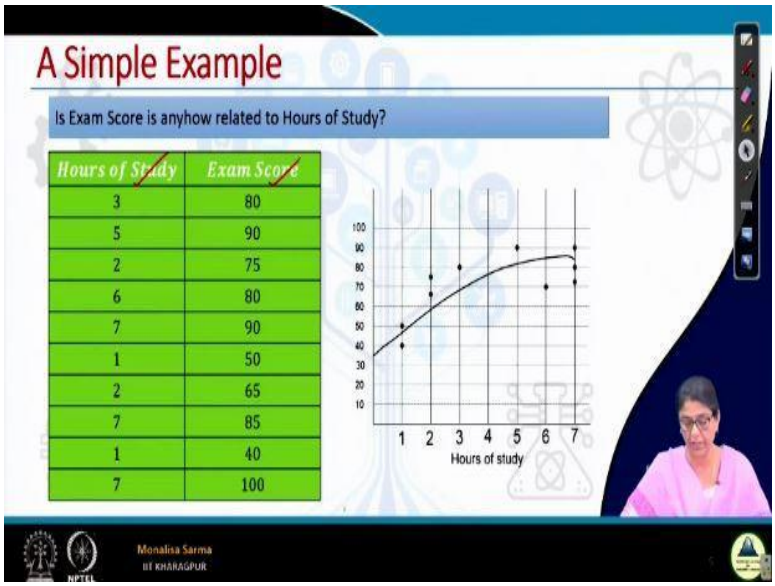
So, correlation analysis we have discussed I am just leaving that for the time being. So, now coming to this regression analysis, so in the regression analysis again there are 3 different types of regression analysis. That is linear regression, non-linear regression and auto regression analysis. So, we will be discussing this 3 regression analysis in 3 different lectures and again coming to linear regression and nonlinear regressions.

There are again simple linear regression and multiple linear regressions. Similarly for nonlinear also simple non-linear and multiple non-linear. So, simple means, when the independent variable is just one independent variable that is just one factor. And when I tell it is in many cases, there is more than one factor**.** like one good example is remember, when I have given an example price of a house in a particular locality others of the same square size, square feet area.

So, we know the price of a house in a particular area as per the square foot area. So, that is one factor. One area is the square foot area that is the one variable. And again the price of the house may vary based on the age of the house as well is not it? It can vary based on the square feet area, as well as it can vary based on the age also. So, there are 2 factors. When there are 2 factors involved and basically when there is more than one factor single.

Then we call if it is a linear regression, then we call it a multiple linear regression. If it is non-linear then we call it a multiple non-linear regressions.

(Refer Slide Time: 04:29)



So, in this lecture basically we will be discussing simple linear regression. That is one dependent variable, one independent variable. So, let us see the example, it is a very simple example. Here, there are basically 2 variables. One is the hours of study and another is the exam score. So, this is hours of study and the exam score. Now here can you tell me, which is the dependent variable, which is the independent variable?

Definitely hours of study cannot be a dependent variable, hours of study now does not depends on the exam score. Basically exam score may depend on the hours of study. So, here my

independent variable is the hours of study that is my regressor. Independent variable is also called a regressor. So, my hours of study are the regressors and exam score is the response or the dependent variable.

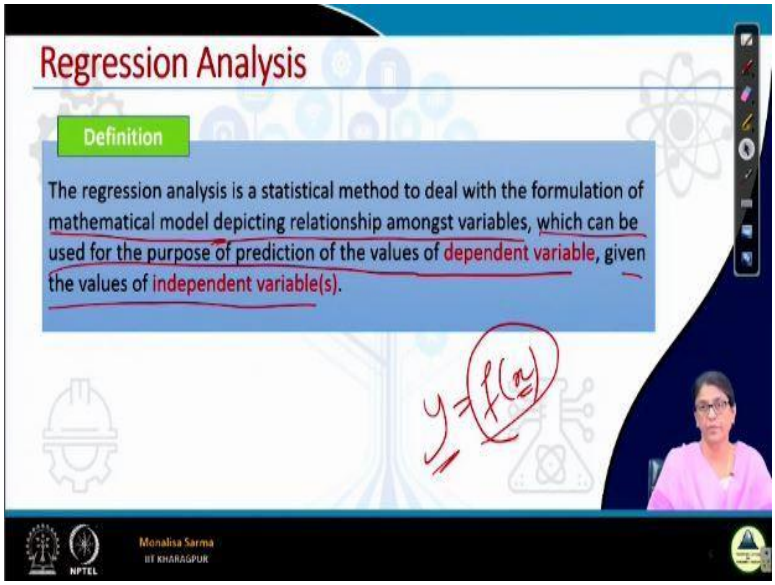
So, here this data is given. Basically how many students say 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 for 10 students, hours they put into studies and consequently the score they get in an exam. So, I want to find out is exam score anyhow related to the hours of study, is there any relations between the exam score and hours of study? So, if I just want to find out is there any relation between exam score and hours of study.

If I just want to find that is there any relation then what I will do? Then I will go to correlation analysis. So, basically correlation analysis will give me is there any relation plus as well if there is relation what is the strength of the relation as well. So, now here that is not my objective. Here my objective is that, of course it has to be related. Once it is related then only we can go to the next part like.

Now my next part is that if I want to find out if I put in say 4 hours of study what will be my likely exam score? So, if I want to answer such type of question definitely it is not a question of correlation analysis it is a question of regression. So, first what I did is I just try to plot the data. It is just simple, it is a scatter plot and I just try to draw a straight line just what to say.

So, that straight line, which more maximum of the points falls in the straight line, if not if the points does not fall in the same plane as at least it moves near by the points what is given. So, from this you can see, it has a linear sort of relationship.

(Refer Slide Time: 07:07)



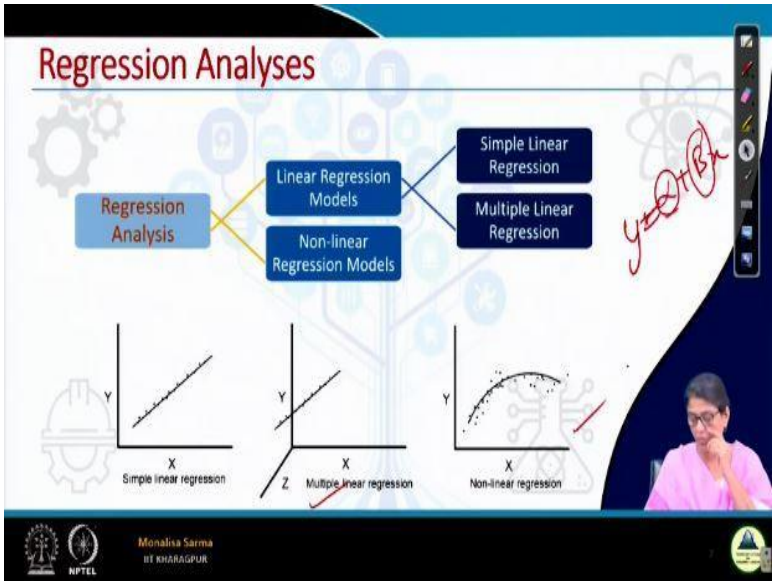
So, now from that I will now try to find out what is the relationship between this x and y. That is the hours of study and exam score. Now, first let us define regression analysis formally. So, what is this? A regression analysis is a statistical model to deal with the formulation of the mathematical model. It is a statistical model, what is this? It formulates a mathematical model.

What mathematical model, depicting relationship amongst variables what variables independent variable and dependent variable, which can be used for the purpose of prediction of the values of the dependent variable, given the values of independent variable. So, this is regression analysis is a mathematical model, which depict relationship between 2 variables. That is dependent and independent variables.

And it is used for the purpose of predicting the value of the dependent variable given the values of the independent variable. So, given the values of the independent variable in the last example what I have seen. Given the value suppose my hours of study, I put is 4 hours then what may be my likely exam score. So, I am trying to predict on the dependent variable. Dependent variable is the exam score.

So, I basically need a mathematical model, which gives the relationship between these two. The model is nothing, it is nothing but just that y = f x. What do I mean by model? Model is nothing, but just y = f x. So, my x is the independent variable, y is the response variable the dependent variable. So, I just need this function. What is this function? How y is related to x.

(Refer Slide Time: 08:42)

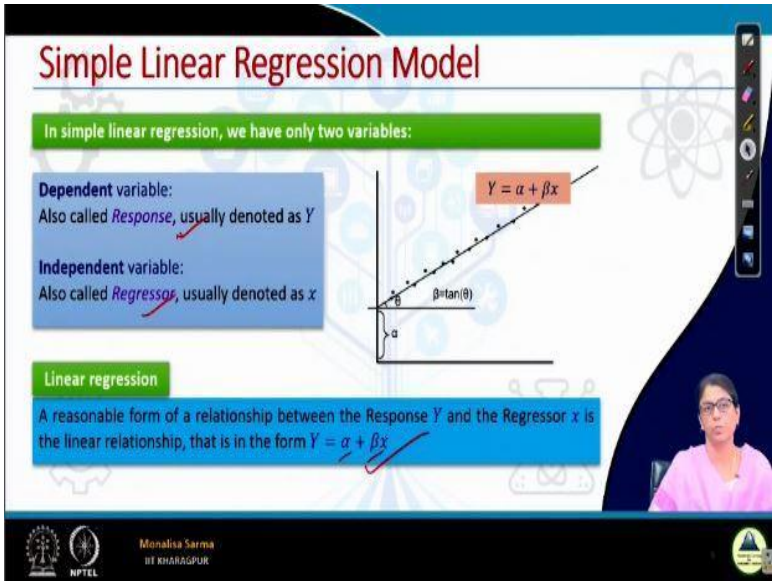


So, here is just a picture, how the different regression models look like. If it is simple linear regression then it will just graph a straight line. Maybe in the form of y = mx + c or y = α + β x. So, if I write y is equals to, this is very rating y = α + β x if I write. So, α is the intercept, β is the slope. Slope of the straight line, α is the intercept. Basically at what rate it moves corresponding to x or y moves corresponding to x that gives the slope β.

So, that is a linear regression model. Now if there are more than one factor as I have given the example, that if the property is not only dependent on the square foot it also depends on the age**.** So, then if it is more than one factor again it may be linear or non-linear. If it is a linear and more than one factor then what happens then again we may graph a straight line, but it will be in a multi-dimensional plane.

So, here you see there is a multiple linear regression. So, if it is a non-linear regression, non-linear regression it will graph a curve sort of pattern. So, this is an example of the figure for non-linear regression.

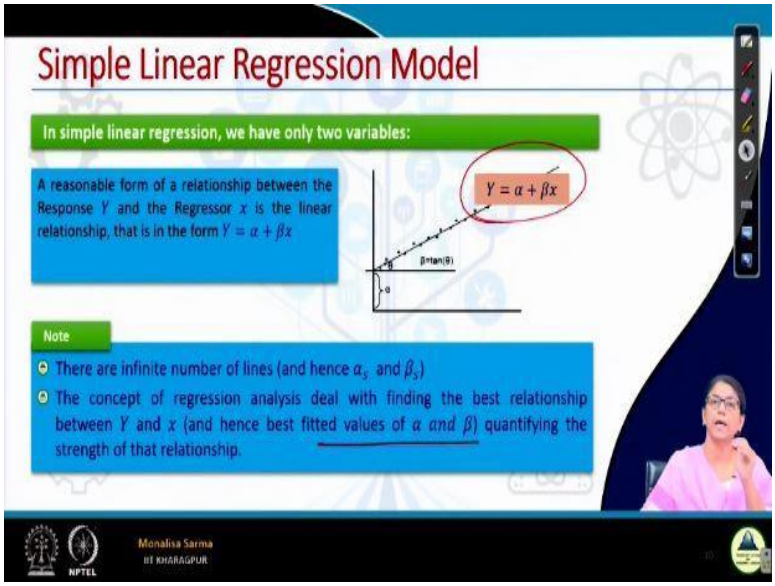
(Refer Slide Time: 10:02)



So, coming to some simple linear regression, as I already mentioned. We have two variables one is called the response or the dependent variable another is called the regressor or the independent variable. It is easy to remember. Response is the output. So, what is the output? Output is basically we are doing something to get the output. So, output will be definitely the dependent variable.

So, that is the response. One is response, other is the regressor. So, a reasonable form of relationship between response y and regressor x is a linear relationship that is a form of y = α + β x where, α is the intercept; β, is the slope.

(Refer Slide Time: 10:52)



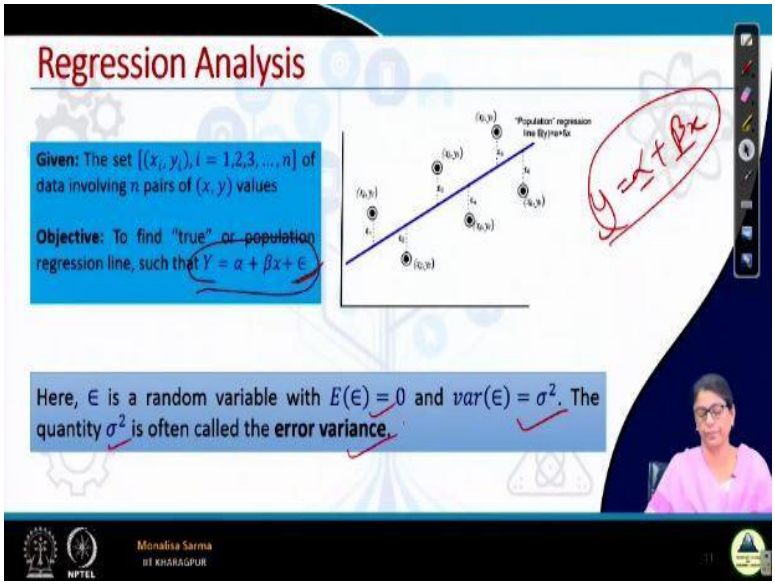
So, now this y = α + β x that is a mathematical model that we have considered. But then based on the points, if you just plot the points whatever data we have, based on the data we have

plotted the point. And we have tried to find out the line that crosses through this point. But you may see that there may be infinite lines that may cross to this point. Because it is very rarely that that is the case that all the points will fall on a straight line that we have made, that we have modeled.

So, when that is not the case there may be infinite lines that will cross through these points. So, when we have infinite lines means, there will be different values of α, there will be different values of β x. So, now the concept of regression analysis what it does is. It deals with finding the best relationship between y and x, there infinite lines, we have to find out the best line. Best line that connects y and x that gives the relationship between y and x.

And hence the best fitted value of α and β. Best lines means, the best value of α and β and that also quantifies the strength of that relationship. So, that is the main objective of the regression analysis which we have to find out the best relationship between this y and x. That means we have to find out the value of the α and β that is the parameters; α and β are the parameters.

(Refer Slide Time: 12:17)



So, now actually, here there is one thing like if I just write y = α + β x. That means, if I know α, if I know β that means given x I can find y. So, that means it is a deterministic equation. So, for given any value of x I will be able to find out the value of y, if α and β is if known. Now see the example like in the first class, when I was talking about regression analysis I have given few example like.

If the amount of money spent on the promotion of a movie depends that will help me in predicting, box of his performance. If I spend more on the more money and the promotion that my box of his performance will be more, means I will be able to predict based on the amount of money, which I spent in the promotion. But sometimes we see if many movies, where we have spent a lot of money under promotion.

But still it tanks in the box office. Again like same is the case to the advertisement, sometimes usually based on the advertisement expense we can predict the revenue of an item. But sometimes though the advertisement is we have spent quite a lot on advertisement, but still the revenue is not less. Still the product would not be sold again. Another one example, what I have given is the engine volume.

Different automobile engine volume is same, but then the mileage is different. Again this price of the house in the same locality, same square feet but, still some house are sold at a higher price, some houses are sold at a lower price. So, you see that this is not a deterministic equation. This is a deterministic equation.

So, basically from this example, what I have just told and there are many such scientific example, engineering applications, where we will see that this relationship it is not a deterministic relationship. There is another some random component to it. That random component I am calling it as epsilon. That is the error basically this random component the engineers or the statistician who calculate this.

Either they do not know what this random component is, why the price of home in the same locality with the same square foot, why it is different? This random component, this random portion why it is so either they do not know or they do not use it for calculation, there might be some other factors which are beyond their control. So, they do not use it for the calculation. So, there is a random component.

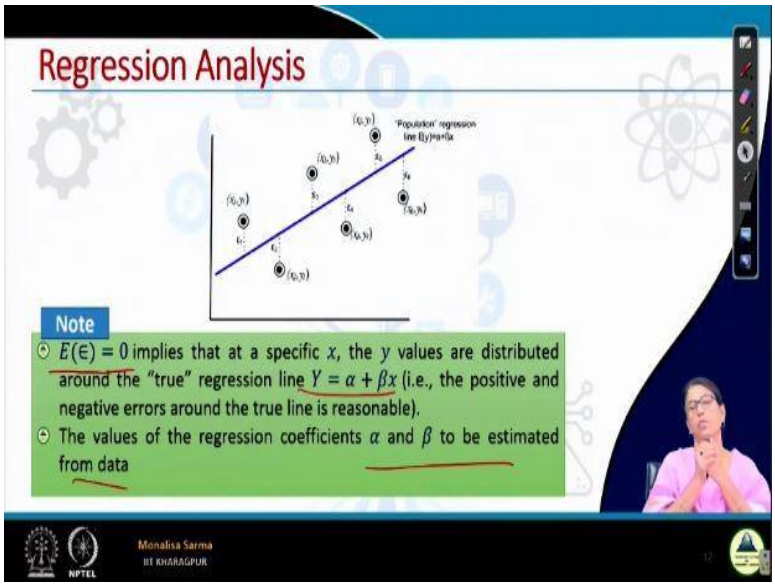
Similarly for all the other examples, so this random component makes the regression line random, it is not a deterministic line. So, our basically the regression line is this $y = \alpha + \beta x +$ epsilon. So, now here x is very much a deterministic value, x we just take it the value, which you calculate, which is already given to us. Like the number of hours, which I will put in study that is 4 hours.

There is no randomness to it that is a deterministic value. So, just there may be measurement, there may be some negligible error in measurement, other than that there is no in-deterministic pathway, no randomness to it. But for the independent and dependent variable there is a random component attached to it because of which the equations that maps the dependent variable to the independent variable is no longer in deterministic equations, because of this epsilon.

Now here this epsilon is a random variable, where expected value is equals to 0. This is the assumption we take for linear regression to be valid; our epsilon should be our error component. This error should be such that this mean of this error should be 0. If the mean of this error is not zero basically the model which you come out that model is not a proper model actually.

So, if we assume that a mean of this expected value of epsilon is 0 and it has a constant variance, homogeneous variance, that is equals to σ^2, a constant variance among all the epsilons, among all the errors. Error means, the difference of the point from the straight line from the line which we have got the regression line. The quantity σ^2 is often called the error variance.

(Refer Slide Time: 16:55)



Now what does this expected value epsilon = 0 what does it implies actually? This implies that, that x and y are distributed around a true regression line. What does that expected below epsilon l = 0 we have assumed that your epsilon is a random variable with expected value of epsilon = 0. So, what does this imply? This implies that for a specific x, the y values are distributed around the true regression line.

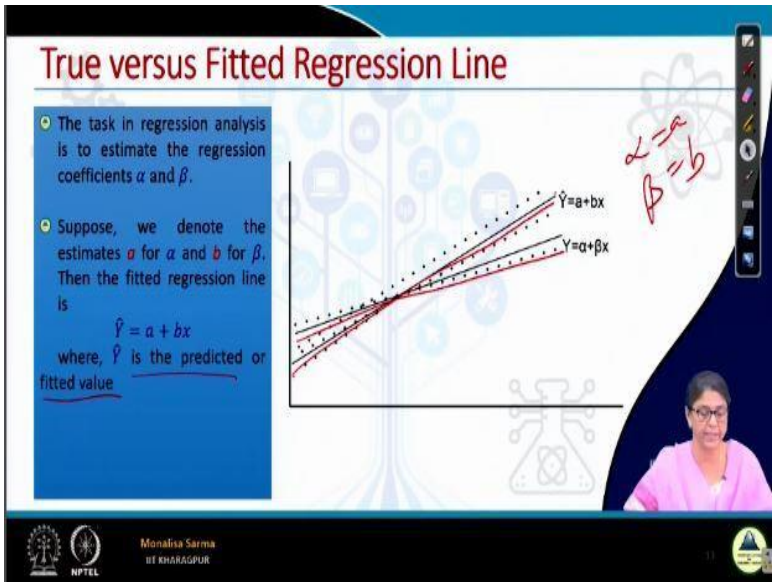
True regression line this is Y = α + β x and that is the positive and negative errors around the true line is reasonable. What does this expected value of epsilon = 0? It implies that for a specific x of my y values, which are distributed around a regression line. It will be above the regression line or down the regression line it will be distributed around the true regression line in such a way. That means the positive errors.

Positive errors means, if it is above, negative error means it is down. If the positive errors and the negative error around the true line is reasonable, it is not very high. And when it is reasonable then we can assume that e of epsilon = 0. And that is a necessary condition actually if we can do not assume that expected below epsilon = 0 then that means our fit is not proper. That means our regression line the model, which have come up is not proper line.

So, because there are lots of error to it. So, now the next what we need to do is that. We found that so we will basically as we know there may be infinite lines. So, we have to find out the best line that fits the data points, that crosses the data points in such a way that expected value of the random component, that is expected value of epsilon = 0. And once we find this line, so basically now the next step will be we have to find out the coefficient.

Finding the line means underline unless we know the coefficient that is the coefficient that is the α and β. The line is not complete our mathematical model is not complete. So, in our next step the values of the regression α and β are to be estimated. So, value of α and β are not given to us, we will have to estimate this value of α and β from the data.

(Refer Slide Time: 19:20)

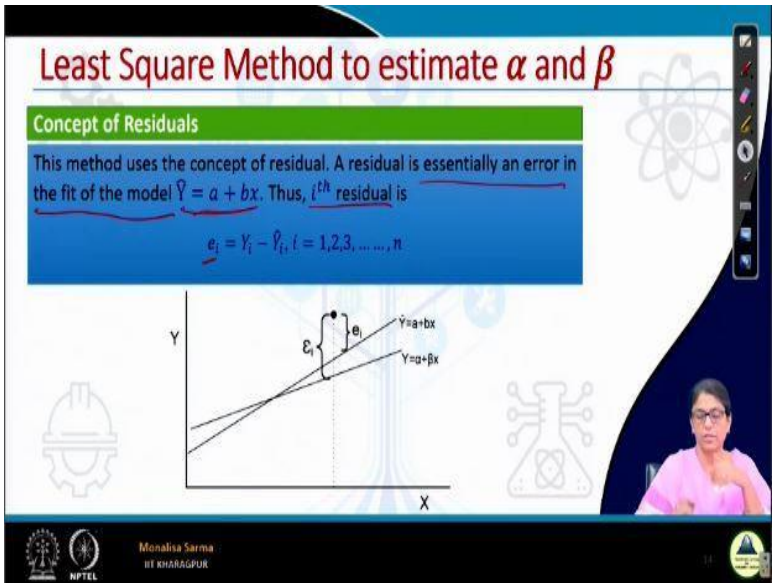


So, suppose so what we will do, we have drawn an arbitrary line**.** Now the line is Y = α + β x. Now that means we have to find out until and unless this line carries no meaning until and unless we know the value of α and β. So, basically we have to estimate the value of α and β. So, suppose we denote the estimate for α = a and for β = b.

Suppose we got an estimate, suppose we take α = a and β = b. Suppose this is the estimate we got. So, if this is the case then what is our line now? Our line is now a + bx. Now for that instead of Y I am writing is Y cap, because this is my assumption. I have assumed that value of α is a, value of β = b. With this assumption the line I got a + b x**.** Suppose I got is Y cap I am using a different variable name that is y cap, where y cap is the predicted or fitted value.

Suppose this is my actual line, this is my actual line Y = α + β x. I do not know what is α, what is β. I do not know there is in the `space there is a line. Basically there is a line, that line is y + β x, Y = α+ β that is my actual line. I do not know. I have estimated the value, say I have estimated a value of α is a and estimated the value of β is b and with that I got a line that is a + bx. So, suppose that is my a + bx gives a value Y caps, this is my predicted fitted value.

(Refer Slide Time: 21:19)



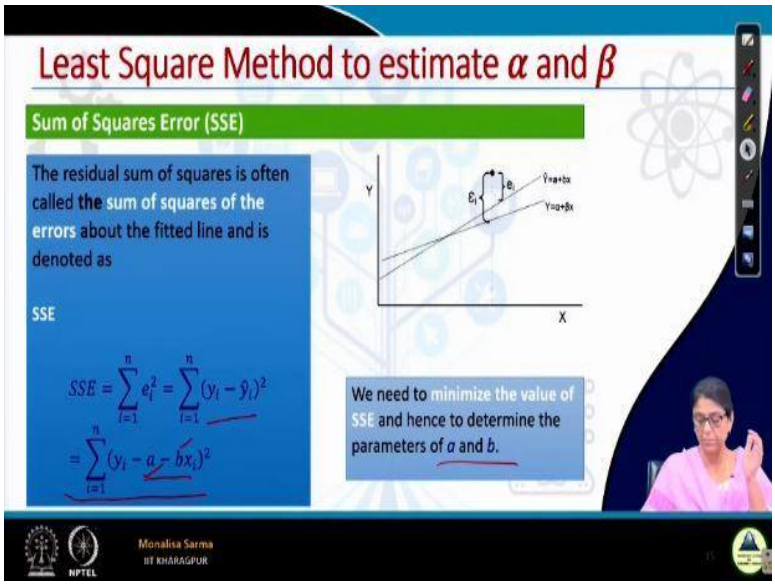
So, now we need to find out, we have estimated that value a and b**.** Estimated that α = a, β is b, we have to find out what is that a, what is that b. For that first we will have to know the term there is a term called residual. So, what is this residual? So, basically the thing is that to find out the values of α and β one of the most used methods is they call the least square method.

So, we will be learning this least square method, least square methods to estimate α and β. So, in this least square method uses a concept called residual. What is residual now? The residual is essentially an error in the fit of the model. My predicted line is this, is not it. What is my actual line? Actual line is equal to Y = α + β x. So, from these two lines, I get a residual.

What is that and that is the ith residual is e of i is Y of i - Y of i cap. That means the difference between these two. This is my actual line, this is my predicted line. So, residual is difference between these two, this portion, this is my residual. So, e of i. So, when I can say that the value whatever predicted is a good fit. So, when my e of i is as small as possible is not it. So, my residual both negative residual and positive residual when it is as small as possible then I can say that is a good fit.

So, my residual should be as small as possible. Now this is I have found out e i just for one value of x i, similarly I will take different values of x i, x 1, x 2, x 3, here like for the example I have taken different hours of study. Hours of study 1, 2, 3, 5, 7 times and many I took different X i. For different X I's I got a Y i value is not it. So, similarly I will find out the residual for all such X i's.

(Refer Slide Time: 23:20)



So, this is to find out so what is the best? I told you the best is when my this residual is as small as possible. So, what will I do with this? Now this residual can be positive also as well as it can be negative also. So, I will take the square. This method is called the sum of squares error. So, what will I do is that this is the residual sum of square is often called the sum of squares of the errors, about the fitted line and is denoted as SSE.

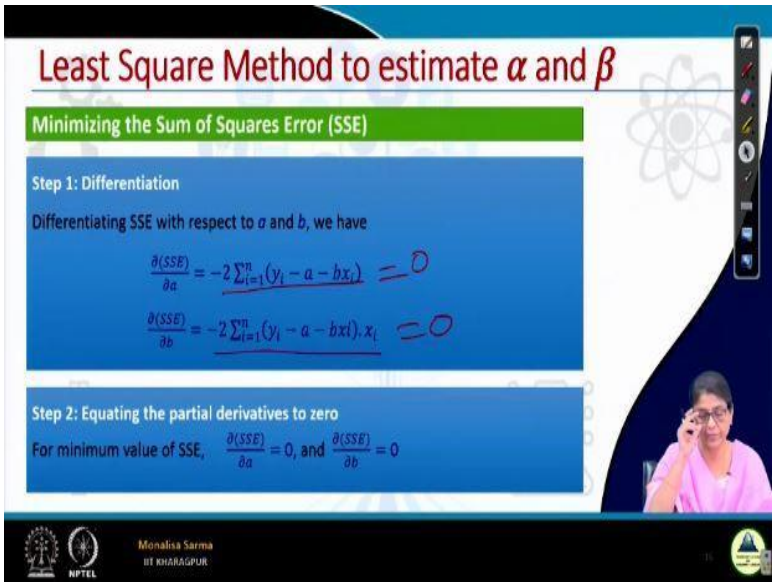
So, sum of squares of the errors, it is called sum of squares error SSE. So, how it is? SSE is nothing but summation of all e i squares. What is e i? E i is Y i - Y i cap actual value minus predicted value. So, similarly for each X i I will get actual value and predicted value. So, SSE is summation of all such Y i squares. So, this is e i ^{2} is nothing but this is the thing and what is Y i cap?

Y i cap is a - bx i is not it, that is what we have predicted the value of α and β we got this. So, this is my SSE is this? So, I need my SSE to be minimum. So, under what situation my SSE will be minimum under for what value of **x** my X and Y and SSE will be minimum. So, to find it out we need to minimize the value of SSE and hence to determine the parameters of a and b.

So, under what situation my this SSE will be minimum, for what value of a and for what value for this difference because a and b these are the parameters which have assumed. So, for assuming this a and b I have got a predicted value. So, now this difference of predicted and original actual value is my e i. So, sum of all this e i that I am calling it SSE. Now I need to minimize this SSE.

So, and for what value of my parameter estimation, because of this different parameter estimation only I am getting a different line is not it. So, for what value of this parameters a and b I will get a minimum SSE? So, how to do when I want to find a value of the variable, so that the function is a minimum.

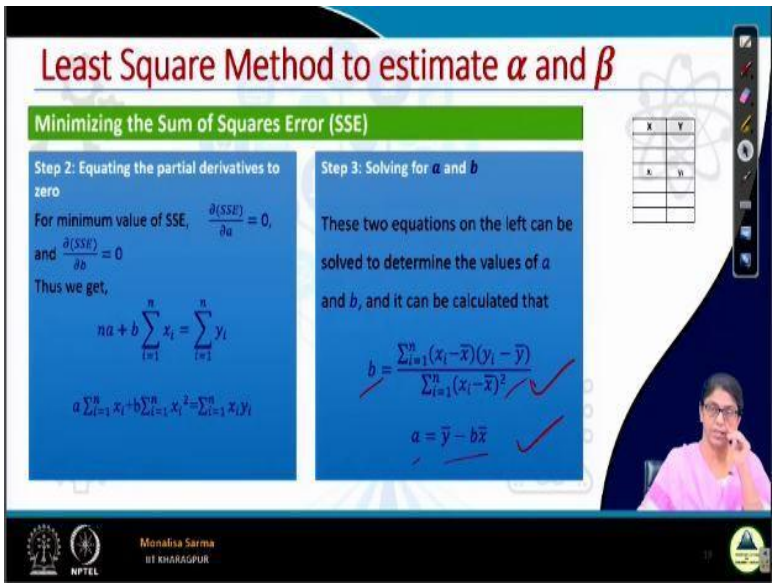
(Refer Slide Time: 25:39)



So, it is just partial differentiation and then we equalize it to 0. So, differentiating SSE with respect to a and b. This is a simple function concept we all of you know. So, I will do partial differentiation SSE with respect to a then I will do differentiation of SSE with respect to b. When I differentiate SSE with respect to a I got this and with respect to b I got this. Now this I will equate it to 0.

This is equated to 0, I got 2 equations. From this 2 equation, I can find out the value of a and b. I want to find out the value of my parameters a and b is not it? So, that is all equating the partial derivatives to zero.

(Refer Slide Time: 26:21)

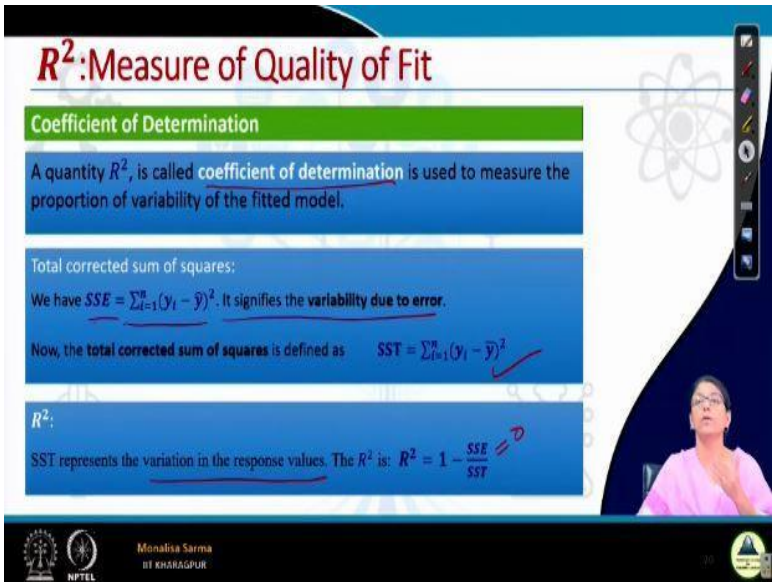


So, I got these two expressions. This simple nothing to explain here, all of you know how to find out the minimum of a function and what value the function will have a minimum value that is all nothing else. So, these are the two expressions. So, once we solve this equation. This is my value of b and this is my value of a. By solving this two equation I got my value of a and b.

So, this is the least square method to estimate α and β. This method is called least square method to estimate the α and β. So, what α value I have estimated a and β value is estimated as b. What is my a value? This is my a value, this is my b value. So, a value and b value all I am getting it in terms of x i's. See x i's and y i's basically, so this is my b value, I got all my terms in x i, y i, x bar, y bar all from the data itself. I have a set of data. Similarly I got a value also.

So, now from this we found out the value of the parameters a and b. So, I got the line y = a + bx that is my line which basically gives the relationship between x and y given a value of x I can be able to predict the value of y using this line Y = a + b x I got the line. Now the question is that, is this line the properly fits the data that is actually given or this line maybe it is not a proper fit. So, for that we need to find out one measure that is called measure of quality of fit.

(Refer Slide Time: 28:08)



So, the one measure of quality of fit is R^2 measure. It is a very commonly used R^2 measure to find out the quality of the fit of the regression line, quality of the fit of the mathematical model that we got. So, this is nothing but a mathematical model is not it Y = a + bx. So, how good the line is? So, this quantity R^2 is called coefficient of determination.

This is used to measure the proportion of variability of the fitted model. The fitted model that I have fitted that means the straight line, the straight line that I have fitted as how much variability is there from the actual. Means actual value, which I have from x i and y i what value it gives and what value this line gives, how much variability is there? So, basically that is the quantity R^2.

The quantity R^2 is the coefficient of determination is used to measure the proportion of variability of the fitted model. So, R^2. To define R^2 basically we need two terms, one is SSE that is that we have just already seen some of this. This one this SSE what is this SSE? Sum of squares error, how did you find out the sum of squares that are fitting the what the predicted value and the actual value.

So, for R^2 calculation we need two values, one is SSE and one other is SST. So, this is SSE, SSE is nothing it signifies the variability due to error. SSE what does it signify? This is the formula for SSE. We have just now seen is not it. What is the variability due to the error? Means it is basically how much is the error between the predicted line and the actual line. So, that is my sum of squares due to error.

And another one term that we need is total corrected sum of squares. Total corrected sum of squares you are not new to this term we have used this term by learning anova remember. So, total corrected sum of squares it is basically again it is also variability of the data from the mean of the data. So, that all the data that we have used for regression analysis how much this data is variable from the mean of that data.

See SST is nothing but y i - y bar. We have taken some data x i and y i is not it. This y i that we have already some data as we have based on this data only we are finding the model. Without the data we cannot find a model. So, now this data how much this data there is variability in this response data. So, that is SST. Now so what does R^2. See SST represent the variation in the response values.

How much this response why I add a response values, how much variation is there in the response value? That has nothing to do with the error or something like that. We have taken different x value, for different x value what is the y value. So, that is this different y values that we get this what is the variability of this data. That is the SST total corrected sum of squares. So, my R^2 is nothing but 1 - SSE / SST.

Now if my SSE = 0 suppose, under what situation my SSE will be 0? If my SSE is 0 that means my line is a very good fit is not it, my line is a very good fit then only my sum of squares due to errors is equals to 0. In that case what I will get $R^2 = 1$. That means it is a very good fit, quality of fit is very good. Even if suppose my SSE is quite less than SST. If it is quite less than SST it is not zero but quite less than SST then also I will get a very less term here.

If I divide this SSE / SST and my R^2 will be closer to 1 that also indicates a good fit. Of course 1 is a very good fit and values which are closer to 1 also good fit. Now if SSE is almost equal to SST. That means the sum of variability due to error as well as the variation in the response;

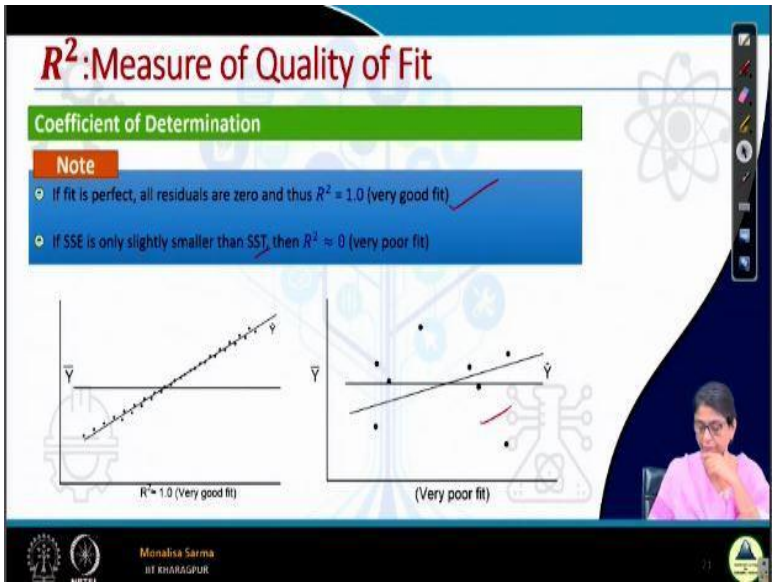
variable due to its error is so high that we are getting almost equal to the variation in the response value.

Variation in a response value is different thing is not it; my variable to do error is so high that I am getting it is almost equals to SST. Then what will happen? My R^2 will be equal to 0. That means my fit is not at all a proper fit. It is not a good fit $R^2 = 0$ means it is not a good fit at all. And theoretically in fact my R^2 value can be less than 1 also. Under what situation my R^2 value will be less than 1.

When my SSE value is bigger than SST see pay attention please when my SSE value is bigger than SST sum of squares due to error is bigger than SST you can imagine that means that the line I have got that is not at all it is showing totally wrong trend. That is not at all the line that actually these two independent variables and dependent variables are fitted. The line that the regression line I have got that is totally in the opposite trend.

It is not showing the proper trend that is why I am getting such a high error that it is bigger than SST. So, theoretically R^2 less than 0 is also possible, negative also is also possible. In that case it indicates my quality for fit is equation of quality of it does not come basically, it is totally then my regression line that I got it is totally showing a wrong trend.

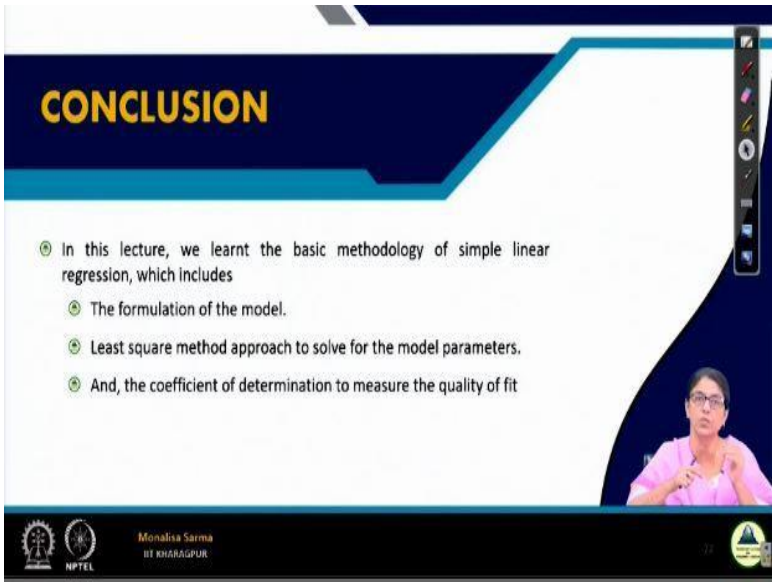
(Refer Slide Time: 34:11)



So, see if the fit is perfect, all residuals are 0 and $R^2 = 1$ as I told you SSE = 0 then all fit is perfect, all residuals are fit, that is of course it is very impractical but getting a value, which is closer to 1 is also very good fit. So, if SSE is only slightly smaller than SST, then $R^2 = 0$ almost

equals to 0 that is a very poor fit. So, you see it is a very good fit. Almost all the points are in the line or very near the line and this is example where it is a very poor fit.

(Refer Slide Time: 34:50)



So, in this lecture, we learned the basic methodology of simple linear regressions. What that includes? That includes the formulation of the model, how we basically formulate the model to get the function, between the dependent variable and the independent variable. Then we have seen the least square method that approach to solve the model parameters, that is α and β. That is the intercept and the slope.

And also we have seen the coefficient of determination to measure the quality of fit, coefficient of determination that is the R^2 to measure the quality of fit. So, we have discussed all these three things in this lecture.

(Refer Slide Time: 35:30)

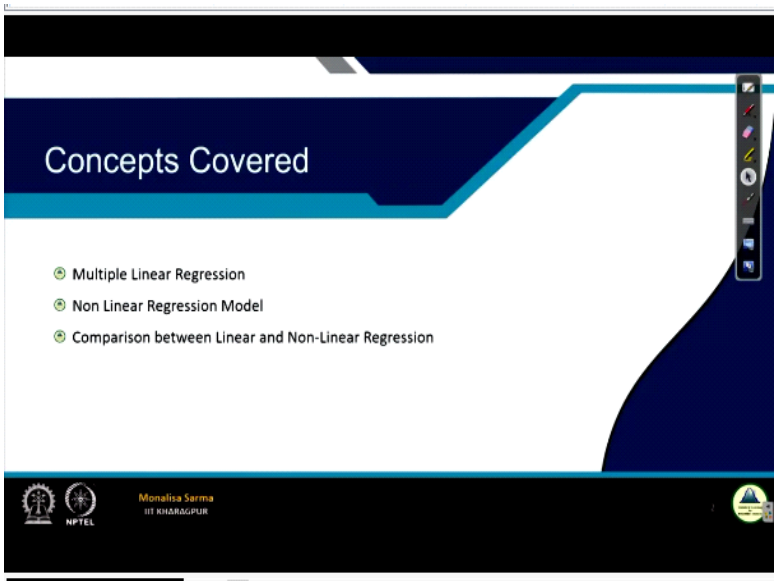
And so this is the reference. And thank you guys.

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

Lecture-42
Regression Analysis (Part-III)

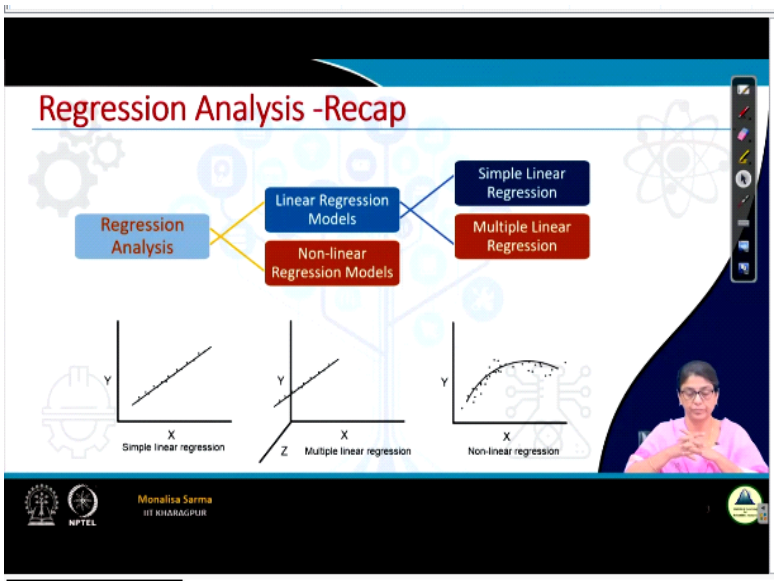
Hello guys, so in continuation of our discussion in relationship analysis, basically we were discussing linear regression. In my last lecture I discussed linear regression.

(Refer Slide Time: 00:38)



So, today's lecture I will discuss multiple linear regression, non-linear regression model and then I will try to see a compression between this linear and non-linear regression model. So, when I talk of multiple linear regression means there will be more than one factor; one factor means which factor more than one independent factor. Like a very good example is the same example I want to give basically which I have given is the price of the house, it not only depends on the square feet it also depends on the age of the house, there are two factors, square feet as well as age.

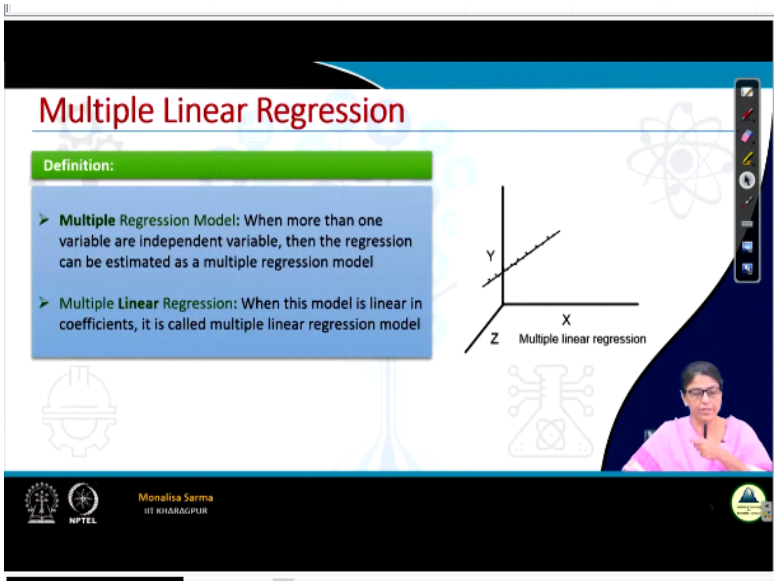
(Refer Slide Time: 01:10)



So, that is basically multiple linear regression, if the regression line is a straight line then it is a multiple linear regression. If the regression line is not a straight line it is a curve then it is a multiple non-linear regression. So, this figure already we have seen in my last lecture also. So, just a quick recap. So, this is the simple linear regression, simply it features straight line y = mx.

So, mx + c or $y = \alpha + \beta x$ whichever you take. So, now this is the multiple linear regressions where if it is a linear regression but more than one factor, so we will consider a straight line in a multi-dimensional plane, it is a non-linear regression where basically it does not graph a straight line but maybe a curved line, it may be a parabola, it may be hypo anything like.

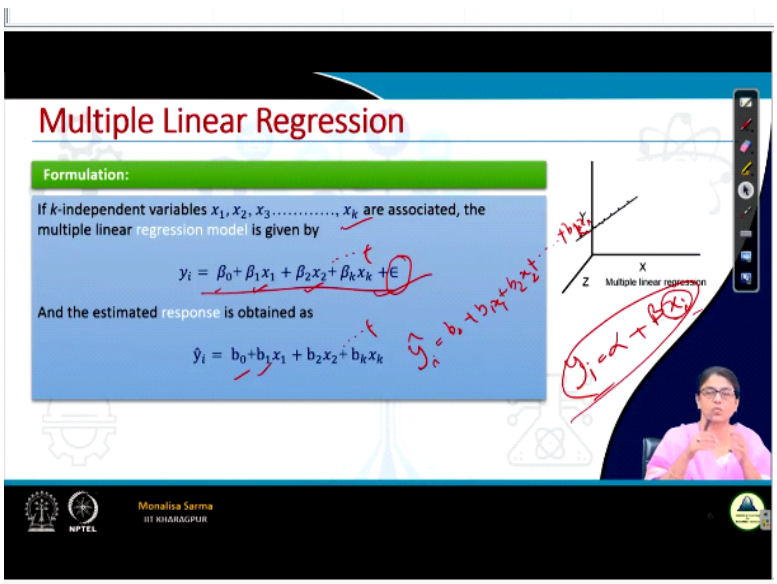
(Refer Slide Time: 02:10)



So, when more than one variable are independent variable, I am just talking dependent variable is one only. So, when more than one variable are independent variable then the regression must be estimated as multiple regression model. So, more than one independent variable, so like in this example in the price of the house there can be some more other example we can think of anyway forget it.

One example is sufficient; you got the point basically that is the most important thing. So, now this multiple linear regression, when this model is linear in its coefficient it is called a multiple linear regression model, when they have more than one factors and the model is linear then we call it a multiple linear regression.

(Refer Slide Time: 03:04)



So, now multiple means there can be more than one factor, more than one independent variable. So, let us consider k independent variable x 1, x 2 up to x k. Suppose these are associated, now the multiple linear regression model is this. Earlier what was my regression model is y i = α + β x i. But at time using i is a subscript there may be n data and x i data, for n x I am getting y and y i data, y data.

So, that is why this subscript i. So, this is earlier my first simple linear regression this was my model, y i = α + β x i. Now there are more than one independent variable; earlier just one independent variable that is x. Now there are more than one x independent variable. Let us take

there are k independent variable. So, if there are k independent variable x 1 to x k. So, the regression model will be in this form β 0 that is the intercept β 1 x 1, β 2 x 2 **+,** so it will be dot dot dot + β k x k.

There is a slight mistake here; it is β 2 x 2 + dot dot + β k x k + ε. Why again this ε is there because this is not a deterministic equation. As we have mentioned why it is not a deterministic equation because of this random component that is why it is making it y i as a random, we cannot deterministically say that for particular value of x 1 x 2 up to x k we can knowing the coefficient will be able to tell the value of y i.

No we cannot deterministically say there is always a randomness to it and now we need as same as what we have done for linear regression. So, we need to estimate the value of the coefficient, here the coefficient is β 0 β 1 β 2 β 3 β 4 up to β k. So, we have to estimate this value of this coefficient. Now how we have to estimate the same method least square method what we have used in the linear regression method, for that first what we will do? We will first assume some values for this coefficient.

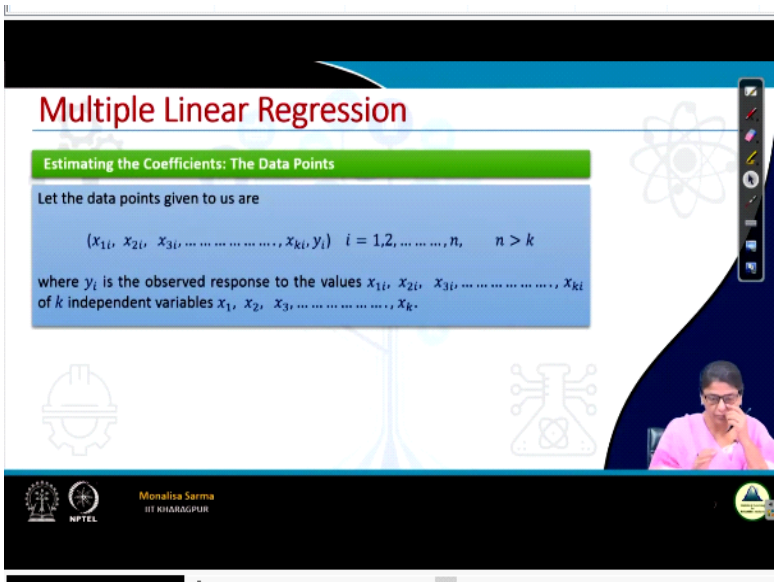
Now since there are k variables, when there are, k variables how many coefficients are there? K variables then there are total k + 1 coefficient, there are one variable, there are two coefficient α and β. So, there are k variables, so there are total k + 1 coefficient. So, what is the k + 1 coefficient β 0 β 1 β 2 up to β k. So, now suppose we have assumed as value for β 0 as b 0, β 1 as b 1.

So, plus this equation I am just writing one equation both the equations are wrong basically y i cap = β 0 + β 1 x + β 2 x and β 1 x 1 β 2 x 2 + dot, dot + β k x k. So, here basically dot, dot means it goes from x to x l, it is not only it is just showing as if x 1 x 2 x k that is wrong. So, this is my if I have estimated the value for β 0 β 1 β 2 I have estimated as to be b 0 b 1 b 2.

So, if that is the case, this is my predicted value y i cap is my predicted value. If my y i cap is the predicted value, I need to estimate this parameter I will be using the least square method only. Now using this same method again I will have to find out the residual. So, what is the residual for each y i value. So, how do I find out a residual? Same method y i - y cap. So, this residual may be

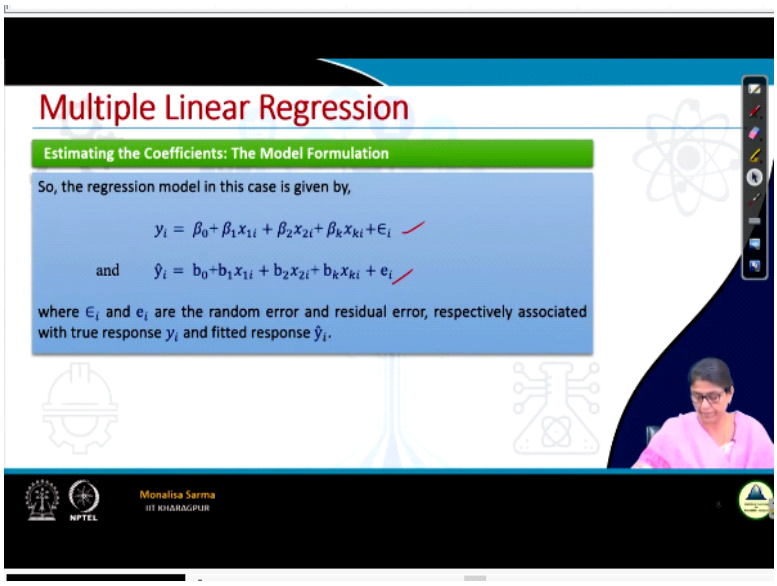
positive, it may be negative. So, that I get a proper residual, I will square it, so, sum of all this residual square.

(Refer Slide Time: 07:24)



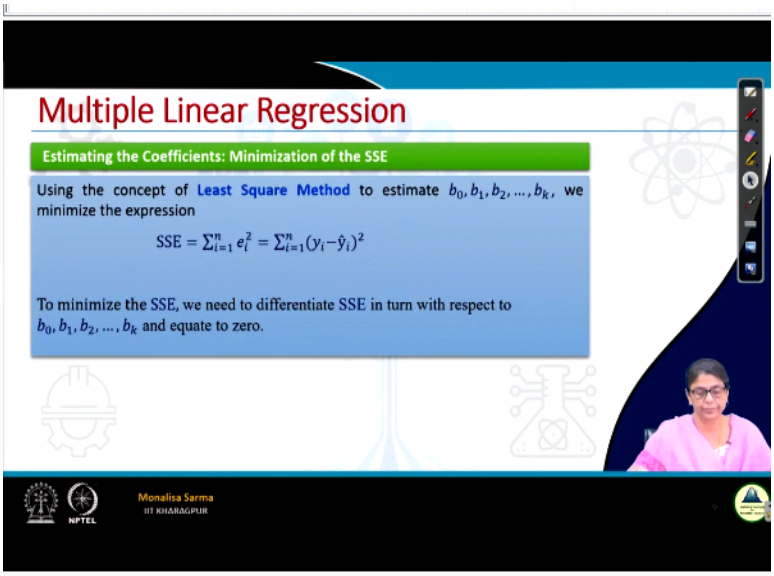
So, say and here let the data points to be used are these are my data points, for y i where i = 1 to n, so y i is the observed response to the values this of k independent variable. This is if you have understood the previous one simple linear equation, multiple linear results exactly the same just that instead of one variable I am using here more than one variable that is all everything remains same.

(Refer Slide Time: 07:55)



So, the regression model in this case is this is the ε i is a random component. Now y cap also and this is e i; where ε i is a random error and e i is the residual error. Residual error means the predicted value and minus the actual value is what I get is that that is the residual error. So, where ε i and e i the random error, residual error respectively associated with a two response y i and a fitted response y i cap.

(Refer Slide Time: 08:29)



So, using the same concept of least square method, I will use the same to estimate this b. So, what is my SSE? SSE is nothing but the residual error square, sum of the all the residual error square. So, what is the sum of this residual error square is nothing but this y i - y cap whole square summation up from i = 1 to n, i = 1 to n means all the data points what I have taken.

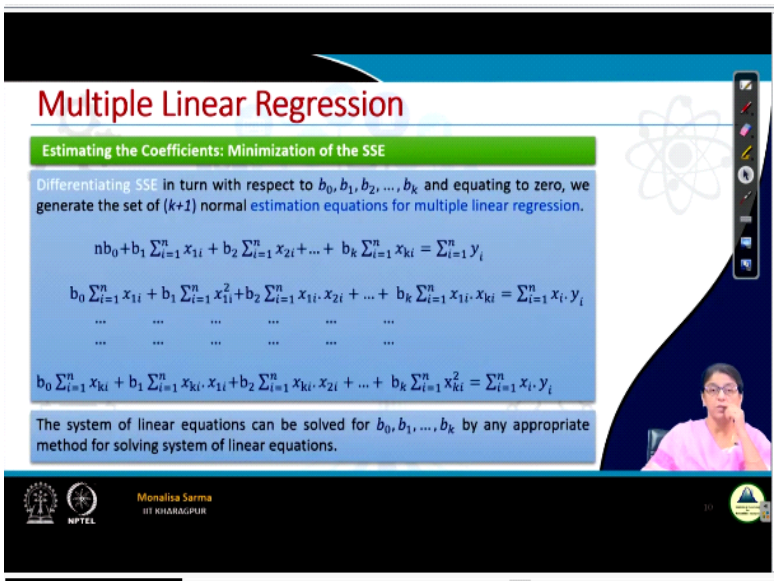
To minimize SSE we need to differentiate SSE in terms with respect to b 1, b 2, b k = 0. Now we have to find out the value of the parameters b 1, b 2, b 3 whatever we have assumed we have to find out the values of b 0, b 1 up to b k. So, how to find out the we have to find out the value means now basically our SSE should be minimum or fitted line should be such that the SSE should be minimum.

So, the same concept how do I find a value for which value the function is minimum, so how do I that to find out the for which value of x the function is minimum basically I do the differentiation of that first order differentiation and then equate it to 0. Similarly, whatever we have done it for

simple linear regression I will differentiate SSE now here earlier I have differentiated SSE for partial differentiated SSE with respect to α and with respect to β.

Now I will differentiate SSE with respect to b 0, b 1, b 2 up to b k and then equate all this to 0 and then once I equate all this to 0, so total how many variables I need to find out?

(Refer Slide Time: 10:00)

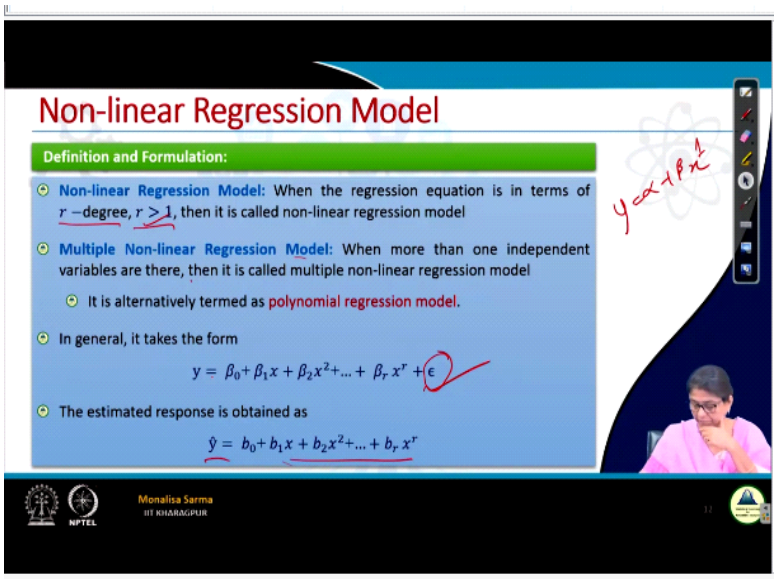


K + 1 variables. So, differentiating SSE in terms with respect to and equating to 0 we generate total k + 1 estimation equation. I will get total k + 1 because a SSE I will differentiate with respect to all these parameters. How many parameters are there? K = 1 parameters, when I differentiate SSE with respect to all these k = 1 parameters I will get to a set of k + 1 equations.

And I will put that equations to 0, equalize it to 0, so that I can get the value of the parameters b 1, b 2. So, now this is a system of linear equations. The system of linear equation you can solve by any method. So, you can solve it by any method and then by solving that and you can find out the values of the b 0, b 1, b 2 which basically, once we find out the b 0, b 1, b 2 then we got the regression model. Basically the mathematical model that gives us the relationship between all the x to the y, x 1, x 2, x k, 2 y.

So, that is nothing the same thing whatever we learn for simple linear regression same thing it is for multiple linear equation. Similarly for multiple linear equations also we can do for quality of it for that we can again find out the r^2 which I am not discussing it here.

(Refer Slide Time: 11:26)



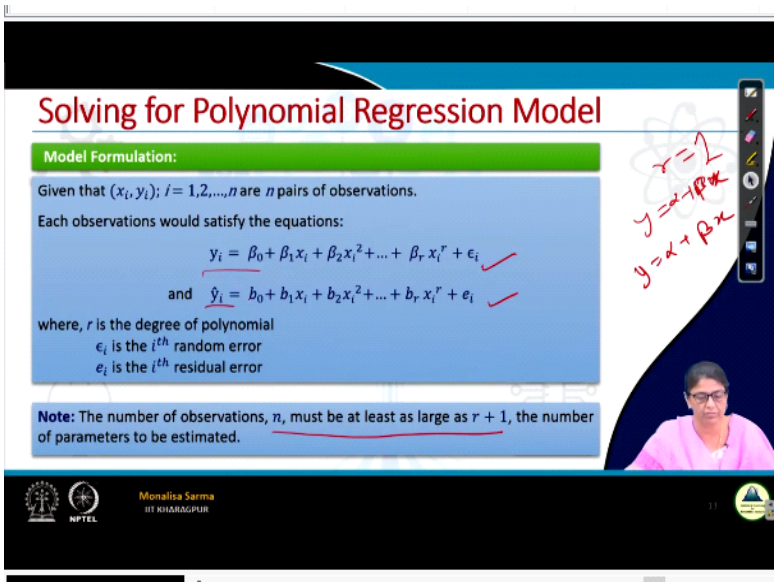
So, now non-linear regression model: So, non-linear when we called it is a non-linear regression model when the regression equation is in terms of r-degree, where r is greater than 1. Earlier y = α + β x, degree of x is what? Degree of x is 1. So, this is linear, we get a straight line. Now when the; regression equation is terms of r-degree, where r is greater than 1 then it is called non-linear regression model. So, x may be x^2, x^3, x 4 whatever it is.

Now it is simple non-linear. Now again multiple non-linear, so multiple non-linear when more than one independent variables are there then it is called multiple non-linear regression model, it is also termed as polynomial regression model. So, when there are more than one independent variable as well as degree is also more than one then we call it as a multiple non-linear regression model or we also call it a polynomial regression model.

So, in general it takes this form b 0, b 1 x + β 2 x^2 + + up to β r x r + ε. This ε is a random component that will always be there. Similarly the estimated response same way square methods, so, this is y cap we have estimated b 0 as the value of β 0, b 1 is the value of b 1 and β 1 b 2 is the

value of β 2. We have estimated this value then we try to find out the residual, how is the residual, summation of all residual square will get it SSE. So, that is y i - y cap.

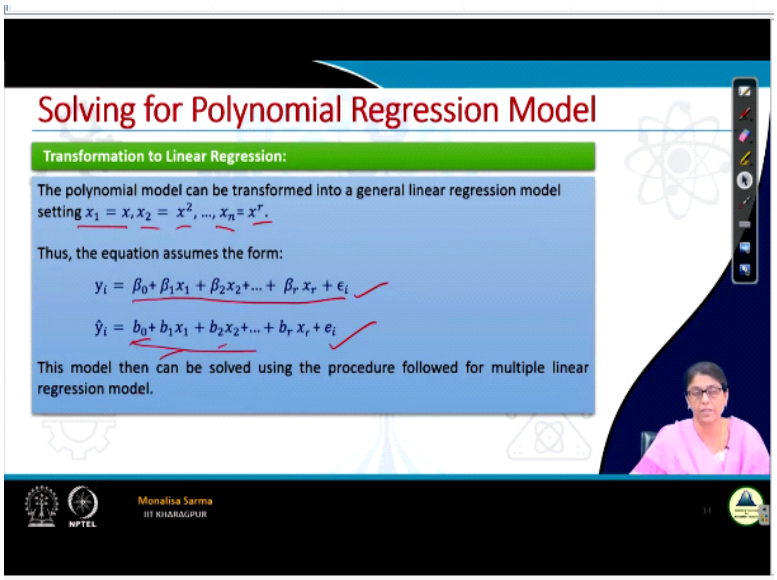
(Refer Slide Time: 13:02)



So, this is the y i, this is y i cap, so now here the number of observation n must be at least as large as r + 1. The number of parameters to be estimated, so, how many parameters we have to estimate? Total r + 1 parameters, when my r = 1 y = α + β x, here my r is 1, when r is 1 how many parameters I have to estimate? Two parameters is not it.

So, total how many equations I got? Two equations, so number of observations that n must be at least as large as r + 1. So, total I have to take number of observation means total data set I have to take at least as large as r + 1 the number of parameters to be estimated then only I will be able to estimate all the parameters. So, this is y i y i cap, same nothing difference here.

(Refer Slide Time: 14:16)



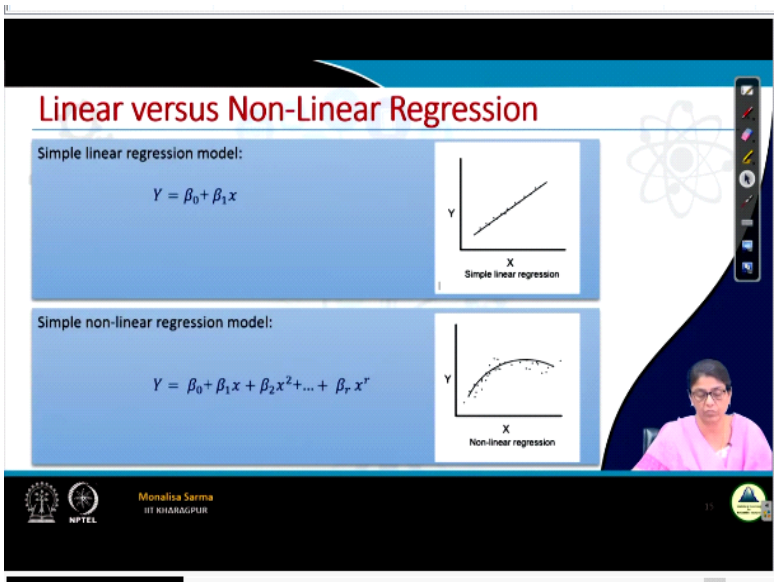
So, now this is in a polynomial form, non-linear form, is not it? Very much in a non-linear form. Nowadays nonlinear form it is difficult to solve it I can convert this non-linear to linear. How I can convert this nonlinear to linear? Say if I write x 1 = to x, x 2 = x ^{2}, x 1 = x r. I have used different variables x 1, x 2, x 3 are different variables to indicate the different powers of x.

X 1 = x, x 2 = x ^{2} then x n = x r = x to the power basically. So, initially this was my equation. So, this equations this b 0 + this equation β 0 + β 1 x i β 2 x i ^{2}. So, this equation I can write it in this form in a linear form. Similarly my y cap will be in this form. So, this model can be solved using the procedure followed by multiple linear. But with the same procedure what we have followed for multiple linear regression models because here we were getting a polynomial expression where our degree is more than 1.

So, if I want to estimate the parameters then it is becoming complicated. So, I want to solve it using the same method, the least square method estimate. So, that what I have done? This equations from linear non-linear I have made it linear, just that what I have done? I have used more number of variables, different variables I have used to indicate the different degrees. So, this expression is same as what I have used in my multiple linear regressions the same expression this too is not it? So, how I have solved it? I will find out the SSE then basically what is SSE is the summation of the residual square.

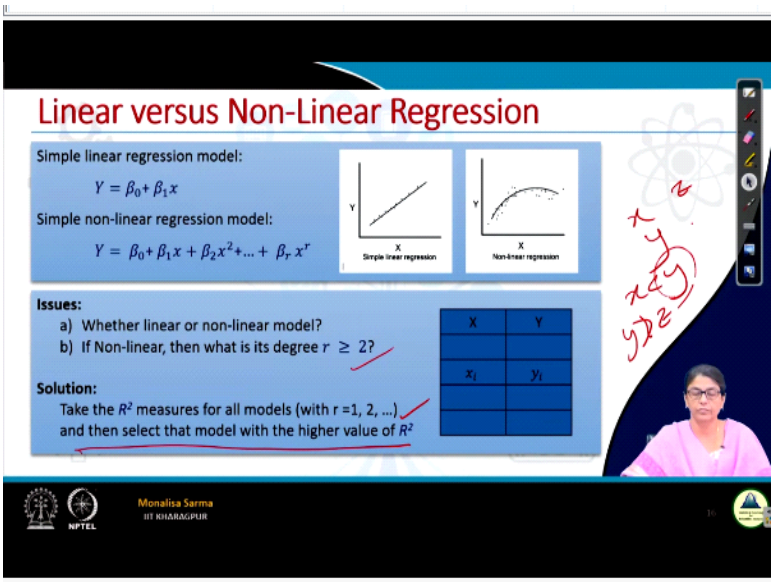
And then to find out the different values of this parameter what I will do? I will differentiate SSE with respect to all this parameter β 0 b and β 2 and all. Then based on that I will get the value for this and once I get the value for this I can find out the regression line. That is the regression model.

(Refer Slide Time: 16:33)



Clear; same thing just that things is getting a bit more complicated that is all the theory behind everything is same, the procedure is same, theory is same, everything is same. Just the calculation is becoming a bit more intensive. So, this is my linear regression model, this is my non-linear regression model.

(Refer Slide Time: 16:55)



Now the thing is that given a set of data x and y let us assume one parameter, one factor only. It is not multi-normal, this is just one what to say here we are using multiple nonlinear regression model, multiple non-linear models. So, let us not consider multiple, just consider it is just linear only, single only simple linear. So, here just one factor it is not multiple, it may be linear, it may be non-linear, but it is not multiple.

Suppose so and this is the case. Now the issue is given the different values of x and its corresponding y values are given, now which model we will try to fit? Whether we will try to fit linear or non-linear model where this y is map thing x is mapped to y based on a linear model or based on a non-linear model. So, if non-linear then what is his degree? Its degree is 2 or 3 or 4 what it is.

So, first what we will do? So, solution to this is because given this value of x and y we really do not know first it is really very difficult just if I try to plot the scatter plot also, from the scatter plot also visually sometimes very difficult to know whether it is the linear, it will fit a linear regression line or it will fit a non-linear regression line. So, if moreover even if you come to know that it will fit a non-linear regression line then what will be the degrees?

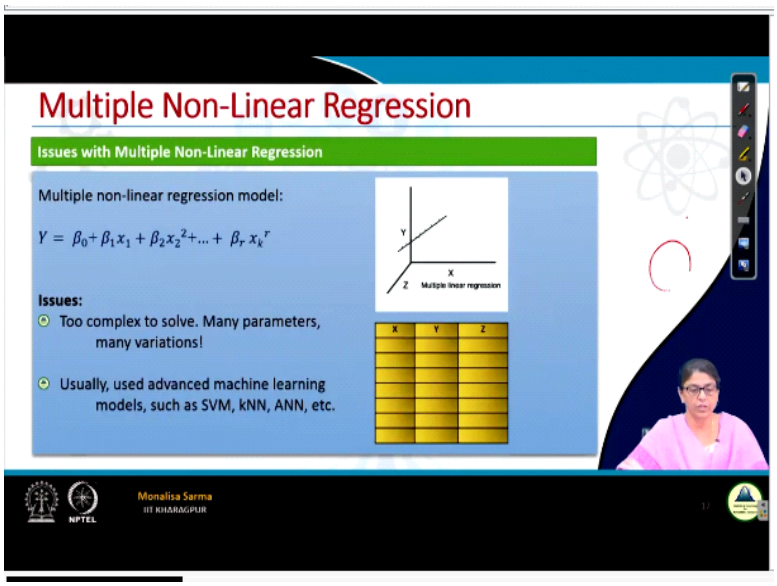
There will be 2 or 3 or 4 or what, so in that case what we will do? First we will take the r^2 measures for all the models. First we will try to find out the simple linear regression and will find out the r^2 model. That will be degree 1, then again I will find out for degrees equals to 2, that is the non-linear I will again find out the r^2 value.

Similarly I will again take r = 3 and I will find r^2 value and then select the model with the higher value of r^2. By taking a simple linear regression model if I get a particular r^2 value say x then I try to fit in a non-linear regression model where degree is 2, I get a value suppose y. Now my x is less than y. Then I am sure that it is not a linear regression model, it definitely it will fit a non-linear integration model.

Then I will again try it for r = 3 nonlinear fine. But let me just check for r = 3. For r = 3 I got z, however y is greater than z, then I need not to continue further then I know my regression is it is a

non-linear regression model whose degree is 2. That model will fit best for the given set of data action y.

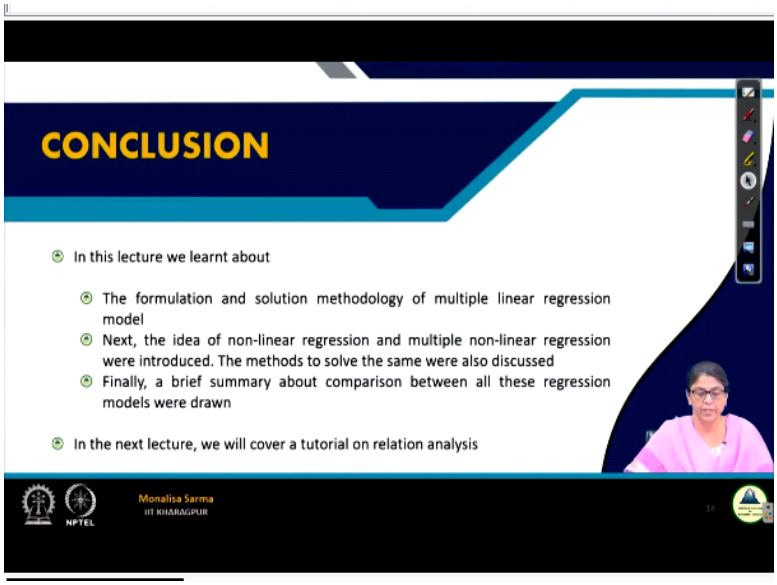
(Refer Slide Time: 19:59)



So, issues with multiple nonlinear regression model. So, what is the issue with multiple nonlinear? Too complex to solve, many parameters, many variations, so all this when we have multiple nonlinear even if we are choosing simple nonlinear regression model then also it is quite complex. But for multiple non-linear regression model it is too complex to solve we have many variations.

That is really not possible to solve using simple statistical matter. So, we use advanced machine learning models such as SVM, KNN, ANN etcetera which we will be discussing in coming few lectures.

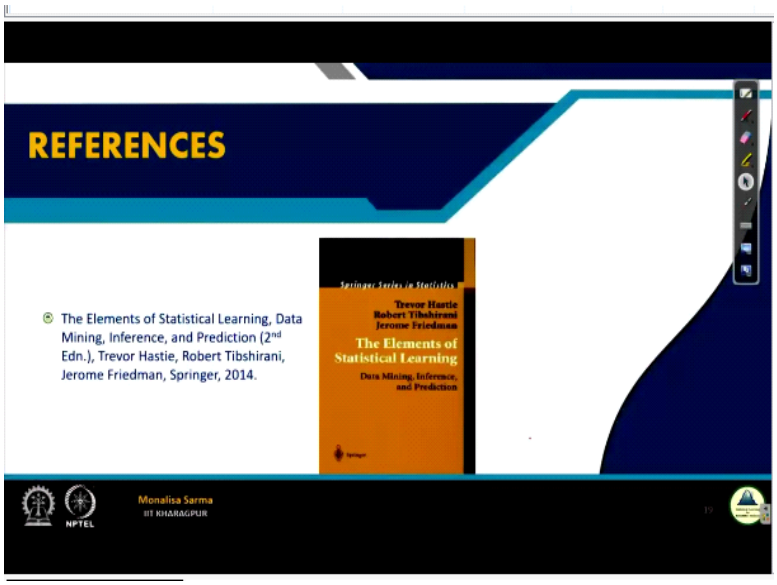
(Refer Slide Time: 20:36)



So, now to conclude this lecture; in this lecture we learn about the formulation and solution methods of multiple linear regression model. Next idea of non-linear regression and multiple non-linear equations were introduced, the methods to solve the same that also we have discussed and then we have given a brief summary about the completion between all these regression models.

So, in the next lecture basically we will cover a tutorial on the relation analysis, by relation analysis means I will cover a tutorial on correlation analysis and regression analysis still what we have learned till, auto regression is still not covered, logistic regulation is still not covered, till whatever we have learned. So, correlation analysis and regression analysis everything put together will take a tutorial in my next lecture. Thank you, guys.

(Refer Slide Time: 21:22)



So, this is the reference, thank you.

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

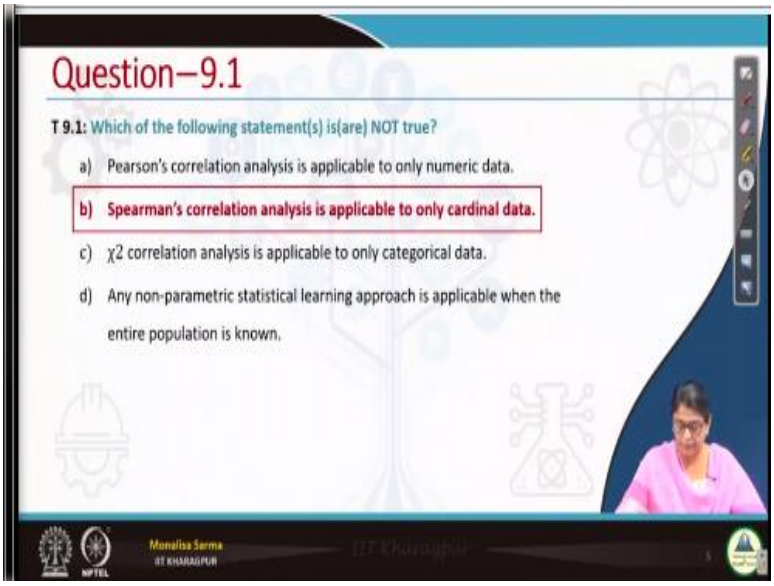
Lecture-43
Tutorial on Relation Analysis

Hi guys, so today is the tutorial session, we have done a couple of lectures.

(Refer Slide Time: 00:33)

So, we have learned like correlation analysis, different types of correlation analysis we have done regression analysis, so now it is time for tutorial. So, in this tutorial as like my previous tutorial first we will be doing few objective questions and then we will be solving a bit bigger problems. So, now coming to the objective type question, my suggestion is that please do not see the answer immediately.

(Refer Slide Time: 00:55)

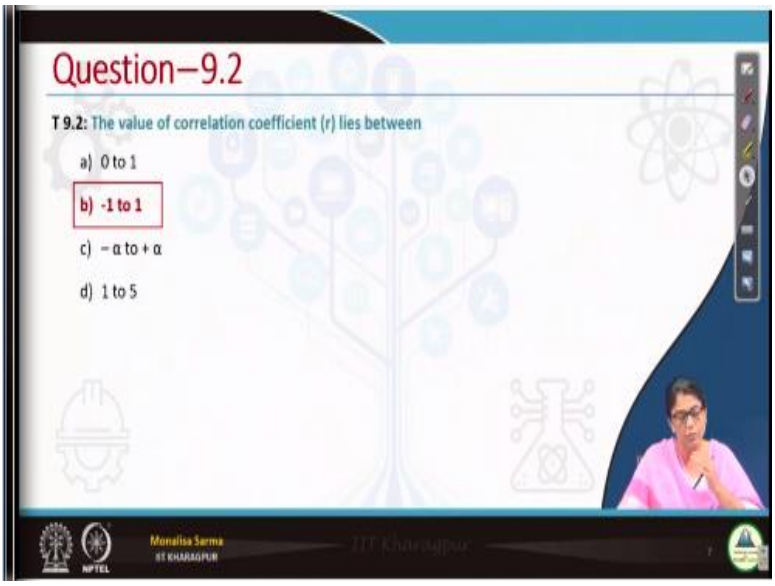


First try to think and then if you cannot then only you go to the answer, that is a very easy questions. First is which of the following statements are not true? So, Pearson correlation analysis is applicable to only numeric data, yeah, that is correct right, Pearson coordination analysis applicable to numeric data. Spearman correlation analysis applicable to only cardinal data, no, what is cardinal data?

Cardinal data is the numbers which we use to count numbers 2, 4, 5 those are where we used to count some numbers we call it as a cardinal number. Then Chi square test is applicable to categorical data, yeah of course chi square test is used for nominal data that is categorical data. And any non-parametric statistical learning approach is applicable when the entire population is known, very true and the entire population is known or we have a very big sample size unlike the parametric case.

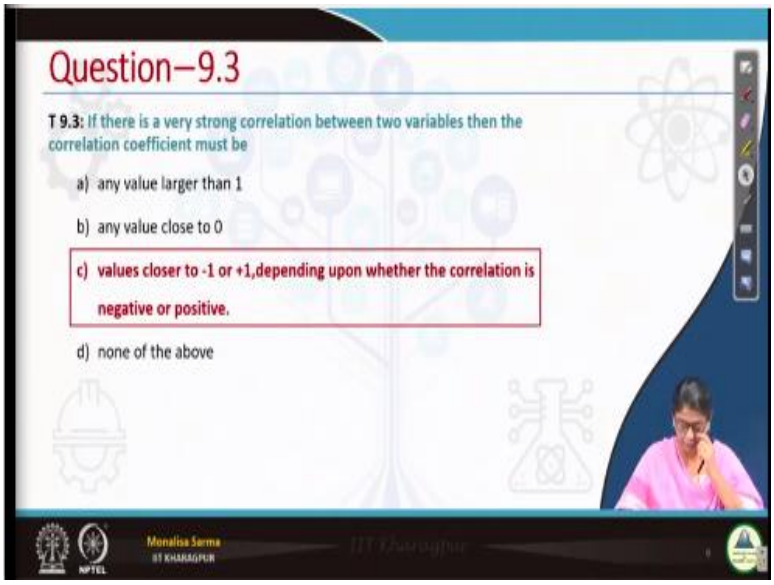
So, here only one statement is wrong that is b, Spearman correlation is analysis is applicable to only cardinal data, Spearman correlation lines is applicable to what sort of data? It is data which where there is an intrinsic ranking among the data, so that is ordinal data. Spearman of course we can use for numeric data also like for ratio, scale data for interface scale data also you can use it but then better is that we use for the ordinal scale data. And the numeric data when we try using Spearman concept we lose some of the information basically.

(Refer Slide Time: 02:24)



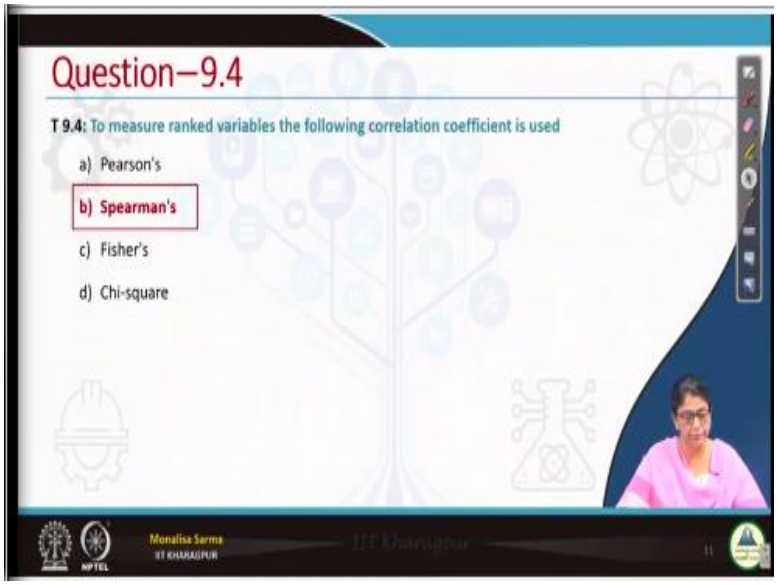
So, next question the value of the correlation coefficient lies between, the values of the correlation coefficient lies between what? Correlation coefficient, so it lies between -1 to +1, is not it. So, because correlation there can be negative type of correlations, there can be positive correlation, is not it, so the value lies from -1 to +1.

(Refer Slide Time: 02:45)



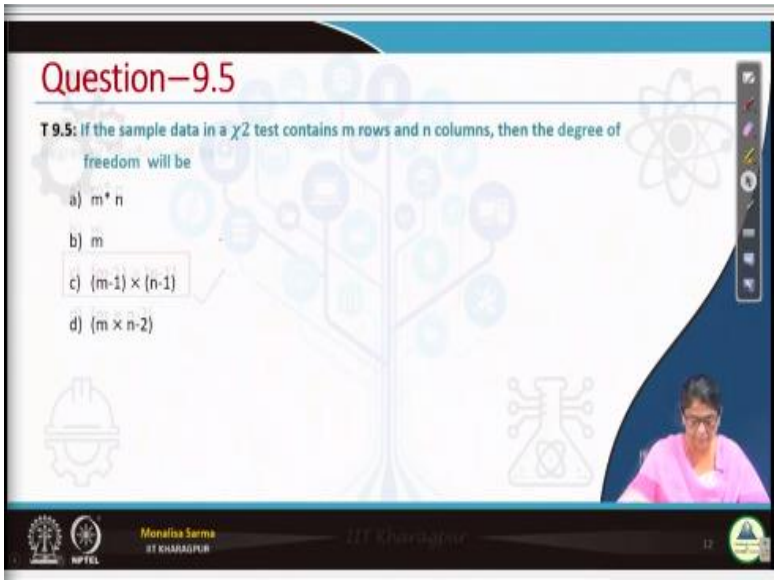
If there is a strong correlation between 2 variables then the correlation coefficient must be strong correlation, when there is a strong correlation then it will be closer to either -1 or +1. If it is a negative correlation then it will be closer to -1, if it is a positive correlation then it will be closer to +1. So, for strong correlation when it is closer to -1 or +1, weak correlations closer to 0.

(Refer Slide Time: 03:12)



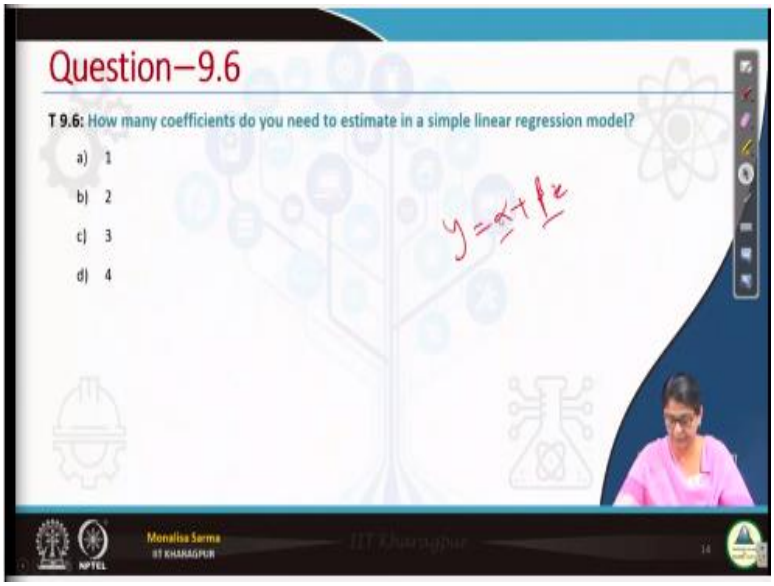
To measure rank variables the following correlation coefficient is used, when we have to find a correlation between the rank variables. So, rank variable is the ordinal data, ordinal data is a Spearman.

(Refer Slide Time: 03:27)



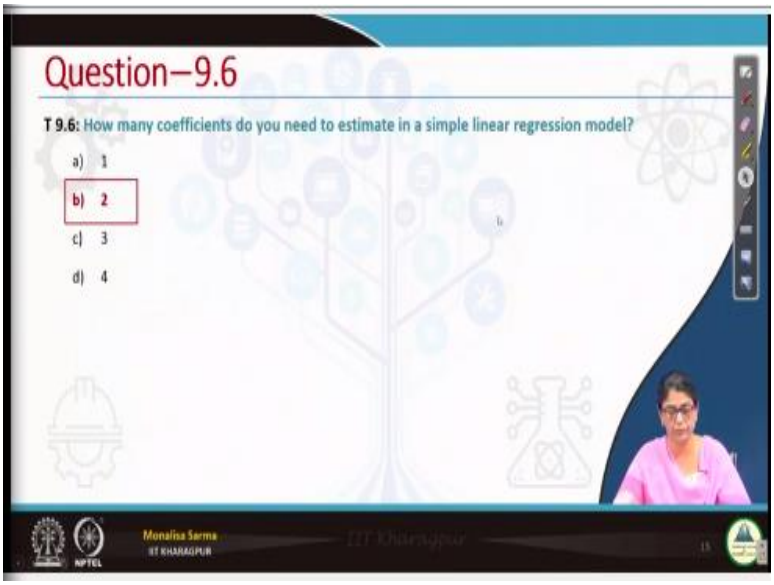
If the sample data in a Chi square test contains m rows and m columns, remember Chi square test we form a column where there are the independent variable we put it in the column and the dependent variable we put it in the row. And so it is we form a m cross n matrix, so then the degree of freedom what will be the degree of freedom? Degree of freedom is m -1 cross n -1. So, this is the degree of freedom.

(Refer Slide Time: 04:02)

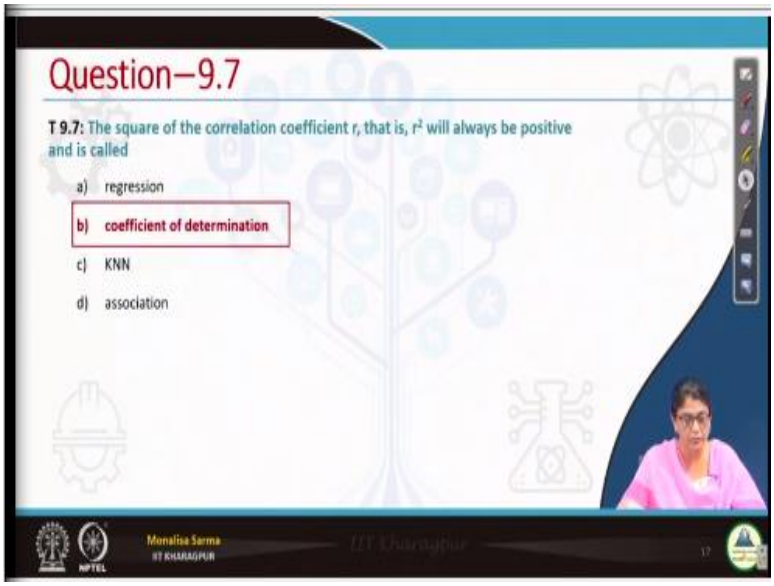


So, how many coefficients do you need to estimate an simple linear regression model? Simple linear regression model what is the expression for simple linear regression model? $y = \alpha + \beta x$, so what are the coefficient? Coefficient is α and β, for simple linear regression we need to estimate 2 coefficients.

(Refer Slide Time: 04:22)

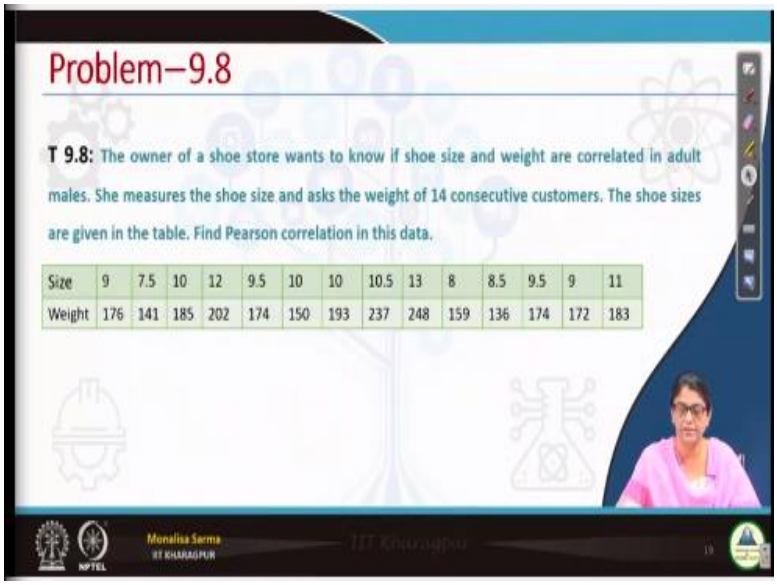


(Refer Slide Time: 04:24)



So, the square of the correlation coefficient r, that is r^2 will always be positive, square of the correlation coefficient r, we have done an r^2 test. Remember, what r^2 test the way we do, and what is r^2 called? R^2 is called coefficient of determination, is not it. So, now we will be doing some problems.

(Refer Slide Time: 04:53)

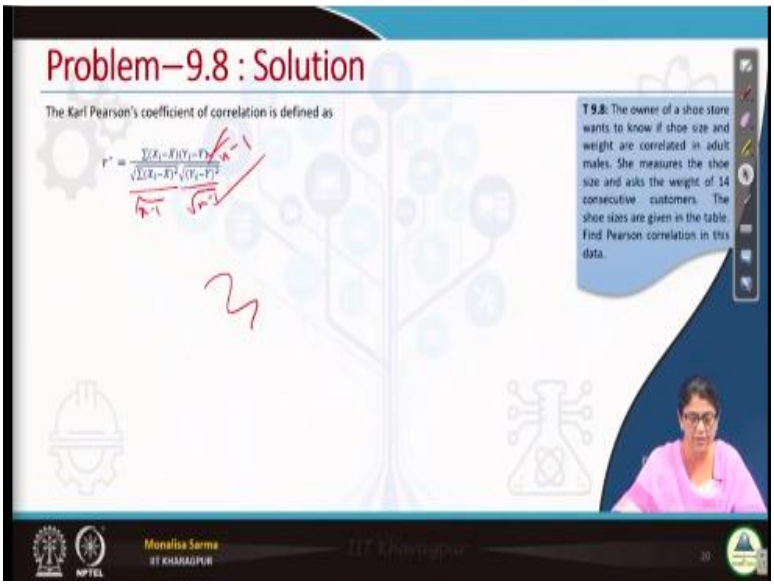


So very easy one just simple if you know the technique, you know the formula, you will be able to do it, nothing much to understand here as such just you have to understand what this and some problem it is the direct straight forward, it is directly given use this and some you will have to find out what we have to use and accordingly you should know the formula for that. Now here the owner of a shoe store wants to know if the shoe size and the weight are correlated in adult males.

If the size of the shoes and weights if there is there any correlation between them the owner wants to know. So, she measures the shoe size and asks the weight of 14 consecutive customers, for the 14 conservative customers she took the weight as well as the weight of the customers as well as the size of the shoe. The shoe sizes are given in the table; find Pearson correlation in this data.

So, it is just given you have to find the Pearson correlation. So, just if you can remember what is the formula for the Pearson correlation you do not have to do anything else just put it in a formula simple formula.

(Refer Slide Time: 05:55)

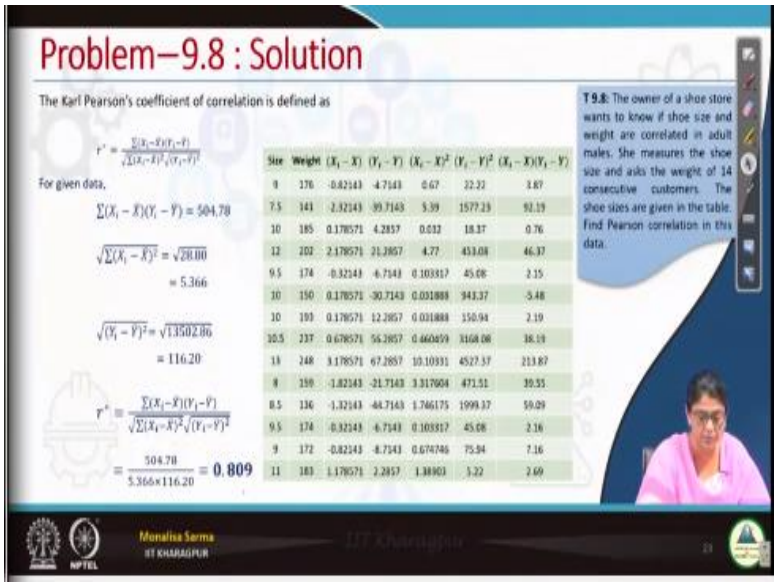


So, this is the Pearson correlation coefficient what is the formula? Covariance of X and Y divided by the standard deviation of X into standard deviation of Y, covariance of X and Y. When we write covariance of X and Y, this is the formula this numerator what we see and will be divided by the degrees of freedom. What is the degrees of freedom? Degrees of freedom is n -1, similarly standard deviation of X into standard deviation of Y.

So, standard deviation is this again under √ I will be having n -1 and here also I have √ n -1, here also I will have √ n -1, here I just have add n -1. So, √ n -1 into n -1 is n -1, this n -1 and n -1 gets cut, so what remains is this. So, this is the Carl Pearson coefficient is just this is the formula. If you cannot remember this formula very easy just remember covariance of X and Y divided by the

standard deviation of X into standard deviation of Y. So, now we what we need? We need X i, from the table will be getting X i value and Y i value.

(Refer Slide Time: 07:05)



And also we need X bar and Y bar, that is the mean of X and mean of Y, just put it in the formula you will get the r^2 value. So, this is the r^2 value, r value are this correlation coefficient that is 0.809. Now see here, from this value what can you tell whether the correlation is a strong correlation or weak correlation? This value is closer to 1, so that means it is a strong correlation.

Now again do you also remember one thing like this is the value, this correlation coefficient we got it, okay, fine but again we have to do the significance test also whether what this correlation coefficient what we got is from a sample. This correlation coefficient what we got from the sample, is it also applicable to the population or it is just by coincidence we got that.

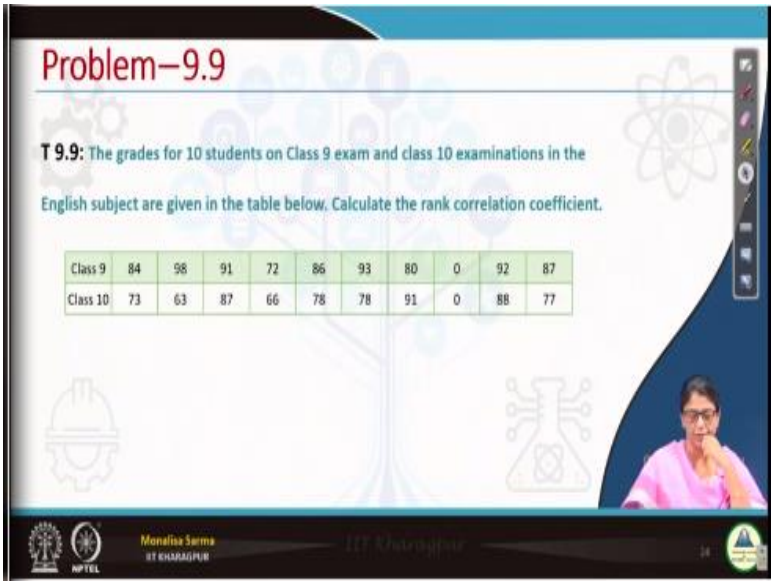
We have taken a random sample because samples are random; we have taken a random sample. In the random sample we got this correlation coefficient but actually it is not so in the case of a population. So, whether it is so whether this correlation coefficient what we got in the sample is it also applicable to the population, so we do the significance test, if you can remember that.

Significant test means we have find out the null hypothesis, we find that the alternate hypothesis but as a null hypothesis there is no correlation coefficient and alternate hypothesis **is** there is a

correlation coefficient. Now here for Carl Pearson we use the t statistics, so we compute the value of the t and for a particular significance level we see if the value falls in the critical regions then we reject the null hypothesis, remember.

So, anyway here we are not doing it just asking for the correlation coefficient, so with this, this is our answer, that is all. If it has asked to find out whether this is, see it passed the significance test or not then you will have to do that.

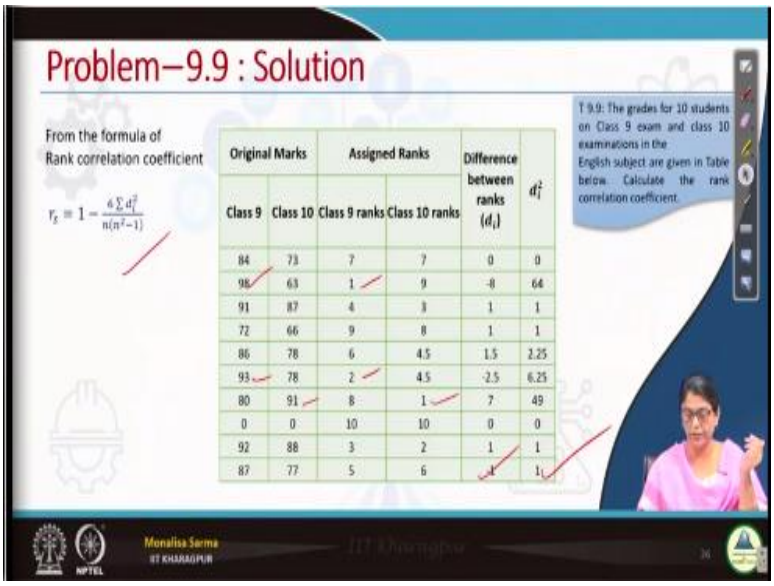
(Refer Slide Time: 08:56)



Now the grades of 10th students on class 9 exam and class 10 examination in the English subjects are given in the table below, calculate the rank correlation coefficient. Which is the rank correlation coefficient? Rank correlation means Spearman, Spearman we do it rank wise. So, different student's marks are given 10th student marks are given marks for class 9 exam and for class 10 exam, same set of students.

Same set of students they are given marks for class 9 and class 10. So, now we have to find out the rank correlation coefficient. So, actually this question looks sounds a bit weird, the same set of students for class 9 exams and class 10 exams. Take 2 different types of exam or the same subject instead of 9 and 10 they say for maths and science that way.

(Refer Slide Time: 09:51)



Original Marks		Assigned Ranks		Difference between ranks (d_i)	d_i^2
Class 9	Class 10	Class 9 ranks	Class 10 ranks		
84	73	7	7	0	0
98	63	1	9	-8	64
91	87	4	3	1	1
72	66	9	8	1	1
86	78	6	4.5	1.5	2.25
93	78	2	4.5	-2.5	6.25
80	91	8	1	7	49
0	0	10	10	0	0
92	88	3	2	1	1
87	77	5	6	-1	1

Then how to find out the rank correlation coefficient? Rank correlation coefficient is this is the formula. So, how do I find that d i ^{2}? d i is the difference in the rank of 2 numbers, how do we give the ranking? Ranking highest number gets the rank 1 and accordingly we come down, is not it. So, first we will have to find out which is the highest number and accordingly we give the rank 1, then rank 2, rank 3 and gradually.

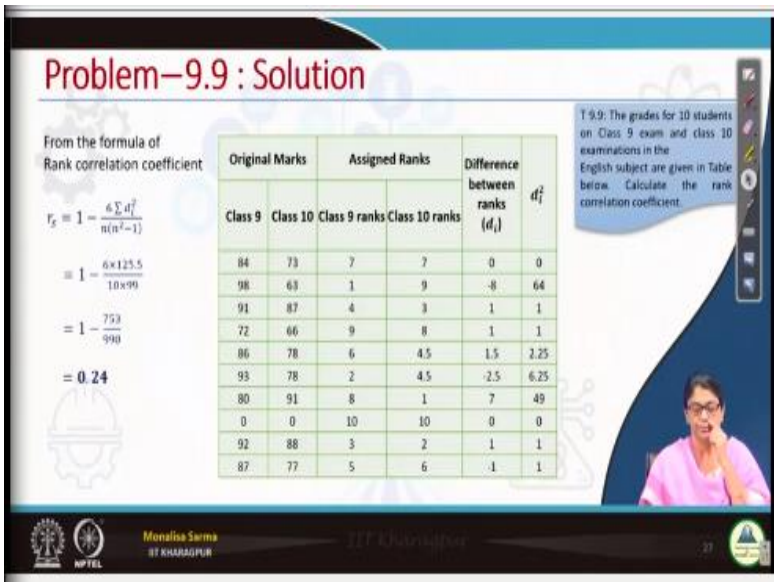
And then we will find out the difference of the rank of the 2 variables what we are trying to correlate. Is there any correlation between the class 9 marks and the class 10 mark? So, we will find out the difference of the ranks. So, once we find out that difference of the rank is d i and what is n? n is the total sample size, so that is how just putting it in the formula we will get the r s value. Now again for the significance days what I will do?

Here we will be using this Spearman correlation graph is there and for a significance level that I remember I told you in the class last in the one of the lecture when we have discussing Spearman correlation. It is available in the standard textbooks, the graph for Spearman correlation. So, from the graph if you see if it falls into a rejection region, then there is no correlation between the 2 variables.

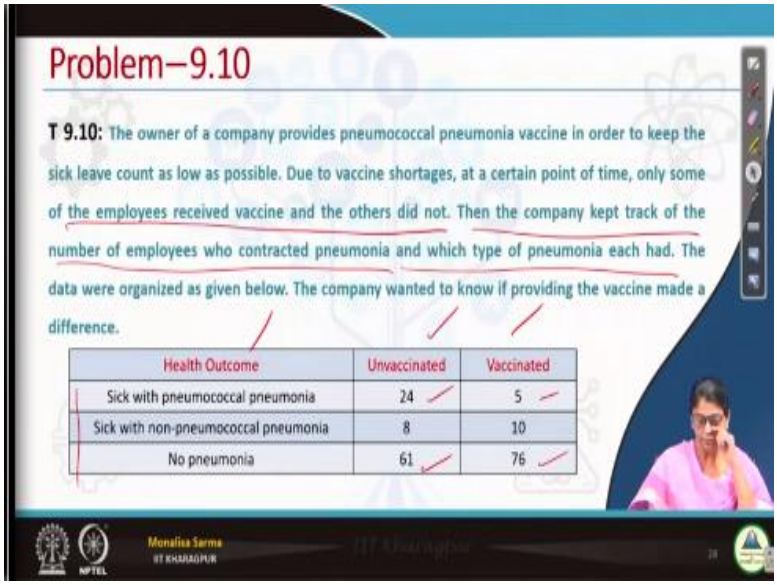
So, see here, say the class 9 this is 98 is the highest mark I have given the rank 1, after 98 it is 93 I have given rank 2. Similarly here 91 is the highest marks I have given rank 1, so likewise and

then we find d i, then I find d i ^{2}. Then accordingly put it in this formula and get the value of r s. Here it is the starts to find out the rank correlation coefficient, so we will not do the significance test if it asks. So, find out the significances then accordingly left to it.

(Refer Slide Time: 11:38)



(Refer Slide Time: 11:40)



Say here the correlation coefficient 0.24 it is a weak correlation basically. So, this is a very interesting question, see here. The owner of a company provides pneumococcal, pneumococcal is a type of pneumonia basically, pneumonia vaccine in order to keep the sick leave count as low as possible. Due to vaccine shortage at a certain point of time only some of the employees using vaccine and others did not.

Find as a vaccine like provide a vaccine sort of initially, so all did not get vaccine few got and few did not get. Then the company kept track of the number of employees who contracted pneumonia and which type of pneumonia each had. There are different types of pneumonia the vaccine is for pneumococcal pneumonia but there are different types of pneumonia. So, company is trying to keep track of the people who are contacting pneumonia and for what type?

Are they contacting the type for which the vaccine they have given the vaccine or they are contacting some other type of pneumonia. The data were organized as given below. So, the company wanted to know if providing the vaccine made a difference, what the company wanted to know? That providing a vaccine does it made a difference? That means by taking the vaccine does it have any effect on the occurrence of the pneumococcal pneumonia, does it have?

By taking vaccine that means we are having less number of pneumococcal pneumonia or we are having same having vaccine or not having vaccine we are having the same number of pneumococcal pneumonia? That is what we need to find out. So, this is very much equation of Chi square distribution c because these are the different categories and we have to find out the correlation of these 2 categories, what are the 2 categories here?

You have to find out the different categories are the for the independent variable if you see the different there is the vaccination status that is the unvaccinated and a vaccinated, does this vaccination status have any effect on the health outcome? Health outcome we have 3 different categories, fine, one is sick with pneumococcal pneumonia, sick with non pneumococcal pneumonia and no pneumonia.

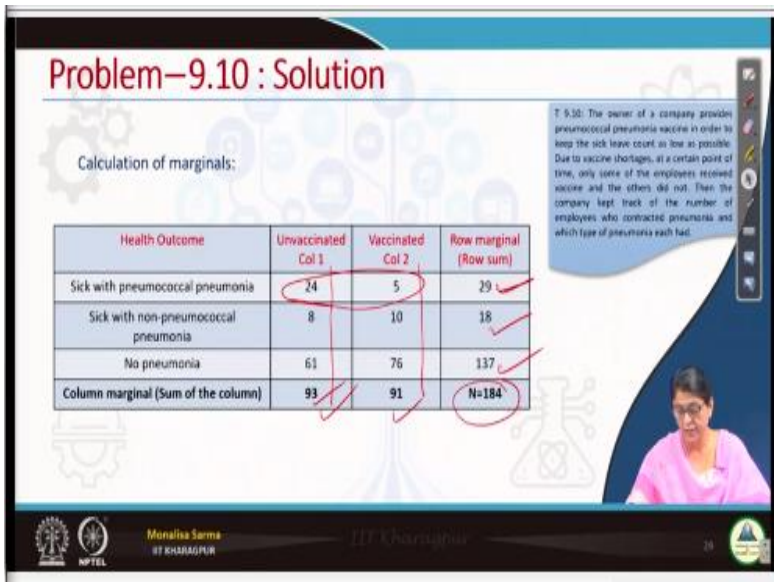
So, these are the 3 health in 3 different categories for the dependent variable and for the independent variable there are 2 different categories. So, this is very much equation of Chi square distribution that means we will have to find a Chi square correlation. So, we will have to find out the Chi square value of this whole table and then we will be able to tell it whether this providing the vaccine made a difference or it does not make a difference?

So, if that Chi square value that we get if it falls in the rejection region then providing a vaccine really makes a difference if it falls into acceptance regions that means providing vaccine does not make any difference. Basically there is no correlation between people contacting pneumococcal pneumonia with whether they are vaccinated or not vaccinated; there is no correlation between these 2.

So, first thing is that we will have to find out the Chi square value. Remember in this table I have already told you in the class the first thing is will fill in the table with observed value, these are the observed value 24, 5. 24 people caught pneumococcal pneumonia which are unvaccinated, 5 people caught pneumococcal pneumonia were vaccinated, similarly 8 non-pneumococcal, 10 here for vaccinated and this is no pneumonia 61 and no pneumonia for a vaccination is 76.

So, we need to find out the Chi-square value. So, for finding out the Chi-square value what is the first step? The first step is we will have to find out the row total and the column total if you can remember. If you do not remember I just say please go through the lecture once again.

(Refer Slide Time: 15:41)



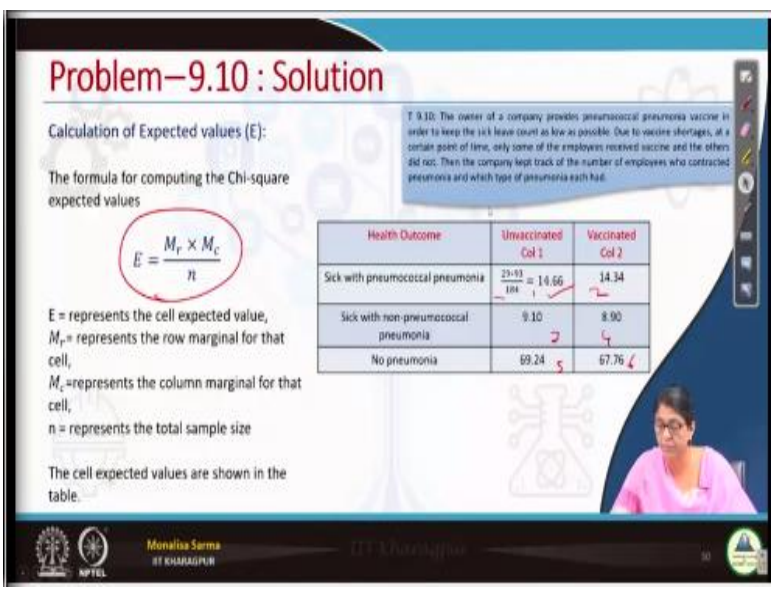
So, first thing is that we will be finding the row total and the column total. This row total and the column total we call it a marginal, so row marginal and column marginal, so this 2 put together is 29, this is the row margin of this row and then the row marginal of this row is 18, row marginal of

this row is 137. And similarly the column marginal of this is 93, column marginal of this is 91 this is sum and the total data is 184.

So, now first in the table we have the observed data and then from our task is to find out the Chi square statistics. To finding out the Chi square statistics first thing what we will have to do? Our first step is calculating the row marginal and column marginal. Then after we calculate the row marginal and column marginal then next step is to find out the expected value. Why we will find out the expected value? Expected value how we will find?

Well, the expected value is that value which will get considering that there is no effect on the vaccine. So, how do we find out the expected value? If the expected value is something meaning that there is no effect on the vaccine, vaccine or no vaccine what value we are expecting that people will be contacting a particular type of pneumonia or people will not be contacting.

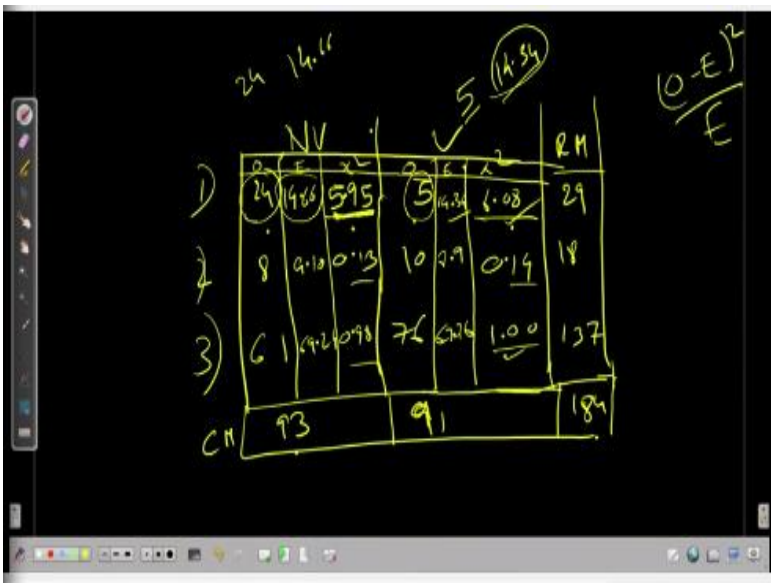
(Refer Slide Time: 17:00)



So, to find out the expected value for each cell this is the formula, row marginal into column marginal divided by the total number of values. So, now here for this case, this is 1 cell, this is the second cell, this is the third, this is the fourth, this is the fifth, this is the sixth cell, fine. So, now to find out the expected value for this cell how do I find out the expected value for this?

Row marginal into column marginal divided by the total number of values, you see what is the row marginal here 29 and column marginal here 93, 29 into 93 divided by 184, that is what we got it 29 into 93 divided by 184 that is how we got the expected value. Similarly we will have to calculate expected value for each column, let us do and see because.

(Refer Slide Time: 17:51)



See here this is the vaccination status that is I think this is the first one is non-vaccinated and second was is vaccinated and this side we have row we have the health status, another one is first let me just call it as a first category, second category and third category. First category is sick with that is pneumococcal pneumonia and second is sick with non pneumococcal pneumonia and third with it stayed healthy, that means they will did not have any pneumonia.

So, for that what was our observed value? Initial I will make the whole table that will be better actually. What was the observed value? Our observed value was initially this was 24 and this was 5, this was 8 and this was 10, this was 61 and this was 76 this is our observed value. First I will write the observed value, what is the observed value? 24 and then we have here 8, we have here 61 and here we have 5, then 10, then we have 76.

So, this total is 93 and this total is 91 and what we have this total here is 29 and here is how much it is 18 and here how much it is 137 and total we have is 184. So, this is the observed value, this is the row marginal and this is the column marginal. So, now once we get row marginal and column

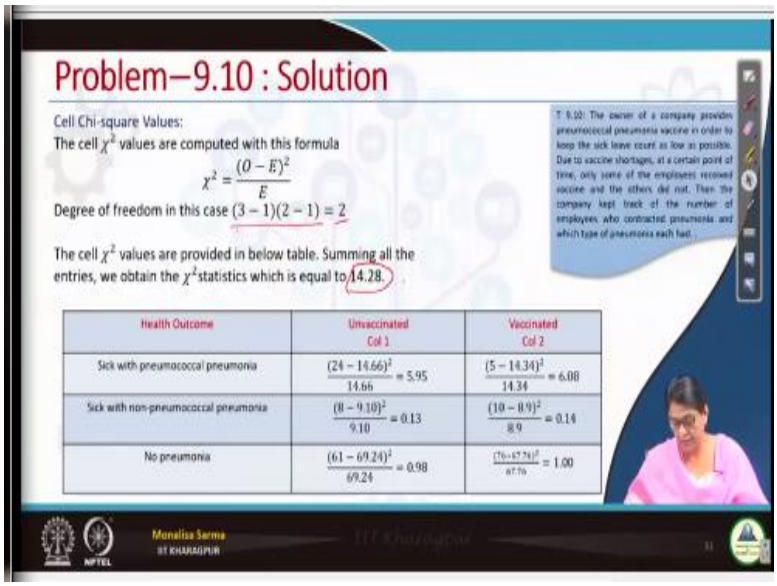
marginal, next we will find out the expected value, this is the observed value, next we will find out the expected value.

So, what is the expected value for this? Expected value for this is 29 into 93 divided by 184, 184 is the 4 total data. This if you calculate it you will get this 14.66 I am just putting the data you can calculate it afterwards, this we will get 9.10, this we will get 69.24 for this 5 you will get observed value is 14.34 this will get 8.90, this will get 67.76, fine, 67.76 this is the expected value.

Now once we calculate the expected value our next step is to calculate the Chi square value, how do you find out the Chi square value? $(O - E)^2$ by E, so this is the formula for the Chi square value $O - E^2$ by E, O is the observed value, E is the expected value. So, when you can calculate the X^2 value we get a X^2 value is 5.95 for this, this is the X^2 5.95 this we got 6.08, this we get 0.13, this we got 0.14, 0.98 and 1.00. So, these are the X^2 value.

So, now we have to find out the X^2 status, the X^2 status means sum of all this X^{-2} value. If we do the sum of all this X^2 value, so what we will get? Let us come back I will come back just to this table again. So, see here.

(Refer Slide Time: 22:03)

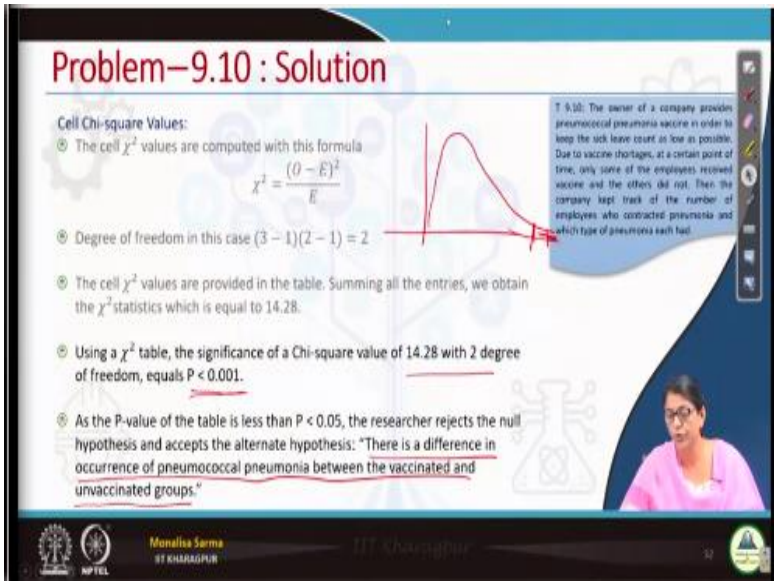


So, this is how we calculate the X^2 value. Once we get the X^2 value these are the X^2 values given here. So, the sum of all this X^2 value is 14.28, we will have to find out the sum of all X squared

that is my total X^2 value. So, total X^2 value I got is 14.28, now I will have to see it in the whether it is 14.28 values is a significant value or not, for that we will have to console the X^2 table.

We will console the X^2 table what is the degrees of freedom for that? How many rows and how many columns? Let us say m - 1 into n - 1. So, 3 - 1 into 2 - 1, so the degree of freedom in this case is 2. So, for 2 degrees of freedom for value 14.28 if you see the X^2 table what you will see?

(Refer Slide Time: 22:50)



The P value, the probability for that is basically less than **0.0001** 0.001 if you see the X^2, I am not showing the table here, I have showed it many, many times and I am sure by now all of you are in comfortable with that. So, if you console the X^2 table you will see for this 14.28 with 2 degrees of freedom it has a P value of 0.001, very less. So, and significance value usually we consider around 0.05, so 0.05 and this is less than that 0.001.

That means if we draw something here this is my significance level, if it falls in this area then I reject the null hypothesis and my value is falling here much lesser than the significance value. So, that means I reject the null hypothesis, that means the researcher rejects the null hypothesis and accept the alternate hypothesis, there is difference in occurrence of pneumococcal pneumonia between the vaccinated and unvaccinated group.

That means there is the vaccination, what was my earlier question, this vaccination does it have any effect on the occurrence of pneumococcal pneumonia? Yes, vaccination has an effect on the pneumococcal pneumonia in the people. Because null hypothesis does not have effect, alternative hypothesis it has an effect. Now this story does not end here, but that is one of the beauty of Chi square distribution I should say actually.

So, here from this okay we found out that, we rejected null hypothesis but we got a X^2 value by taking the sum of which cell? Actually which cell has contributed to this rejection? We can find it out, in earlier test there is no specific which data has contributed to this, now we can find out which cell has actually contributed to the rejection of the null hypothesis.

Now if we go back to our table, see whichever column we get a X^2 value high very we got a X^2 value > 1 is in this and this, less dollar this is very much less than 1, very much less than 1, this is less than 1 and this is just 1. For this 2 value we got a X^2 value which is very much bigger than 1. What does the highest value of X^2 is 6.08, why did I get a highest value of such a highest value?

Because you see, here my expected value considering that the vaccine has no effect, my expected value was 14.34. I expected if vaccine did not have any effect then I expected that the people will have this type of pneumonia around 14.3 but how many people had? Only 5 people had, so by chance this cannot be a random event, there has to be is a significant value, where expected is 14.34 and people who got is very 5.

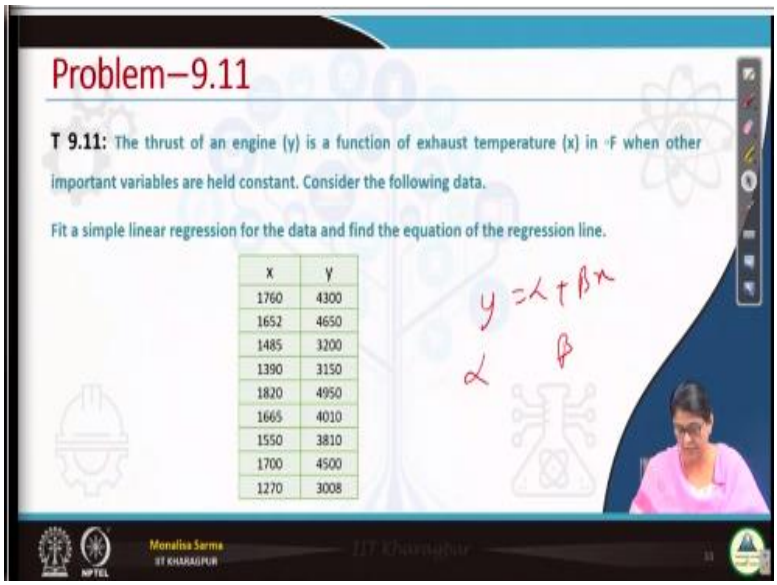
That means vaccine had an effect that is why we got such a less value, whereas if the vaccine did not have an effect the expected value is 14.34. Similarly next highest value is here X^2 value 5.95. Here see our expected value is 14.66 and how many observed value is 24, this also it cannot occur by chance. So, if the vaccine did not have any effect the expected value is we are expecting people around 14.66 people will be having the pneumonia that particular type of pneumonia before which the vaccine has been developed.

But the number of people who actually had that pneumonia is 24, so this cannot be a value by chance, we cannot just consider it a random event. So, that means the vaccine has an effect that is

why people who has taken vaccine, very less people got the disease whereas expected is so high. And similarly here because vaccine have an effect say these people since they did not take any vaccine, so though the expected value was 14 but 24 people had the disease.

And rest all the values it is very less, lesser than 1, just 1 value is just 1, so if this other cells has did not contribute to draw a decision, the only factor which contributed to our decision is this cell only the cell 1 and cell 2. So, interesting, right, this is how we find a Chi square distribution. We can not only be able to tell whether there are some sort of correlation or not, we can also specifically point out this presence of this correlation is for which category of data basically. So, this is by correlation using Chi square distribution.

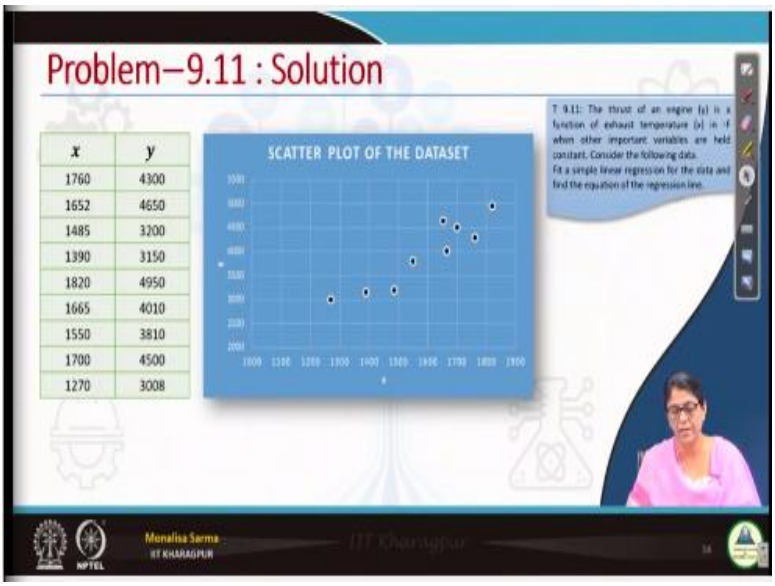
(Refer Slide Time: 27:31)



So, quickly we will go to the other problem. Now this the thrust of an engine is a function of exhaust temperature Y is a function of X basically in degree Fahrenheit, when other important variables are held constant, consider the following data. Fit a simple linear regression for the data and find the equation of the regression line. So, we have just asked to fit a simple linear regression for the data and find the equation of the regression line.

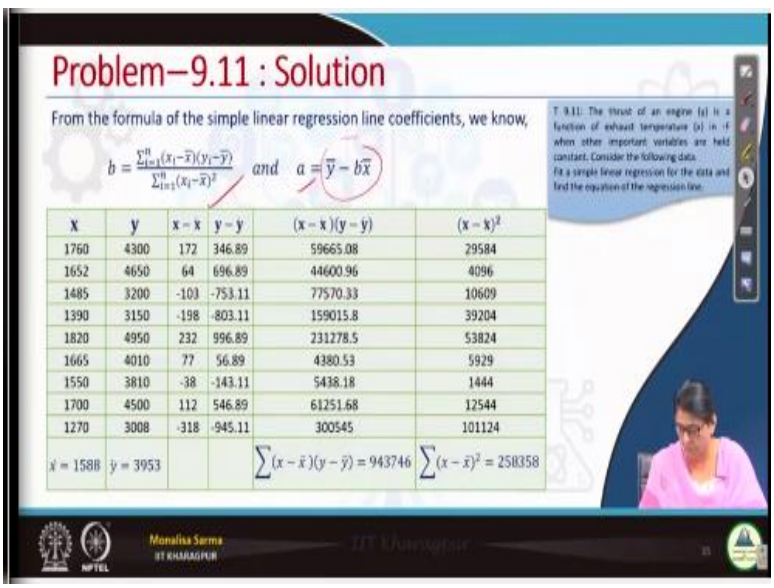
So, fitting the data to a simple linear regression that we do not have to do anything, just we know simple linear equation expression is $y = \alpha + \beta x$. So, we know what is α, we know what is β, we have already calculated using the least square estimation method, is not it.

(Refer Slide Time: 28:12)



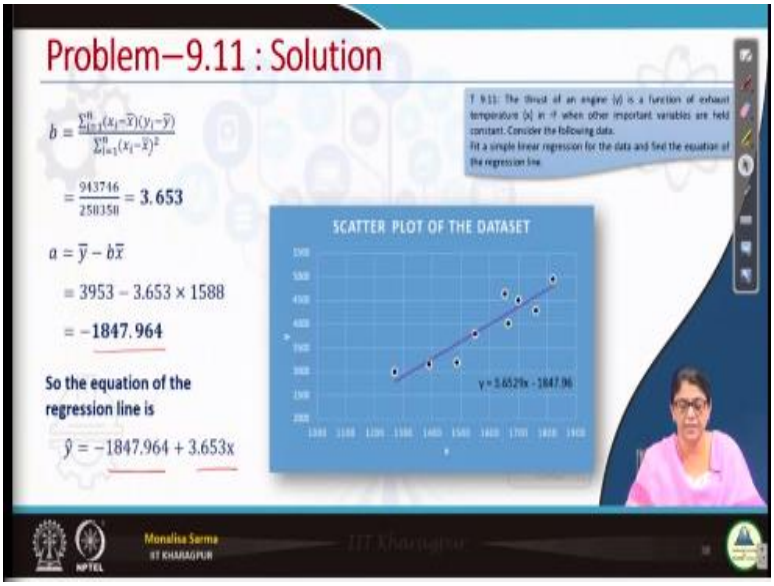
So, let us set the data we can just use the scatter plot just to see.

(Refer Slide Time: 28:16)



And what α and β that value which we have assumed, for β we have assumed b, for α we have assumed a, these are the values we have calculated using least square estimation method. We have done it I think in the last class only, we have last to last lecture maybe. So, this is the parameter b, this is the parameter a, so we have here what we need? We need x i, x bar, y i and y bar, from the data we can find out x i, x bar, y i, y bar is nothing so great. So, we found out the data, we found out the a value, we found out the b value.

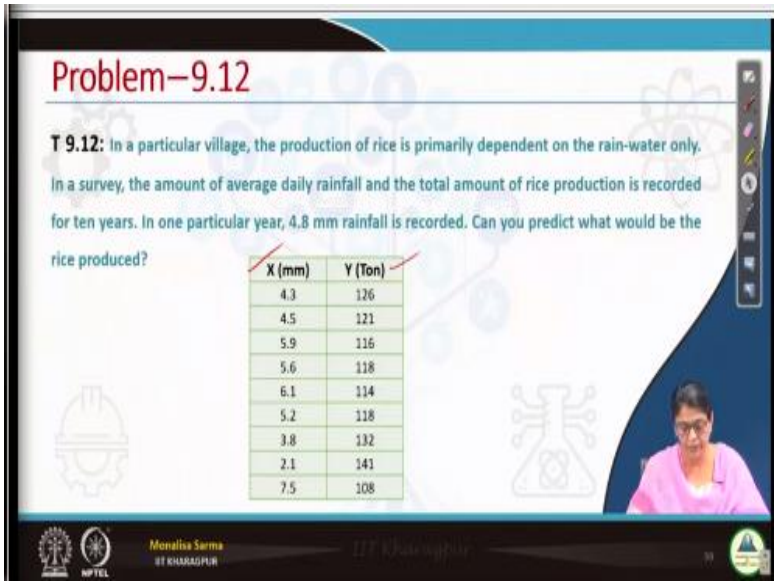
(Refer Slide Time: 28:46)



Once we found out a value and the b below our equation is nothing but this whatever this is c, our this is our a value, a value + b into x, this is my expression. So, this is my linear equation regression line, cool. Just we need to remember what is b, what is a, if you do not want to do the whole process of least square estimation then the minimizing the least square estimation how do you minimize it by doing first order derivation, partial half first order derivation with respect to each parameter 1.

Then if you equalizing it to 0 and then finding out the value instead of if you want to do it, do it great, if you do not want to do it just still have to remember this formula. From there only we got this a and b value, there is no rocket science here.

(Refer Slide Time: 29:36)

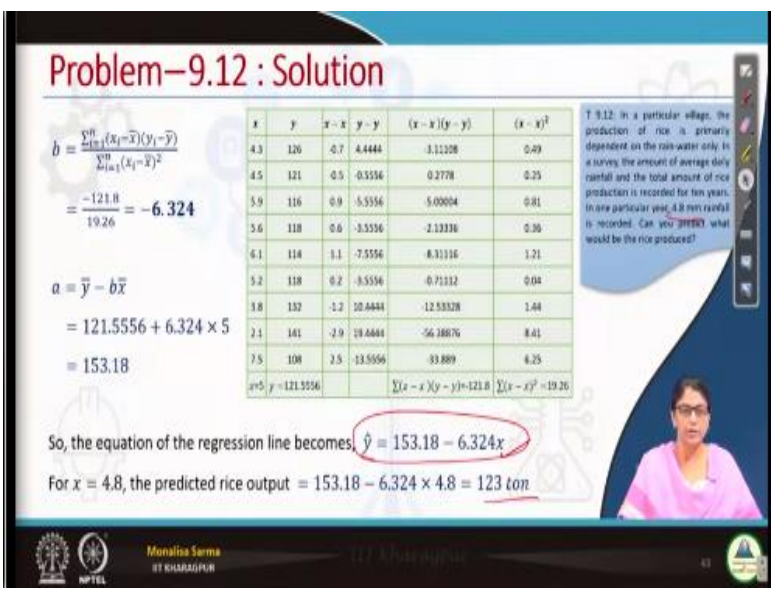


So, next question in a particular village the production of rice is primarily dependent on the rain water only. In a survey the amount of average daily rainfall and the total amount of rice production is written, recorded. So, this is the rainfall and is the rice production. In one particular year 4.8 mm rainfall is recorded. In a particular year 4.8 mm rainfall is recorded, can you predict what would be the rice produced?

Simple question of regression analysis is not it. Given the value of the independent variable we will have to find out the value of the response. So, how we will do that? For that we will have to need to know the function, that means we will have to find out the regression line that is y = α + β x. That means we need to know the value of the parameter, that is α and β, how we will know the value of a parameter?

That the way we have done in the last example, same way find out the value of the parameter. Once we know the value of the parameter we got the equations, once we got the equation in the equation put the value of the x and you will get the value of y, cool.

(Refer Slide Time: 30:39)



So, this is b, this is a, we calculated b and a, so this is our equation, expression this is the linear regression line. Now here given our the x value is 4.8 mm, so put 4.8 mm in x then what value you get for y you get 123 ton.

(Refer Slide Time: 31:03)

So, that is all, it is a very simple tutorial, is not it, so this is the reference and thank you guys.

Transcriber's Name: Prabhavathi
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-44
Auto-Regression Analysis

Hello guys. So, today in this module on relation analysis, on the module on relation analysis, we have basically learned correlation analysis, we have learned relation analysis, I means regression basically, we have learned correlation analysis, we have learned regression analysis. Today we will be learning a special type of regression analysis that is called auto regression analysis. (refer time: 00:50)

Now the question is why it is called special. Why I told that it is a special? It is not all special, but why I told it is a special type of regression analysis. That is basically this is here a regression analysis will be doing on a different type of data. What is the different type of data will be doing this regression analysis and the time-series data okay. So, now what is time-series data I will come to it shortly.

So, and now when we do correlation on this time-series data we do not call it the simple correlation analysis, we call it autocorrelation. Similarly when we do regression analysis on time-series data we just do not call it regression analysis, we call it auto regression. So, basically in this lecture, we will be learning what is time-series data; we will see autocorrelation, auto regression and will also see some of the applications of this time-series data okay. So, now first what is time-series data? (refer time: 01:45)

First let us understand what is time-series data. Basically okay I will go to the figure. First if we consider time-series data it is like when we tell data what is the data? It is data means we collect we collect data on some elements or some on some variables like rainfall. So, it is a data we this is data of some particular variables right. So, like our date of birth and so just just if I write 19 say 1990 what does?.

It does not mean any meaning or just if I write 22nd July, it does not carry any meaning. What this data is of what? So, we collect data on some or some element or some ahh observation or some variable. So, when we collect data of some observations repeatedly we do repeated measurements, over a period of time when we collect data of some observations, over a period

of time we do repeated measurement then we call it a time-series data.

Like this repeated this repeated observation can be anything. Like all of you I think you know when we when I told you the time-series data like rainfall direction, where time-series data. So, yeah there are simple very time. some simple time-series data may be like ahh if I want to find out the sale of a particular product in a shop, sale in hourly basis. If I can calculate the sale of a particular product in hourly basis or daily basis, that also becomes a time-series data.

When I am Basically when I am collecting data of some observations through repeated measurement of time, then I call it a time-series data okay. So, now basically so and okay Now you got an idea what this time-series? This is an example of a time-series data what I have given here in this picture. See it is a rate of price inflation. Price inflation is also time-series data; because we are collecting the data of the price the inflation of the price repeatedly we are collecting this data.

Repeatedly we are calculating this, taking repeatedly we are measuring this data on a over a period of time or in a particular time period. So, here suppose we have calculated the price inflation, we have calculated the price information for a couple of years from 1960 to 2005. And 1960 means we did not just take the data of 1960, if you can zoom it further then you will see we have this inflation data for each day of 1960, similarly each day of 1965.

We have calculated this. We have measured this inflation data for each day of 1970, each day of 60, 61, they base in 60, 61, 62 for each day. So, does if you zoom it further you may see it this way, but of course in this slide I could not keep it that way. So, this sort of data we call it a when we are repeatedly measuring this observation over a time period, we call such type of data we call it a time-series data.

So, here in the the vertical axis, what is that here we are taking the percentage. So, it can be in a different format. Now here we are taking the percentage. What is the rate of price inflation for each year? So, it is here we are giving in percentage. This percentage may be it may be negative inflation, it may be positive inflation and we get in percentage okay. (refer time: 04:58)

So, some other examples of time-series data like okay aggregate consumption and GDP for a country, for example 20 years of quarterly observatories. Each year we take 4, 4 observatories, 4 observations means after we take observation for each 3 months. So, how many total observations we will get? For 20 years, we will get 80 observations. This is also in time-series data.

What to say We are collecting data of this aggregate consumption and GTP and we are doing sort of repeated measurements for this, repeated measurements so, how much how much repetition is that each year we are calculating it 4 times, after each 3 months we are taking the value. So, this is also a time-series data. Then again Yen per Pound, Yen per Dollar, Pound per Dollar, Euro per Dollar exchange rates.

All this exchange rate if we take daily data for 1 year this is also a time-series data. This may be if you just take it for one it may be very small time-series data. Usually time-series data we consider here for many years. Then cigarette consumption per capita in a state, by year for many number of years. This is also time-series data. Rainfall data, rainfall data is a very good example of over time-series data.

Rainfall data over a year or a period of years, this is an example says when time-series data on does not need to be a very great thing it can be very small thing also very small concepts like like sales of tea. Sales of tea from a tea shop in a season that can also be also time-series data. If you consider the sale of tea each hour or sorry let us not take each hour if we consider a sales of tea each day in a particular season. That can be also a time-series data okay.

Not okay These are the examples of time-series data. Can you give me an example, which is a which is which we do this repeated measurement, but we which I will not take repeated I should not say it is a repeated measurement, but ahh data which varies over time. But then we do not call it as a time-series data, can you just give such example? okay ah okay Once this example is video data, audio data these are not time-series data okay. So, why they are not times this data? okay I will come to that later OK. (refer time: 07:16)

Which are the following graph is due to time following graph is due to the time-series data here. They are given 4 graphs as this graphs they represent their time-series data, well unfortunately no. Non or the graph represent the time-series ticket data, because here we see is 200, 400, 600, it does not seem to be like a periodic data. Means we are taking the where it does not specify that it is we are taking measurement repeated measurement over a period of time, none of the data. Here also is 1, 4, 7, 10 same. So, none of this figure is that it is represents a time-series data okay. (refer time: 07:57)

okay So, now this time-series data it has many effects it has sorry it has many uses. It is a very, very useful data I am like we can use it in many different things. First one of the very good uses of time-series data are which almost all of us know is that as forecast model. How can we focus the ah whether it will rain, how much it will rain next year. Before the monsoon comes people forecast that it is not general people that people who deals with this they really focus, they say

how will be the monsoon this time.

So, ahh so this this type of forecasting model what will be the rate of inflation in the next year. This All this this forecasting model it is the one some it is a really very good application of time-series data. So, basically how it is forecasting means we are basically we are predicting. So, predicting remains us we are doing regression linear regression. So, this is when we talk of focusing it is nothing but linear regression only.

We are The thing given an independent variable, based on the independent variable we are giving the value of a dependent variable. Here the the what is the dependent variable, dependent variable? is a In this stage dependent variable is a inflation in the next year. And what is the independent variable? Independent variable the inflation rate of many years down the line maybe.

So, this is one of the uses of time-series data that is the forecast model and to estimate dynamic causal effects, dynamic causal effects very good. So, like one good example is like this this mobile service provider. There are different mobile service provider, so now the this this service provided, they introduce different schemes, they introduce different what to say pro schemes and they sometimes reduce the rate, they sometimes increase the rate and because of this what how how it affects the customer?

Basically what is the customer chance, because of this different proposal a different scheme they have introduced? So, customer chance means I think you know like what is the percentage of customer living that mobile service provider. So, because by analyzing those it can find out the causal effect means, because of this ahh thing it is having this effect. Because of introducing this scheme people are leaving the service provider.

It is finding out the causal effect. So, this is time-series data also used to estimate the dynamic causal effect. Like again here given a good example, if the rate of interest increases what will be the effect on the rate of inflation and unemployment in 3 months, in 12 months? If the rate of interest increases, if the rate of interest increases then what happens people will stop expanding their business, people will stop opening new business.

And then what happens, if people become very conservative. Then what happens unemployment will increase unemployment will increase, inflation will increase is not it? So, basically how will you know? We will be able to predict by seeing the data of previous years. How this is happening having an effect on this. So, these are causal effects. But OK this is ahh This is from the perspective of a business perspective.

Now even we can also see it is a relation it is application and reliability analysis also. Reliability analysis, suppose given a system given a system and we have the data of system unavailability on a daily basis or on a monthly basis or weekly basis. System unavailability means when a system goes down it is obvious each and every system is at some point or other will go down is not it?

And suppose system is in continuous use and the system goes down, there has to be some repair action. When the repair action is done, then again the system will be up. So, when the system goes down till it is corrected, till it is repaired it is the system remains unavailable. So, this is the unable unavailability time. So, if we have this information what is the on an average what is the unable unavailability of in a day or in a week whatever it is.

And suppose I need to use this system for a very critical application, where my unavailability should be a particular value, it cannot exceed this particular value. Then I will be able to analyze this my past data and I will be able to tell whether this is suitable for that application or not. Because if the unavailability time increases, then definitely I will have to work something and maybe I will have to put more people on the repair team.

Or maybe I will have to use some advanced repairing actions. So, that my unavailability get reduced so there okay. Those are different management; my thing is that first if I want to find out ahh based on the unavailability data of past history I will be able to predict how much it will be system will be unavailable for how much period of time or maybe this is from the unavailability perspective.

Or maybe we want to similarly from the failure perspective also. If I know the daily basis the system fails how many times. So, that if I take the data for a year then basically I will be able to predict for the next year okay. So, these are some of the dynamic causal effects then time dependent analysis. Time depend analysis like rate of inflation, unemployment it can be observed over a time period. These are the different uses of the time-series data okay. (refer time: 13:28)

So, we have seen the uses of different uses of time-series data. Now how we will get this it is this use this application to get this application of time series, about data we need to we need some sort of modeling right. We will have to model this data then will be then only will be able to get those uses. So, first thing when I tell it is a forecasting model. Then definitely it is a regression model, will have to have a regression model, then only forecasting is nothing but prediction right.

So, if it is if we need if you want to use this time-series data as a forecasting model then we will have to basically find a regression model. Similarly if you want to find out the correlation over time. Correlation over time that means we will have to find an autocorrelation among this data. Autocorrelation it is also called serial correlation. We will have to find out the correlation among the, suppose I am I want to find out the correlation of this year data with the previous year data.

I want to find out the correlation of this year data and previous to previous years data. Basically I want to find a correlation between say 2020 data and 2022 data. I want to find a correlation between 2022 data and 2010 data. So, this way if I find a correlation, this is also called serial correlation. So, I have to find out, I will have to need the correlation model also. So, then again you can model to find out the dynamic causal effects also.

So, this for this different application basically we need to model this data. So, in this lecture we will not be discussing about the dynamic casual model, for dynamic causal effects, will be mainly our focus will be on the forecasting model. (refer time: 15:04) So, now ahh how to model it, like here since suppose there is a time-series data what it is given? Dollars per English pound, so this data ahh this time-series data it significance it signifies dollars per English pound.

So, if this is the value given to us, if this is the data this is the data basically. And this is means 2007 means not one value, basically 2007 means it is maybe data is collected on each day of 2007. So, that is why we can see it in a continuous form, because if we see the data as data is actually discrete. Here we are seeing it in a continuous form. Why because we are collecting the data for each day.

So, when we have this sort of data then can we predict the trend at a time say 2017. In 2017, what will be the value of the dollars per English pound can we predict the trend? Yes, of course we can predict this trend. We will be using regression, auto regression. We would not tell it is regression, we will tell as auto regression okay.

So, now to we need to understand what is auto regression? So, to understand auto regression basically, we will need to do lots of calculation. And this, what is calculation? We will need to understand autocorrelation. Basically, as I told you our focus here is auto auto regression, but for doing auto regression we need also to know auto correlation. So, we will have we will learn that also, but before that let us see some concepts and notation before we go to those modeling in details. (refer time: 16:33)

So, what are the some notation, when we tell Y t is the value of Y in a period of t. Say Y 2022 rainfall data, that is rainfall data. I know that is that if I use the variable Y that Y 2022. Now this Y t means they can be again is data set for each when that means there can be data for 365 days okay. So, data set this is the T observation on the time-series random variable Y. For 1 year, Y 1 to Y 365, again for the next year Y 1 to Y 365.

So, this is one notation. Then I see the example here Y t denotes the daily, weekly, monthly rainfall in year. More precisely, for the year 2021 say 365 days data. So, the rainfall data from 2011 to that is, for the last 11 years data okay. (refer time: 17:30) Then some assumption for time series, then we can call that the data is a time-series data, where it has to satisfy certain assumptions and what is that.

We consider only conjugative, evenly spaced observation. So, we consider for any forecasting model, for any correlation model our time-series data needs to be conjugated evenly spaced observation. For example, say monthly data in 2010 to 21 for each year and without any missing month. If you are considering monthly data from 2010 to 2021, then we should have data for each month from 2010 to 2021 each month of 2010, each month of 2011, each month of 2012 up to each month of 2021.

There should not be any missing month and there should not be any other data. For example, on daily basis for a year is admissible. We are taking monthly data. So, we do not need another there should not be any other data like daily data daily data. We It is not admissible. Then, another one property with this time-series data should follow; the time-series it is is stationary if it is probability distribution does not change over time.

Remember this audio data; video data I told it is not time-series data, because of this property. This time-series data audio data this is does not satisfy this stationary property. But that means what its probability distribution does not change over time. Then now means like say the rainfall data in 2020 we found it follows a normal distribution, in 2021 is found is follows an exponential distribution.

No, data may vary but then there is probability distribution it does not change over time okay. So, then we call it just stationary. A time-series Y t is stationary if its probability distribution does not change over time, that is, the joint distribution of this does not depend on i okay. It does not In rainfall data it does not depend on each year. In 1 year, we will see that rainfall follows a normal distribution, other exponent distribution other year some other distribution.

No, it is it follows a particular distribution then only we can consider the time-series data. Then

only we can use it to use it for for modeling those data. If the problem if the data is not stationary, we cannot model it we will not be able to predict it properly. So, stationary property implies that history is relevant. In other words, stationary requires the future to be like past in a probabilistic sense of course okay.

Auto regression analysis assumes that Y t as both uniform and stationary. So, if you are interested in doing the regression analysis, we are assuming that Y t is both uniform and Y t is stationary. If With this assumption only we can carry out the auto regression analysis. Else if, it is not stationary and if it is not uniform, we will not be able to carry out auto regression analysis, the result will not be correct okay. (refer time: 20:29)

So, now there are 4 ways to have the time-series data for auto regression analysis. For auto regression analysis there are these are the different 4 ways. How our data will be. One is we call it lag. The first leg of Y t is Y t 1. The rainfall data of 2022, it is first lag is rainfall data of 2021. It is second lag, rainfall data of 2022, it is second lag is rainfall data of 2020, that is, we call it lag. The first leg of Y t is Y t 1, it is jth lag is Y t - j okay.

And tth lag is 2022 - 5 that is the jth lag. Then we have a term called difference. The first difference of a series, Y t is a change between period t and t 1. This is how we call it a difference. Log difference sometimes the data changes abruptly. So, it is very difficult to just to take the data. So, in those cases we use the log of it. Then we call it a log difference okay. Then again similarly we can have also percentage. Like in the inflation rate, we have used percentage. So, this is the percentage okay.

So, now we will see autocorrelations. (refer time: 21:48) The correlation of a series with its own lagged value is called autocorrelation, also called series correlation. Remember we have to use the term series correlations, earlier when I used when I told ahh this modeling of time-series data, so this is a series correlation. When we are doing correlation of a series with its own lag value, 2021 rainfall data I am doing correlation with two its own lag value it is for 1 year lag that means, with 2022 I am I want to predict the rainfall of 2023 maybe.

And I am sorry we are not predicting this correlation. I want to find out the correlation between 2022 rainfall data with 2021 rainfall data or I want to find out the what to say ah correlation between 2022 data with 2020 data okay. So that is called as the serial serial correlation. Here we use that and this lag value, this lag. So, formula for the jth auto correlations, this is the formula for the jth autocorrelation okay.

So, we found a covariance between Y t and Y of t j. So, remember how we found out the

correlation? Correlation is covariance of x i and y i divided by standard deviation of x into standard division of y is not it? So, same thing it is just what is x i here? X i is the time-series data Y t and ahh what is y i? Y i is this lag data. So, 2021 and 2020, this is 2021, it is 2020. It is just I am taking the example in the from the rainfall perspective.

So, it may be not in in terms of year, it may be in terms of months, whatever it is. So, this is how we found a jth auto correlation Okay. So, covariance of the Y t and Y t j is the jth auto covariance. We do not call it covariance and simple data we call it covariance. Here we call it auto covariance. (refer time: 23:57) So, how to find out the covariance between variable Y t and Y t 1, this is the formula.

Why it is x i = summation of i = 1 to n, because for that particular year or whatever it is there will be many data right. If you consider here maybe if we have day wise data so i i = 1 to 365 right. So, this is the this is how we find it auto covariance. (refer time: 24:26) So, now suppose the same example what we have seen the value of dollars per English pound. So, for is given that is suppose we have given a ρ 1 that means, between 2 consecutive years ρ 1 that is between two two consecutive 12 and 13, ah 14 and 15, 21 and 22.

So, for given data, ρ 1 is 0.84 between two given consecutive years. This implies that the we got this is the auto covariance value factors is ρ 1 is 0.84. This implies that a dollars per Pound is highly severely correlated. 0.84 is very near to 1, so this is highly correlated is not it? So, we call it highly serially correlated. If it is if its value is near to 0, then we will call it that is weekly serially correlated.

Similarly we can determine ρ 2, ρ 3 etcetera. P 2 means lag of 2 years, ρ 3 means lag of 3 years fine. So, now we will come to auto regression model. In auto regression model, basically why we have seen auto covariance, because we in auto regression model we will be to solve the to find in the regression model we will have to find the value of the parameters right.

So, once we find a value of the parameters then our job is done. We found a function. So, to find the value of the parameters, we will basically be using this auto covariance. This ρ, ρ 1, ρ 2, ρ 3 and all those stuff. (refer time: 25:53) So, auto-regression model and autoregressive model is used to model a future behaviour for a time-ordered data, using the data from past behaviours.

Concept of simple concept regression analysis we will be developing a model, which we call it ARM, AR model, autoregressive model. So, we will be developing AR model to model the future behaviour for a time order data using the data, from the past behaviours. Using the rainfall data

of past years will be able to predict the rainfall data of the next years. So, that is the autoregressive model.

Essentially it is a linear regression analysis, it is nothing but a linear regression analysis of a dependent variable using one or more variables in a given time-series data fine. So, one or more variables like if I want to find out predict the rainfall of 2022, I will be using 2021, 2018, 2019. If I use only 2021, this is 2022, if I use if I want to use 4 4 years, last 4 years 4 years down the line 2021, 2020, 19, 18 and 17. So, that is it. So, Y t is equal to function of this. How many years you want to use that is t - p okay. (refer time: 27:14)

So, a natural starting point for forecasting model is used to is to use past values of Y, that is, Y t 1, Y t - 2 to predict Y t. And auto-regression is a regression model in which Y t is regressed against it is own lagged values. So, we have already seen this. What is auto-regression model where, Y t is regressed against it is own lag values. What is the We call it independent variable is a regressor variable is not it.

So, Y t is here. Y t is the response. Y t is regressed against it is own lagged value. What are the regressors? Regressor is own lag value. The number of lags uses regressor is called the order of auto regression. To predict the rainfall of 2022, if I use 5 years before that, so that means, my order of auto-regression is 5. If I am using 2021 data 2021, 20, 2019, 18 and 17 then my order of auto-regression is 5 okay.

The number of lags used as regressor is called the order of auto-regression. In first order auto-regression denotes as AR 1, Y t is regressed against simple Y t - 1. In pth order auto regression denoted as AR p, Y t is regressed against Y t 1, Y t - 2 up to Y t p. fine So okay fine. We got regressive simple it is just a simple case of regression analysis. Regression analysis means we have to estimate the parameters. (refer time: 28:52)

So, in general pth order auto-regression model is defined as this way. This is how we define a pth order regression model. If it if we regret pth order means we will be regressing with the pth lag data. So, this is the intercept and then how many what to say parameter we have that depends on the order of the regression level. So, BHB is B 0 β 0, β 1 up to β p is called the auto-regression coefficient.

And epsilon t is the noise term or residue and in practice it is assumed to be Gaussian white noise. Epsilon t is just the noise term. It is assumed to be Gaussian white noise. What is Gaussian white noise? In simple if you see Gaussian white noise you do not have to go through such details. Just there is that this where its value is independent and its expected value of

epsilon t is 0, that is mean of epsilon t is 0. For your understanding that is sufficient. So, that is Gaussian white noise.

So, we have exp assumed that epsilon t is assumed to be Gaussian white noise. And this β 0, β 1 up to β p is the auto regression coefficient that means; we will have to find out the value of this coefficient. Now the task is to find out the best value. We have already seen in regression analysis our task is to find out the best value of this different parameters. How did you find out the best value for these different parameters?

Remember what we have done for linear multiple linear regressions how we have done? Simply we have to find out the sum of the squares error and then we have to differentiate first order differentiation. We get it to 0. And then we found out the value, that is what we have done it form. But here it is it will not be so easy, because the data is not so simple. (refer time: 30:39)

So, here actually basically just to find out the coefficient, one very efficient method has been very in general it is used, it is very popularly used it is called it is basically least square method. It is used under concept of really square method, same least square method what we have used in regression analysis. But least square method the way we do and where we find a expression how just we do it partial differentiation we find the expression to finding out the partial differentiation of age and finding the expression of this is a bit a bit complicated.

So, one very easy technique is Yule-Walker equation. It is called Yule-Walker equations. So, Yule-Walker equation basically it converts from the expression into ahh terms of this autocorrelation. This ρ1, ρ2, ρ3 is the, it converts this coefficient values, at it converts it shows a relationships with this coefficient values with this auto correlation values. So, Yule-Walker equation it is not possible to discuss at in this lecture, where it will be needing I will be needing a lots of board work.

And it will take from 2, 3 hours to explain the whole thing. I suggest you please Google it. You will be able to understand it and moreover if these things actually nowadays all software are available. Ready-made software are available and even you can write your own software using R so so, you just feed the data, you just select your technique and you will be able to solve it.

So, this is this Yule-Walker equation, just this is a this is the form where this we need to find out this values OK. P 1, ρ 2 are this is nothing but the correlations, ρ 1 is the correlation which just is next regress value. P 2 is this correlation from this value to 1 1 lakh okay. So, that way we we will having will have a system of equation, when we have a system of equation how to will be using we can use any numerical technique to ah what to say to ah solve this system of

equations.

Once we can solve the equation we will be finding out the values of β 1 to β p, if the order of auto-regression is p okay. So, now here I have used to solve for solving the equation many times I have suggested that you can use any numerical techniques. If By now if you have not learned a numerical technical (()) (33:00) some of the numerical technique methods which you really need it in many stage of your life will be needing I am sure there are some courses on that it is available and if it is not there it will be finding lots of material.

Please not only for this so in in-fact in many of the courses you will be needing how to solve a system of equations . Very not very simple equation, simple equation so you can easily solve it by if if not you can solve it by matrix method; you can solve it by Laplace method, but then if when you have to use numerical technique methods.

So, what are the different numerical techniques, how you can solve the equation. Please learn these techniques. I am sure some numerical technique methods courses are available. (refer time: 33:41) So, ρ i's are nothing but the ith correlation coefficient. So, from this equation we found out the value of β 1 to β p. And β 0 it can be chosen empirically, usually it is taken as 0.

Now we have this expression, we will solve this expression, we will find out the values of β 1, β 2 to β p. And once we found out the value we have the function. Once we have the function we will be able to predict given the what to say whatever order given 2 years data. Suppose that 2 years auto regression if given the 2 years data will be able to predict the third year data okay. (refer time: 34:15)

So, in this lecture we learned about the time-series data in what are the nature different natures of the time-series data and then we have also seen modeling with time-series data. We have seen auto-correlations and also we have seen auto-regression. In the next lecture, we will be learning logistic regression that is also an interesting topic okay. (refer time: 34:34) So, this is the reference. And thank you guys.

Transcriber's Name: Prabhavathi
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-45
Logistic Regression (Part-I)

Hello guys. So, in continuation of our discussion on relation analysis, today we will start a new topic that is logistic regression okay. (refer time: 00:34) So, like before going to discuss about the different concepts covered and all, I just want to bring out one example. Like, all of you I should not say all of you most of us have visited fair. Fair means mela right. We call it mela in hindi or in other languages also.

We have visit fair at some point or the other and there one very common, what is the activities there one gun is kept and some targets are kept and if we can shoot the target, then we will get some boonty right. We will get some awards, some whatever maybe it is. So, now take that example. There is some targets are there and a gun is there I have to shoot the target. So, I want to know that I want a model, which will tell me ahh like if I shoot from a particular distance, what is the probability that I will hit the target okay.

So, I am repeating again. I want a model that will tell me if I shoot the target from a particular distance maybe in feet and then that is the probably, that I will be able to hit the target. So, when I mean to answer this sort of question, first what will I have to do? First I will have to collect some observation from some data means, first say I will try to shoot the target from 1 feet, 2 feet, 3 feet, 4 feet, 5, 6, 7, 8, 9 or 4.5, 5.5 whatever it is, I will try to shoot the target from certain distance and then I will see what the results. Whether I could hit the target or I could not.

When I started when I start try it from 1 feet I could hit that, yes I could hit the target good. Then from 2 feet yes maybe again, from 3 it maybe not, from 4 feet yes. So, that way I have collected few data okay. Now I have this data. Now I want to know, given that I have started, I tried to hitting the target from particular feet, from particular distance, so, what is the probability of my hitting the target okay.

So, somehow I need a model, which will give me the answer to this. So, this this this is where the logistic regulation comes. Again though this is a simple application right, again let me take you give you an example in medical domain suppose in some medical research wants to find

out the probability, probably that a person will get cancer or whether it is highly probable or probable and not at all probable.

Means, highly some, what to say I should not say probable I will say quite a high chance of getting cancer, less chance and no chance at all maybe. These are the three categories maybe. Earlier example was these 2 categories and maybe here I am thinking 3 categories. So, suppose for that I will be taking different parameters like like maybe the age of the person, race of the person.

The genetic predisposition of the person, family history then what to say exercise history, his lifestyle history then maybe smoking habit, drinking habit many more suspects. Workplace environment at the workplace where does he stay everything put together. So, based on that whether the person is highly likely to have cancer or very less likely or not no chance at all of having a cancer, maybe 3 categories.

So, based on these different factors, so, how will I these are basically I will tell it is a predictor. These are based on these predictions, this predicted these variables how can I tell, how can I classify the person will fall in which category. So, this way again logistic regression comes. Again let us let me give you one example on say engineering applications. Engineering like better if I talk of from reliability perspective right. So, in reliability there is a term called FME a failure mode effect analysis. So, you might not be aware of this FME basically, okay you know no need to be aware of that.

What we do is that sometimes we try to find out what is the probability that the system will fail, a process will fail or a portion of the process will fail. For that what we try to do is that in what way the system can fail, we try to find out all the different ways in which the system can fail and there is a it is a very big technique okay. Let us not discuss that in that it will take 1 hour, if I start discussing FMA.

Now that there there is some some factors like, we have identified from the different ways a system or a process can fail. Now what is the severity of different we have identified different faults. Now what will what is the severity of the different faults, what is the occurrence of different faults and then what is the detection rates of different fall? All this we give some some value in a particular scale maybe from 1 to 10 scale.

This value we give it is usually subjective some experts give these values or we get this value just from the past data. So, this is where and based on this basically we try to calculate the RPN bricks probability risk probability number of the process excuse me. So, now in this case, so

where I am giving the number, the severity of the fault than the occurrence rate of default, when I am giving the number just by an expert judgment.

Maybe an expert is giving or just from the past data. Somehow some of the past data I am doing some calculation. I am giving the data and so that that if I could use logistic regression here then it will be more effective. So, these are the different, so logistic regression you can see it has its application in different domain. So, that is ahh the topic that we will be discussing today. So, for ahh first I will try to bring out a difference between the regressions, because we have learned only regression analysis.

Logistic regression is also something regression right. So, irrigation what is there? So, what is the difference between this regression analysis and logistic regressions? Then we will come to the concept of logistic regression, how to go ahead, how to do this logistic equation, how to model it and then I will also mention the type of logistic regression. Here I will be just mentioning the type of logistic execution. I will not be discussing those I will be discussing in my next lecture so okay. (refer time: 06:54)

Now so we have done regression analysis. Let us consider simple linear regression okay. We have discussed different regression. Let us just consider the simple linear regression. Simple linear regression what we have tried to do, means regression analysis means, we are trying to predict the response variable from a regression variable is not it? From an independent variable we are trying to predict a dependent value of a dependent variable.

So, now for that what we have tried to what we have done. So, for basically we tried to we found out the function. The function that maps the output with the input the function that maps the output with the input, how it maps. So, this function is only y = f x. Now what is this f x? F x is a function, which has certain parameter. Now how to find out the value of the parameter? we Once we found out the value of the parameter then only a function is complete and we can say okay.

This is the function which gives the relation between the independent variable and a dependent variable. Now to find out the value of the parameter, what we have used? We have we use least square estimation method. So, basically we try to find out the ahh error between the observed variable and the predicted variable. Predicted value, observed value and the predicted value we try to find out the error, we have squared this error and then we have to we try to find see that when this error is minimized.

So, at when this error is minimized that is the point where we can consider the coefficient is that

is the best value of the coefficient, because regression there may be different function. Now function depends on the coefficient. So, this is the best value of the coefficient, where this SSE is minimum. So, now we have minimized that function and from that we found out the value of the β whatever is the parameters β 1, β 2, α, β whatever it is we found out the parameter.

So, and one more thing is that, this regression analysis it is applicable when the dependent variable it follows a normal distribution. In case of logistic regression here the dependent variable does not follow normal, here and normal distribution here, the dependent variable is just a category value. If The first example what I have given trying to suit a target, what are the two categories?

Suiting the target, success success or failure. You could successfully suit the target you could not so the success or failure there are two categories. In case of the cancer example there were three categories right. So, similarly for there may be 2 categories, 3 characters, 4 categories are more than that. So, whatever it is. So, here the dependent variable it is a it is a basically categorical variable, which may be 2, 3 or few or more outcomes.

Now one more difference is there, actually both are doing prediction. Regression analysis also predicting predicting the independent variable, logistic regression is also predicting the independent. But regression analysis is predicting the independent variable here, the independent variable is in the same same scale, what we want to predict. But in case of logistic regression here we are predicting the categories with which categories belongs.

Probability that it belongs to which category, we have basically predicting the probability of belonging to different categories, if there are 2 categories, probability of belonging to the success category, probability belonging to the failure category. (refer time: 10:15) So, now here one more thing like in regression analysis we can do regression analysis if we take a small sample of data also, but in case of logistic regression we need more bigger sample.

If the number of categories are more than, even more bigger sample we need, so as to be statistically significant. So, And one more difference is that and what to say in the regression analysis see we have used least square methods. Least square method to estimate the value of the parameter, estimating the value of the parameter that is the coding right and that once we do that we are done.

So, we use here here we use list LSE method that is least square estimation matter to estimate the parameter. But whereas in for logistic regression what we use is that we use a concept called maximum likelihood estimation MLE matter. So, what is basically MLE? In very simple

language, if I tell what is MLE? Like given a set of observed values, like that for the example, that the shooting the target, I have taken and observed values like shooting from different place and whether it is success or failure.

So, given a set of observed values, set of values that we have observed the done we have to find out the model parameters okay. The Select the model parameter, the parameter should be such that, so that the likelihood of occurrence of that value. Whatever we have observed likelihood of occurrence of that value is maximized okay. So, the parameters should be as chosen in such a way that likelihood of the occurrence of that data set what we have observed, likelihood of the occurrence or the observed value is maximized okay. We should select the parameter in such a way.

So, that is the concept of MLE basically. So, we will see MLE in details here. See the principle of MLE is to estimate parameters but the principle of MLE is to estimate parameters by switching by choosing parameters values that give the largest possible likelihood. We have to choose the parameter in such a way that whatever the object values we get that that likelihood of getting that observed value is maximum okay.

We have to choose the parameter in such a way that technique is called MLE maximum likelihood estimation. So, now regression analyze predicts the value of a dependent variable. Logistic regression predict the probability of a given value of a dependent variable. Regression less is predictable value, logistic regression predict the probability of a given value okay. Probability of success, probability of failure whatever is the category.

If you have 4 categories probably a first category, probability of second category, probably a third category, probably a fourth category that is what logistic requisitions gives. Both estimates the respective model parameters, in both cases we have to estimate the respective model parameter, when we respect them ahh estimate the model parameter then only our model is complete essentially. (refer time: 13:15)

So, now see the model. In case of regression analysis if we consider a simple linear regression with just one independent variable, which was one regressor then we can consider its as a simple linear regression, maybe one regress or more whatever it is. We can consider it as a it will graph a straight line, if there are more number of variable then maybe it will be in a multi-dimensional plane. It will be a straight line okay, if it is a linear regression.

So, given a value of x it will give the value of y. This model will given a value of x, this model will give the value of y, if you know the parameter. Similarly for logistic regression model the model

is a sigmoid function okay. So, model is a sigmoid function, now what is a sigmoid function? It is nothing but it is it is a function, which takes an input as a ahh a real value and then outputs a probability.

Probability probably will definitely in the range of 0 and 1. It outputs a value in the range of 0 and 1 basically. That is we can tell this is a probability okay. It takes as input a real value and it outputs a value in the range of 0 and 1. So, this is the shape of a sigma sigmoid function, here you can see. It is taken as say this this it takes an as say sigmoid function okay.

Now what is the sigmoid function? If I can In a very simple way if I can take sigmoid function is nothing but f. f x = 1 by 1 + e to the power z. In a simple form, I can tell that what is the sigmoid function is 1 by 1 + e to the power z. Now this z is that z is again an expression it in terms of some parameters, in terms of some variables. So, here in the figure you see I have used, can you see the figure?

Of course it is very written in a very small font. I do not know whether you will be able to see that, you zoom it if you cannot see. So, see here, here I have used for z I have used - c 1 into x - c 2 okay. So, this c 1 and c 2 are what, c 1 and c 2 are the different parameters, c 1 and c 2 basically this will decide the shape of the curve. Shape of the IT is S shape Definitely it is S shaped, but this value of c 1 and c 2 it will decide the shape of the curve.

So, here you see if c 1 = 2 you see this blue colored line this is the blue colored line this is the blue color line this is the shape. If c 1 = 0.5 it is black solid line, c 1 = 1 oh sorry c 1 = 1 is the black solid line, c 1 = 0.5, this is this red dotted line. So, for different values of the parameter we get different shape and whatever is the shape, whatever it it outputs a value what is what will be the value of f x? Value of x will be in the range of 0 and 1.

So, we will be using the log is logistic regression model use this sigmoid function okay. Because why we use this σ because in logistic regressions we want a probability, probability will lie in the range of 0 and 1 right. So, that is why the sigmoid function comes into use in case of logistic regression model. Now we will see how we have used. (refer time: 16:41) Now for Let us see and first see an example here.

So, here is an example where the hours a student the number of hours a student is put on the studies that is given as well as whether he has passed the exam or he he could not pass the exam is given here right. So, say similar type of example we have discussed earlier where the it is in the data was number of hours that a student put on for studies and the other data is the independent variable is the stay score what you got based on the number.

We try to find out is there any relation between the number of hours studying and that the score of the student. So, we found here there is a relation, there is a linear relation and we could find a really linear relation. Now here it is the student the score is not given, but whether the student has passed or failed two categories, 0, 1, 0 is fail and 1 is pass okay. So, you see here is the data.

So, all this green color dot is the past data and category 1 that is pass and red color dot is the fail, for different values this is hours of study and this is the probability okay. This side will get probability and this is the hours of study. And here till now we did not get the probability. We Just from the data whatever this is our observed data, this data what is given this data what is given this is our observed data fine.

So, based on this observed data we will have to find out the given the given the student has put say ahh recep 4 is given which value is not given here. Let me take a value, which is not given okay 0.25 or whatever it is. Maybe 7 hours, some value, which is not given. Maybe say 1.8 hours okay or 1.9 hours. Maybe a student has put 1.9 hours what is the probability that he will pass the exam. We want to find that. So, first this is the observed value okay.

So, to predict that given the the student has put x hours whether he will what is the probability that we will pass to find that we will be using the sigmoid function. Now this is the what to say first we have plotted the data plotted. The data plotting will not will not be using the sigmoid curve there; σ will give us the output. So, data is this like for whatever hours of study just to what students who have failed we have given it with a red dot.

Where pass we have given you the green dot. Now to calculate the given x we need to calculate the probability whether the probability of passing or probably failing whatever it is. For that now we will be using the sigmoid function. Sigmoid function means f x = 1 by 1 - e to the power z. Now what is this z now? We will have to use the sigmoid function to give to get the result. Now what is this z for us? (refer time: 19:42)

So, basically in this ahh what to say, this concept was developed and popularized by primarily by Joseph Berkson, where he coined a new term logit okay. Now what is logic we will see? you This actually logistic regression has come from this logit only logit regression, logit regression or logistic regression. So, this Joseph Berkson on his coins this new term logit. But some this logit basically ahh logit will is somehow related with z we will see what it is okay.

So, logistic regression is a statistical method, it uses a logistic function to model a binary dependent variable. Now okay, it may be binary or it may be more than one. So, if suppose now we are looking for binary in this example because, it is student pass or fail, pass the exam or fail could not pass the exam. So, it is a binary, so it uses a logistic function to model a binary dependent variable or although many more complex extensions exist.

Complex extension may be more than two categories okay. So, our job here is to find logistic regression estimates the parameter of the logistic model. We found that it will be using the sigmoid graph. Now we will have to estimate, what will be the parameter means z is something, that there are some parameters to it. Now we will have to estimate the parameter. First what is the sigmoid function for us and then what are the parameters okay. (refer time: 21:05)

So, see a binary logistic model has 2 dependent variables. If we are considering binary it has this 2 dependent possible variable, with 2 possible values that is pass or fail, happy or sad whatever it is success, failure okay. So, it is represented by an indicator variable, where the two values are labeled 1 or 0. If it is binary and repeating it again and again. (refer time: 21:29) So, now the logistic function, takes the following form.

So, logistic function, so what is the law this logistic function means we want to find out probability that the student will pass the exam? So, that let me write it this way, probability that the student will pass the exam given certain values of x, x may be the hours. So, that I am writing it this way or let me call it as p x this is let me call it as p (x). Given x what is the probability that he will pass. So, this I am writing is this way given x what is the probability he will pass. So, to find this, I will be using the sigmoid function. So, what is my sigmoid function? (refer time: 22:06)

Sigmoid function is So, what is my sigmoid function, so P x = 1 by 1 + e to the power z. So, this if I multiply both sides by e to the power z but will I get e to the power z by 1 + e to the power z is not it? If I multiply both sides by both numerator and denominator by e to the power z this remains sense I will just get e to the power z by 1 + e to the power z.

Why I bought this in this form. It becomes easier for calculation. We will see that. So, here, so this is in this form. So, my P x given x my probability that the student will pass that is I can tell it y. Y = P x is equals to nothing but e to the power x by 1 + e to the power x okay. Now let us not write x it is let us write as z okay. Now what is this z? It is nothing but a if if there are just one variable.

In in our example, what is given? It is just one variable. Hours put on study, there may be more

than one variable okay. So, for each variable there will be a parameter is not it? So, z is nothing but a linear combination of all these parameters along with an intercept, it is nothing but a linear combination of all those parameters along with an intercept. This is the same thing what we got to see in the linear regression okay.

So, here z is nothing but this, suppose there are m parameters x 1 to x m. So, it will be my z will be β 0 + β 1 x 1 + β 2 x 2 + β 3 x 3 up to + β m x m, if there are more than one parameters. Now in this case there is just one parameter the example hours to study and whether he has pass and fail. So it will be β 0 + β1 x e to the power β 0 + β 1 x by 1 + e to the power β 0 + β 1 x okay. So, this is the probability that the student will pass.

Now we will introduce a new term here that is called odds. It is it is not new, many of you may be ever of this term is what is odd? Odd is the ratio probability of success and failure. Odd is basically a ratio of something happening to the something happening probability of something happening by something not happening. Usually when you try to find out probability, what is this? Probability is this ratio of something happening by means the sorry not ratio when we try to find out probably we how we write?

As the number of times something has happened divided by the total number of times that anything that it can happen. The happening plus not having put together the whole sample space that is how we find out the probability, but in case of odds that is the difference is that there in the denominator it is not the total sample space, the denominator is just the number of ways it can fail, number of ways it can pass by number of ways it can fail okay.

So, odd is the ratio of the probability of success and failure. So, that is it. So, now what is odd? Odd is in case in this example what is this? Probably that the student passed by probably that the students fail P x by okay P x by 1 - P x. This is the probability that the students fail. So, if I write P x by 1 - this is P x, P x by 1 - P x if I use this is P x and if you do simplification you will get odd. this is the value of odd.

You can do it yourself. Say very simple calculation okay. So, this is the value of odd. Now what will do is that this logit logit is basically the log of this odd. This new term is Joseph this particular person Joseph Berkson, who has coined this term logit. Logit is nothing but the log of this odds okay. So, log of odds. If we try to find out the log of odds what we will get? Log of odd is nothing but this term.

This log of odd is nothing but this is called a logit and this is the z for us okay. This is nothing but z here what we have used z, here I am using the term t z or t anything instead of z if I write here

t, it is a variable whatever name you give okay. So, this is the logit. That is t is the logit, so value of t is this okay. So, now we know P x. What is P x? P x is this. And what is this? This is nothing but ahh sorry this b 0 plus this portion not e this portion.

What is this portion? This portion is nothing but t then P x I can write e to t e to the power t by 1 + e to the power t fine, that is my P x probability that the student will pass okay. So, now this e this t has β 0, β 1 ahh terms are there. Now so if I can find out this value of t, how can how I can find out the value of t. If I know the value of all the parameters x value I will get from the data, x is the independent variable, which I want to know.

If I put this many hours of study will a student pass or will I pass? So, x value I know. The only things I do not know is the β the parameters. So, if I can find out this value of the parameters, then I will be able to find out the t. If I can find out the value of t, then I will be fine able to find out the probability is not it? This is nothing but the sigmoid function is not it. This is nothing, but the sigmoid function. Sigmoid function was Our sigmoid function how we have written 1 by 1 - e to the power of sorry 1 + e to the power 3.

This is just in writing in a different format. I just got this e to the power t by 1 + e to the power t. This is nothing but a sigmoid function. So, we use the sigmoid function to find out the probability that the student will pass given that he has put x hours in study that is e to the power t by 1 + e to the power t. Now if I am if I can calculate t, for calculating t I need to know my the value of the parameters right. (refer time: 28:29)

So, a small example is here. A sample is collected to examine the effect of toxic substance on tumor okay. A subject is examined for the toxic content in the body under the presence or absence of tumors. We have some data, first to do anything in any any sort of statistical studies we need some data first right. So, what is the some subjects if we call this some people are examine and they try to find out what is the concentration of particular toxic substance and the person has tumor or not.

So, these are the datas, which you got. So, concern for around see here. The number of subject at each concentration N and the number of number having tumors is shown in the table. So, for concentration of 0 we found total 50 subjects, total 50 subjects as concentration 0 out of that two had tumor. Similarly for concern 54 subject had concentration 2.1 and 5 had tumor.

Similarly 46 persons had concentration 5.4 out of them 5 at tumor similarly okay likewise for the outer values. Now the first thing is we will have to find out the logic that is the value of t. If we can find out the value of t then we will be able to solve the find out the probability right. So, to

find out the probability what we need? We need to find out the odds. How What is what is odds? Odds will be probability of having tumor by not having tumor okay.

So, what is this? How many person is having tumor? Two. So, it will be odd will be 2 by 48. So, here odd will be 2 by 48 for the first term this is this. Second is the odd will be 5 by 49, likewise we found the odds. Then logit is what? Log of odds and log of odds is that this value okay. So, here we got everything right. We found a logit we found the logic value. This is the logit value that is the value of t okay. (refer time: 30:40)

So, now we can find out the value of the parameters. So, what is t? T is nothing but is there is only one independent variable. So, there will be two parameters β 0 and β 1. So, t = β 0 + β 1 x. So, this logit is a linear function and with this data we will be able to find out the β 0 and β 1 value. Here I will have to explain something first listen to this. So, with this with this data, because just see we know t, we know x just we do not know β 0 and β 1.

Let us take two of this and we will be able to find out β 0 and β 1 okay or we may plot the data what is β 0? B 0 is the intercept, β 1 is what? B1 is the slope. We can plot the data, from there also we can find out the β 0 and β 1. Once we know the β 1, β 0 and β 1 then we know the t value. Then given any x will be able to tell whether the person has tumor or not okay.

So, now in this question see, we could easily found β 0. First of all if the points fall in a straight line just by finding out the intercept under slope slope we could find out the β 0 and β 1 or by since there is two unknown parameters, just take two values and will be able to find out the β 0 and β 1. But the question is if I take these two values I will get one set of β 0 and β 1.

If I take these two values maybe I will get some other set of β 0 and β 1 maybe if I take this value and this value maybe I get another set of β 0 and β 1. So, there may be different values of β 0 and β 1. Now, which is the best value? okay, To find out which is the best value of β 0 and the best value of β 0 and β 1 is that which will make the likelihood of getting the value what we have observed is maximum.

It will maximize the likelihood of occurrence of the observed value understood. So, for that we will have to use the MLE method. Now let us not use that. Now let us just stick to this we got the β 0, β 1 value. (refer time: 32:29) and then given that for example a subject is exposed to a concentration of 10 has an estimated probability of tumor. So, we could found its estimated probability of tumor is 0.36 okay. (refer time: 32:41)

So, now the log the log-odds that is the logarithm of odd is a linear combination of one or more independent variable that is that we call it predictors okay. There in the earlier example the concentration is just only one independent variable. Concentration is the predictor. The independent variable can each be a continuous variable it can be any real value okay. The probability of the value can vary between 0 such as certainly false and once this is certainly true, if it is just two categories. (refer time: 33:19)

Increasing one of the independent variable, if in the example what we have seen there was just one independent variable. If we increase one of the independent if there is more independent variable, multiplicatively scales the odds of the given outcome at a constant rate. If we increase one more independent variable if there is one more independent variable, which also decides the dependent variable then what happens.

We will be using one more parameter and what is its multiplicity multiplicatively scales the odds of the given outcome at a constant rate, plus β 2, x 2. Earlier it was β 0 + β 1 x. Now it is β 2, x 2. Another one variable x 2 and as a constant rate that is β 2, with each independent variable with each independent variable having its own parameter fine. For a binary dependent these generalize the odds ratio.

For binary we have seen just it is just two, binary logistic the dependent variable has two level, multinomial, outputs with more than two nomial ah. more than two levels and see if it is two then we call it binary logistic regression. If it is more than two we call it multinomial logistic regression. Then we have something called ordinal logistic regression. If the multiple categories are ordered, ordinal its ranking you knows?

So, if we can order the different categories then it is called ordinal logistic regression. Here also we will be using the sigmoid function only, which is just give the probability maybe we will be using some threshold value. In case of binary maybe the threshold value is 0.5. If the probability of occurrence is ahh greater than or equal to 0.5 then maybe it is success. If the probability of occurrence is less than 0.5 then it is failure. We can consider it that way.

Similarly if the 3 levels, accordingly will set the threshold for each level okay. So, similarly we can have an ordinal logistic regression where there can be ranking among the the categories. (refer time: 35:06) Logistic regression models the probabilities for classif ah is Logistic regression a classifier now that is the question okay. So, logistic regression it models the probability for classified classification problem with possible outcomes.

What does logistic equation gives? It gives the probability of the possible outcomes. What is the

probability of being in this class, what is the probability of being in this class being in this class, this class, this class. What is the probability of being in different classes? Logistic regression gives that it is basically an extension of linear regression model for classification problem.

But it is not a classifier. Classifier, classifies into different classes, it is this gives the probability of being in a different classes okay. The logistic regression simply models probably of output in terms of input and does not perform statistical classification. It is not a classifier. It is strictly not a classifier. It just gives the probability of being in different classes. In that way we can say it is a classifier actually.

But it is actually not a classifier. It gives the probability of being in different classes. Although it can be used to make a classifier, how we can use to make a classifier. That is what that is what just now what I told by choosing a cut-off value and classifying inputs with probably greater than the cut-off as one class, below the cut-off as the other. Logistic regression, give us the probability of occurrence of a particular value, probability of passing the exam, probability of failing the exam.

Now if we choose a particular result say suppose 0.5 we have choose that result. Now if we get a probability more than 5 or more 0.5 or more than that then we can this it belongs to success class. If it is less than 0.5 it belongs to the failure class. That way we have you so means we have used this to make to act as a classifier. But actually what the logistic regression gives, it gives us the probability of belonging to different classes. It is strictly not a classifier, but we can use it to make a classifier. (refer time: 37:04)

So, there are different logistic regression techniques. So, one is binary logistic regression. Under binary logistic regression, there are definitely two categories of outputs. And there may be one explanatory variable there may be many explanatory variable. For multinomial there are many explanatory variables and many categories or maybe one explanatory variable, then many categories that is not possible right okay. So, these are the different types of logistic regression, so, which we will be discussing in the next coming few lectures. (refer time: 37:25)

So, now coming to the conclusion, so what this lecture provide an introduction to logistic regression, that comprises a background of the development of logistic regressions, compression between linear and logistic regression, how logistic regression is used as a classifier, how we can use logistic regression as a classifier by making the ah by having a threshold value right. so and different types of logistic regression that also we have just mentioned here okay. We did not discuss. In the next lecture we will learn more about logistic regression. (refer time: 37:55) With that this is the reference. And thank you guys.

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

Lecture - 46
Logistic Regression (Part - Ⅱ)

Hi guys, so in continuation of our discussion on logistic regression last lecture we have discussed logistic regression. So, there we have mentioned different type of logistic regression but we did not discuss in detail so today we will be discussing that.
(refer time: 00:37)

So, one type of logistic regression that we have seen is binary logistic regression. That is binary logistic regression again there were two different types that is one is one explanatory variable two categories or more than one explanative variables and still two categories binary means it will have to get two category pass fail success failure happy sad whatever it is two categories. But explanatory variable may be one or explain a variable may be more. So, today we will be discussing binary logistic regression with one explanatory variable.
(refer time: 01:09)

So, now say this same example I bought here again so a group of 20 students spends between 0 and 6 hours 0 to 6 hours studying for an exam. The table shows the number of hours each student spends studying and whether they passed or fail. So, if it would have been a linear regression instead of pass or fail if a score is given the amount of hours, it is the student has put what is the score.

So, if there is a relationship between the hours and dress code and we would have done gone for a linear regression analysis and simple regression analysis. And if the what is the relationship is linear then it is a simple linear regression analysis this is just one variable but unfortunately no we are not given the test score what is given is the student has passed or failure two categories pass category fail category.

So, now here meaning we want to find a model given the student has put this many hours of study say some x what is the probability he will be passed whether we will succeed. So, we have to find it out. Here also we are predicting there also we are predicting but here prediction is different their prediction was different, their means I mean regression, simple regression.

(refer time: 02:25)

So, how does the number of hours spent studying affect the probability of the student come to the fag end of this course still I am not comfortable using this pin actually. So, how does the number of hours see again I think it is done, now how does the number of hours studying affect the probability of the student passing the exam. That is what we find out we need a module for that.

The reason for using logistic revision for this problem is that values of the dependent variable pass and fail while represented by 1 and 0 and are not cardinal numbers. The values are not cardinal numbers the values are not some scores 120 140 whatever it is. If the problem was changed so that pass field was replaced with grade from 0 to 100 certain in some score that is cardinal numbers then simple regression analysis could be used but unfortunately here we cannot use simple regression analysis.
(refer time: 03:33)

So, we used to fit a logistic function, logistic function means a logistic function means a sigmoid function to the data consisting of the hours studied and outcome of the test y i = 1 for pass and 0 for fail. We need basically probability of that probability that the student will pass that is y i = 1 and probably that the student will fail that is y i = 0. We need to find such a sigmoid function the data points indexed by the subscript i which runs from i 0 1 to 20 because there are 20 students.

The x variable is called the explanatory variable and the y is called the categorical variable. The x variable is called the explanatory variable it is also called a predictor because this variable predicts the outcome. So, x variable is also called the predictor or explanatory variable and y variable is called the categorical variable because we have different categories output is in different categories. Consisting of two categories pass or failed corresponding to category values 1 and 0 respectively.
(refer time: 04:39)

So, here see I have if you find it difficult you may not use this also, I have used a different sigmoid function. It is not a different σ function but just a different form we are used to having this sigmoid function f of x = 1 / 1 - e^{-z}, we are used to this sigmoid function. So, if you are comfortable with that go with that but just for binary logistic regression, we can use this form as well.

It is just the same thing just since it is a binary, we have bring out the concept of μ I will just tell you. So, now if this is the thing then what is t that is the logistic is log of odds that is the odd or

this p / 1 – p. We have seen it in the last lecture is it what is odd, odd is p / 1 - p so logic is log of p / 1 - p. So, here for generalized sake I have written b base is b usually base I you can use base e you can base use this 2, 10 it totally depends on the range of the data.

If you have very high at a high range of data then definitely will go for log 10. So, it depending on the range of the data you can use whatever base it is. So, it is not necessarily you will use base e only. Usually when we come for more explanatory variable more than one so it may be that log base e may not suffice then maybe we have to use a higher base maybe 2 or maybe 10. So, that is why just for generalizes I have kept here base b this is obvious like all of you know that.

So, the graph of the logistic regression curve fitted to the line does this is the curve the curve shows the probability of passing an exam versus hours of studying. So, versus hours of studying business hours of studying and decide it is the x axis is the hours of studying y axis is the probability of passing the exam. So, this is the sigmoid function this graph gives us that for how many hours of studying what is the probability of passing.

So, in this usually $1 - 3^{-z}$ when z goes to ∞ our probability value goes to 1 when z goes to minus ∞ our probability value goes to 0.
(refer time: 07:00)

So, this is the logistic function which we have just now considered. So, this basically this expression since if you are not comfortable you can just convert it into this form. It is the same thing because it is just one parameter so it will be $\beta 0 + \beta 1 x$ so $\beta 0 + \beta 1 x$. If we; just substitute this then what will be my $\beta 0$, $\beta 0$ is nothing but μ / s and intercept is nothing but $1 / s$.

So, what is μ actually why I have used a μ, μ is a location parameter the midpoint at a curve where p of μ is equals to half. In binary there will be just two categories which one to use as a classifier so maybe we can use this (07:43) no issue if you are finding a trouble you can just use that only no issue at all. So, μ is nothing but it is just a location parameter that is the midpoint of the curve, μ is the middle of the median of the curve.

So, p μ is half probability of μ is will be always half so and s is the scale parameter. So, what is $\beta 0$ here? If I substitute it in this the standard format this is the standard format so $\beta 0$ is $- \mu / x \beta 1 = 1 / s$. So, now we have already seen what is p of x i what is the probability given that x i values, what is the probability that it will this student will pass or the it will hit the target.

Basically, we are trying to find out the success since in categories it is more than two categories then we will take it differently. So, we may define fit to y i, fit to y i maybe means that a success pass that is one at a given axis p i we can define is this way just for simple calculation instead of carrying out p of x i all the way just simply I can just write p i, p i means p of x i. What is the probability given x i what is the probability that it will be success do not get confused with that.
(refer time: 08:57)

So, now p i had a problem is that the corresponding y i will be unity and 1 - p i are the probabilities that they will be zero, nothing no doubt here, I hope. So, we used to find the values of β 0 and β 1 which gives the best fit to the data. We have seen one example in my last lecture we had a concentration under the odds and the if a person has certain concentration of some particular whatever it is and then what is the thing that he has (09:27).

So, there from the data given I have already shown you that we can find out the different values of β 0 and β 1 but we want to find out that the best values of β 0 and β 1 which will maximize the likelihood of the occurrence of the observed value because the values that we have observed as a real value that we have observed that is. So, we want such parameters which will maximize the likelihood of occurrence of this observed value.

So, there will be different β 1 so now we need the best β 1 β 2. Best β 1 β 2 means we will have to use MLE estimation method the MLE is nothing just that it finds out that parameter value it finds out the parameter value in such a way that is maximize the occurrence of the observation what we have seen. It maximizes the likelihood of the observation that we have observed likelihood of the occurrence that we have observed.

So, in case of linear regression whatever we have seen that is written here we know how we try to find out the minimize the error and then we try to find out the parameter. Similarly, here we will try to find out the, we will try to maximize the likelihood of the observation.
(refer time: 10:43)

So, this is the likelihood function, so this likelihood function what is that is the probability that the given certain test is produced by the logistic function, so this is the likelihood function. So, we want to maximize the likelihood of occurrence of the observed value. So, what is that, so basically it is a joint probability distribution of all the values what we have seen join probability distribution of all the values what we have seen.
So, this is product of all the p i, p i where y i = 1 that means where we got the success probability of all the products all the product of all the probabilities where we got the thing there is a success and 1 - p i where y i = 0 this is nothing but the likelihood function. So, when we maximize this likelihood function that is after maximizing this likelihood function whatever value

we get for the parameters that is the best parameter for β 0 and β 1.

So, understood this is the likelihood function nothing it is just the joint probability distribution of all the probabilities, basically we are trying to multiply all the probabilities. Now here when you try to multiply all the probabilities it might the equation might become unstable because when some probabilities if you try to multiply some very small probabilities maybe the value becomes maybe and may become zero also.

So, when we try to multiply too many probabilities it senses that the equation become unstable. So, in such case what we do is that so this is my likelihood function and I need to I want to maximize this likelihood function. So, because there it involves a lot of multiplication of probability, joint probability distribution is nothing but a multiplication of the conditional different conditional probabilities.

So, what I do is that I take the log of this that is called the log likelihood. So, if I take the log of this if I take the log both side then this product becomes sum is not a log when the product if we did not take the log this side now this product will become sum that is all.
(refer time: 12:50)

So, this is the thing log likelihood. So, this also I can write it in this form. So, this is only instead remember what a p i was I am using p of x i I am writing this p i. Now just for your understanding I have kept this step you see what does this step means this step means just this and so up here I bought again p x i what was because if you get confused what is p i, p i is basically probability of success given x i probability of success means y i = 1 given x i probability of y i = 1 given x i this.

So, if we simply and whatever p of x i p of x i has this b 0 and b 1 term what is p of x i p of x i is $e^{t/1+e^t}$ and what is t, t = β 0 + β 1 x. So, simple simplification here nothing I will not explain these simple simplifications but actually you do not have to do it manually all these ready-made programs are there, in the net nowadays all these ready-made programs are there you will have to you do not have to do this many MLE you do not do it manually just feed the data and you get the results.

But even if you should know also how to do just, I have simplified it, now when I am talking of maximize this likelihood, I made log likelihood so that my equation remains stable since I am multiplying the probability is so I have made it log likelihood now to maximize this. So, basically to get the optimum value. So, what I will have to do? I will have to differentiate these equations partial differentiation with respect to each and every parameter here there are just two

parameter I will do the partial derivation with respect to β 0.
I will do the partial derivation with respect to β 1 and then I will find out the value of β 0 and β 1 by equalizing to 0.
(refer time: 14:54)

So, that is what, I am just differentiating it with respect to the different parameters and once I have what to say differentiated it with respect to the different parameters then I will equalize it to 0 to get the values of the β 0 and β 1 that is the technique standard technique which you know but here there is a problem. What problem is there? You see the expression here this p i when I talk of p i there is a term of exponential terms.

So, this is this whole this expression it is not an algebraic expression. So, when this is not an algebraic expression so when we differentiate it, we cannot because since it is not an algebraic expression it is not it will not give us a finite sequence of some algebraic operation is not it exponential terms are there and the exponential terms there then sin term cos and when these terms are there.

When it is we will not get a finite series of algebraic operations it will give an infinite series. So, when this is an infinite series definitely though I have shown it here because this is the standard way of equalizing is to zero and then finding out the value but actually not actually you cannot equalize it to zero because this is not a finite series. It is not an algebraic expression that is why this is not an finite series.

So, what you will have to do? You will have to after differentiation you will have to use one of the numerical techniques methods the best is Newton Raphson method you will iteratively using some approximations iteratively you have to solve this expression and find out the values of β 0 and β 1 by using Newton Raphson method. Newton Raphson method will best suit here. So, now please do not expect me to explain the Newton Raphson method.

That is a totally a concept of numerical techniques numerical methods that all of you should know as engineering student all of you should know that, not only in this subject in many subjects you will be needing this numerical technique. So, please go to it if you have not gone through it please go through it.
(refer time: 17:03)

So, now so by that whatever Newton Raphson method we have used and we found out the value of β 0 and β 1, we found out the value of β 0 and β 1. So, now what was for this since it was binary, we have used the concept of μ and s and all those stuff (17:28) so I can find out

also what was μ you can if you can remember here see here β 0 is - μ / s β 1 = 1 / s from there if I find out μ what will be μ.

So, this is my μ value I have the μ value I can find out since I know β 0 and β 1, I can find the as value similarly I can find out the logic expression that is the t. So, now the next question you have to find out I found out the coefficient. Now again I need to do here also I need to do the significance test; significance test means here and here what is my null hypothesis. My null hypothesis is that there is no correlation there is no the what to say the predictor does not the coefficient of the predictor to give the dependent variable the coefficient is almost equals to zero.

That is there is no dependency between this predictor and the dependent variable and the response variable. So, that is my null hypothesis. Alternate hypothesis the predictor contributes for the dependent variable. So, now how do I find a better at this whether there is a dependency between the predictor and a dependent variable significant dependency or it is not. So, for that I need to find out the p value.

If my p value is very small that means if it falls in a rejection region given a significance level usually significance level is 0.05 is considered so even if my significance p value is much less than this it is less than the 0.05 that means if it falls in a critical region then I reject the null hypothesis. Null hypothesis means that there is no dependency between the predictor variable and the dependent variable.

So, now how do I find a p value here? So, I found out the coefficient so after finding out the coefficient I will have to find out the standard error. What is the standard error? Standard error means the how the coefficient the variance of the coefficients. So, remember the first example in my first class on logistic regression when I have given so concentration cancer concentration and having tumour.

So, as I told you we will be able to find out different β 0 β 1 values is not it. So, now let us take just one parameter suppose the β 0 or let us not consider the intercept because what is important is the slopes the different slopes that we have. The parameters that are that basically that is involved with the independent variable that is the predictors. So now this β 1, so now we will have different values of β 1 is not it so there will be variance among these values.

So, this variance among these values is the standard error. So, now how do we find out the standard error? So, it is really not possible to find out the standard error manually there are soft wares for that, so you know you should know the technique how we find out is that supposed to

find out the coefficient for we have used maximum likelihood estimation and whatever it is and we found a cost coefficient for that suppose we took a sample a bigger size sample.

Suppose we took a sample of size 100 for example of size 100 and we have found out the coefficient β 0 and β 1. So, now what from this sample we will take a sub sample, say maybe of size say we have taken a sub size say 60 or size 50 whatever it is. We will take another from this sample we will take is another sample that is lesser than our original sample. We will by maybe will take a sample in such a way so that by replacing we will take one again will replace that value again will take will replace that value.

So, we will take a sample and then from that sample we will calculate another β 0 and β 1. Similarly, again we will resample again we will take another sample from that so that way we have taken many samples. For each sample we have used this MLE estimates and we got the β 0 a β 1 value. One largest on the largest sample we have calculated β 0 β 1 values, now from this smaller samples from the same sample but we have what to say done sampling of this sample.

We have done the sampling of the sample means we have taken a sub sample of the sample this sub sample we have taken maybe say size 50 so we will take differences sample of size 50. We will take sample it different times (21:43) so we will take a different sample of size 50 also be size equal so then we will find out different values of β 0 and β 1. Now this you know different values let us consider β 1.

So, we found out different values of β so we will find out the variance among this β 1 there will be this β 1 the values are different this difference what is this difference this variance is nothing but this is the standard error that is the standard error. So, here we have this is the coefficient value what we got it by (22:13) then we calculate the standard error. Once we calculate the standard error then we will be able to find out the z value.

What is the z value? There is one technique by proposed by well this is called well technique to find out the z value is nothing but the coefficient divided by standard error gives us the z value. So, once we get the z value definitely, we will be able to find out the P value from the z value given a z value we will be able to find out the p value that is no issue at all. So, now here from the z value itself we can come to the conclusion.

If the value of z is further away from zero that is the value of z is away from zero then only, we can say this it is large enough to be significant that means there is a relation between the predictor variable and the dependent variable. So, this z value is quite large this z value is quite

large so this itself it shows that that null hypothesis is rejected. So, when we got the p value also which is very less that is how we do the significance test if we have to do the significant test.

And once we find a z value now β 0, we got this value β 1 we got this value these are some high level topics so I am just mentioning it I am not going to explain in details. So, we found a β 0 value we found a β 1 value we found z value. Now remember confidence interval while teaching hypothesis testing while discussing hypothesis then I have discussed confidence interval confidence interval means what value b 0 will take within in the we give the interval of β 0 value.

We cannot because this is we are just giving one estimate point estimate that is not good probability of a point estimate is zero. So, it is better if it is always better if you can give an interval. So, once we have the z value once you have the coefficient value, we can calculate the interval also. Same technique what we; have used to calculate confidence interval while discussing hypothesis testing.
(refer time: 24:20)

So, that is what here this value is 0.0167 it is very less value by Wald test it is the output indicates that hour studying is significantly associated with a probability of passing the exam. Since the p value is very less that means it is significantly hours and passing the exam this association is significant association it is not an insignificant association. If the p value is greater than 0.05 then null hypothesis is satisfied that means hour studying it does not it is not a predictor for the passing the exam.

So, β 0 β 1 coefficient may be entered into the logistic regression equation to estimate the probability of passing the exam. So, for example for a student who studied 2 hours entering the value at x = 2 into the equation gives us the estimate probability for passing the exam. So, here this is t this is my logic equation so I found out I have found the b I have I know the β 0 value β 1 value I know what is x axis 2 so I found out what is t then how what is how do I find out probability of passing given 2 hours.

This is the formula is not it the sigmoid function I know the value of t putting the value of t I got the probability of passing the exam is not it. It is not difficult once you find out the β 0 and β 1 finding out the probability is not difficult at all.
(refer time: 25:43)

So, similarly suppose a student studies 4 hours so what is this estimated probability of passing the exam. If you studies 4 hours that means x = 4 I have β 0 and β 1 value so I found out the t value this is my t value. So, probability of passing the exam $p = 1 + e^{-t}$ it is the same thing as e^{t}

$/ 1 + e^{t}$ same thing this form or this form whichever form you use.

So, 0.87 and probability of this is the 0.87 probability of passing the exam. The table shows the problem you are passing the exam for several hours values of our study. You see here for different hours here I have specifically kept the value 2.71. Why for binary we use this depth form which I have given the sigmoid representation what we have seen no need to use that do not get confused.

But if you use this for your understanding see for μ is equals to 2.71 what I got probability I got 0.05 it is the mean of the curve basically 2.01 I found t value is 0 then log of e 3 is 1 probability is 0.5. So, for different values hours of study I could find out the probability of occurrence and probably have passed in the exam.
(refer time: 26:54)

So, output from the logistic analysis gives a p value which is based on the Wald z score, Wald z score is nothing but the coefficient by standard error. Coefficient for each we find out for each individual coefficient standard error we find out for individual coefficient we find out the individual coefficient for β 0 β 1 β 2 how many other coefficients are there. This simple model is known as binary logistic regressions with one variable and two categories and has one explanatory variable.

The above example of binary logistic regression on one explanative variable can be generalized to binary logistic regression on any number of explanatory variables. Here we had just one explanatory variable the example we have seen is just the hours of study that is one explanatory variable. Now we can generalize it to any number of relative variables maybe hours of study and maybe the age maybe both these ages also may be the predictor of passing the exam.

So, that way many number of expanded variables and any number of categorical values category values may not be only just zero and one here we have seen just 0 and 1 fail or pass there may be more than two categories as well.
(refer time: 28:02)

So, in this lecture we discussed binary logistic regressions, we learned the concept of logic very important what is logic. Now you understood from logit we got the logistic regression logit is nothing but the log of odds, what is odd probability of success by probability of failure ratio of success by failure basically. So, that is I should not use the word probability odd is the ratio of success by failure that is on.

Then we understood a concept of logit we understood a concept of odd then what is and maximum likelihood estimation. And also, the use of sigmoid function in logistic regression, these are very important things which you have learned in this lecture. In the next lecture we learned binary logistic regression with more than one explanatory variable and also, we will learn multinomial logistic regression.

(refer time: 28:50)

And this is the reference and thank you guys.

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

Lecture - 47
Logistic Regression (Part Ⅲ)

Hello guys, so in continuation of our lecture on logistic regression on the module relation analysis.
(Refer Slide Time: 00:37)

We have already seen the binary logistic regressions where we have considered one explanatory variable and two categories. So, today we will be seeing the second on the case that is many explanatory variable two categories. Many explanative variables meaning like the predictors in the earlier case which you have seen for categories are two, successfully a pass fail or whatever it is presence absence whatever you can say which we designated as 1 and 0.

The categories are two but then explanative variable we have just one like though it is not a very good example, the example what I have sited is like the number of hours you have put to study and with the you the student has passed or failed. So, it is like the two categories pass fail and that is the response variable and what is the predictor variable. Predictor variable is the number of hours that a student has put on the studies.

So, that was one explanatory variable only what was the predictor, that is the number of hours of study. So, now we will be seeing the second case where the explanative variables are more than one and categories are two only. So, we call it binary basically binary means it has to be two, so many explanatory variables meaning like for the same example maybe if age is also one of the predictor.

So, number of hours and an age of the student if we have two explanatory variable two predictor which will affect the response variable then we will call it a when it is more than one then we will call it a many explanatory variable two categories but that also falls under binary logistic regression. So, we will be discussing that and then finally we will discuss multinomial logistic regression.

Multinomial logistic regression means where the number of explanatory variable is definitely

more than one and as well as number of categories also more than two.
(Refer Slide Time: 02:30)

So, now first coming to this the second case there is many explanatory variable and two categories.
(Refer Slide Time: 02:34)

So, to begin with we may consider a logistic model with M explanatory variables. So, M explanatory variables means x 1 x 2 up to say x M and two categorical variables what are the categories variable y, y = 0 and y = 1. For simple binary logistic regression model, we assumed a linear relationship is not it for the when we have considered a simple binary logistic regression model that is one explanatory variable and two response variable.

We assumed a linear relationship between the predictor variable and the log odds. Remember we wanted how we find found out the log odds. So, first we found out the log odds, what is odds success number of success by number of failures. So, that is odd and then we have done the log or we have taken the log of odds log of orders, what is that is nothing but that is the logic from logic only.

We got the logistic regression name I think you can remember that, if you do not remember kindly go through my lecture number one and two for this logistic regression, this is the lecture three logistic relation. So, and this linear relationship, now whatever linear relationship we have seen for one predictor variable for the predictor variable and the log odds the same linear relationship we can extend to the case of M explanatory variable.

If you can remember, let me write it just here remember we have used the sigmoid function. First of all, why we have used the sigmoid function, because sigmoid function what is the sigmoid function our sigmoid function is something of this sort 1 / 1 + e where this is our logic function, we have already seen in the last class. So, now what we got logic that is t we call it t, t we got it something this form β 0 + β 1 x.

This is what we call as we got the logit function for the simple binary logistic regression model, this was our logit function remember. So, now similarly this we can extend it for M explanative variable as well, so here our logit that is t, t is nothing but log of p / 1 – p, p is 1 p is the; what is this p / 1 - p it is the odd of getting 1. What is the odd of getting 1? Odd of getting 1 in so odd of getting 1 / odd not getting 1.

There is a difference between odds and the probability, I have already mentioned. So, this is when we take log of this then we got is whatever we call it we call it a logit that is t. So, this is nothing but this if we simply extend this to the multi variable we call it we get is in this form. So, here you see the parameter β 0 to β M if there are M parameters M explanative variable x 1 to x M, M explanatory variable then we have parameters β 0 to β M that is we have parameters as well so, M + 1.

Similar here we have seen we have only one explanatory variable the number of parameters is 2 β 0 + β 1 where t is the log odds and β are parameters of the model.
(Refer Slide Time: 05:53)

So, now and here you see I have used the log base b, b means I have just want to generalize it while considering the x many explanatory variable why I had just want to generalize it in say earlier cases what we have used. We have just used log e base e is not it. Here, basically it is that if you can use either e or 2 or 10 it totally depends on the range of your data, what range your data falls into.

So, expecting that for many explanatory variable it may have a very higher range if it is a very high range then, we will go for base 2 accordingly what if it is in the range of 2 to power 10 then we will go to the for the base of 2. So, likewise I have kept a base as b so as to reflect that it can use any base 2 e or 10 depending totally depending on the range of the data values.
(Refer Slide Time: 06:50)

So, now what as I told you our explanative variables are total M explanatory variable x 1 + x 1 x 2 x 2 x 3 up to x M total M explanative variable and for M explanatory variable how many parameters we need we need M + 1 parameters that is β 0 to β M. So, just to generalize it we to make it more compact notation here I have introduced one more predictor. We call it a predictor or explanatory variable whatever it is and the output we call it a response variable.

So, we have introduced one more that is x 0 and x 0 is nothing but x 0 = 1. So, it will not make any change because you see simple equation when we have t 0 = β 0 x 0 + β 1 x 1 where x 0 = 1 it is just equal to β 0 + β 1 x 1. So, it does not make any difference that is why just to make it compact we have introduced one more explanatory variable x 0 which is the which is we have assumed it to be 1 and so the logit will take a very compact form.

So, what is the now logit is M summation of all β M x M say this function, what is t is this one. So, to make it compact we have just introduced x 0 nothing no great rocket (()) (08:13) science here just then our t = β x.

(Refer Slide Time: 08:20)

Now, solving for the probability p that y = 1 given a particular value of the explanatory variables. So, what is the probability that we it falls in the category 1. So, how we get that we probability we get just by using the sigmoid function is not it. So, let us go in a backward reverse direction. So, this is what is my this is my sigmoid function is not it. So, from this basically I can write it in this pattern we have already seen that.

So, basically p x is this it is nothing but the sigmoid function. So, where S b is the sigmoid function with base b see earlier, I have used e now to just to generalize it I have used it e b do not get confused with that, the above formula shows that once the b m's are fixed. We can easily compute the log odds that y = 1 for a given observation. So, once we know the value of the parameter β 0 b 1 b M.

Once you know the value of the parameter, we can easily compute what is p x is not it what is p x probability that it falls in there. What is the probability that it falls in the category 1 is not it, it is taking that there are two category p x is the probability that it falls in a category 1. Then 1 - p x is false in the category 0. So, the main use of the velocity model is given an observation x estimate the probability p x that y = 1.

So, why we can make the model, so we make the model so that given an observation x we can estimate what is the probability that it will fall in the category 1 like the example what I have given for in a fair when you try to hit the target. Given that I try to hit the target, I am trying to should from a particular fit say 4.5 feet what is the probability that I will hit the target. So, if the regression model is there, I will be able to do that.

So, now the question what we will have to find out the values of the parameters what the parameters different parameters what the parameters will take what values once we find that values will be able to given any explanatory variable given any value of the any predictor will be able to give the what will be the response, we will be able to give the probability of the different response variable.

So essentially, we need to estimate the values of the parameters. So, for estimating the values of the parameters whatever we have seen for single explanatory variable two categories same applies here also. What we have done there we have used MLE method.
(Refer Slide Time: 11:01)

So, see here also we will be using MLE method, same nothing no difference here the same MLE method just that here explanative variable is that it is instead of one here our explanative variable has M classes total sorry, M explanatory variables not one explanative variable but more than one explanatory variable that is the only difference that it has nothing other. So, we will have to find out the M will be using MLE estimation that is maximum likelihood estimation.

So, you I am sure you can remember what is MLE? So, once you know basically from that first we find out what is the likelihood and, in this likelihood, basically, when we try to multiply by the probabilities it becomes unstable so we take the log of that. So, it is called log likelihood log likelihood nothing but the maximum likelihood estimation and I am not explaining going to more details here.

Please if you cannot remember I request you please go to my second lecture of logistic regression which I have covered in details what is MLEM. So, after that once we find the log likelihood expression done to optimize the; which is the optimized value of β what we will do? We will have to differentiate it or differentiate it for different parameters and usually how do we find at optimum value.

So, build from difference first order differentiation and then we try to equalize it to 0 and once we equalize it to 0 then we get the then we will get a set of equations then from this equation we find out the values of the for particular value for which we can say that function is optimum. Here, but as I told you this the expression here that we get it is not an algebraic equation so, here it is if so, we just cannot directly equalize it to 0.
(Refer Slide Time: 12:52)

Though, I have mentioned it is putting l as well so putting this equal to 0 but because that is the standard format of standard wave when we try to find out the optimum value of a particular thing when first order differentiation then equalize it to 0 and find out the values of the parameter that is the standard form that is a do I have written it to 0. But actually, you cannot just equalize it to 0 because, it is not a finite series it is not a finite series of some algebraic operations.

So, we will have to use approximation method but you will use Newton Raphson method. Newton Raphson method is the most suitable method you can use any there are other numerical techniques also but Newton Raphson method is suitable. So, with that you will be able to find out the different β, 0 β 1 β up to β M different β values. Once you found out the β values then you will be able to find out the t that is the logit once you find out the logit then you will be able to easily find out the probability.
(Refer Slide Time: 13:45)

So, here an example is given. So, consider an example with M = 2 explanatory variable and suppose base is 10 and so we got by MLE method we found β 0 is also t β 1 is also β which have been determined. Now, here I want to tell you this MLE method is that the thing is that you do not have to do it by hand. Of course, you should know how to it is done but you as I have already mentioned before you do not have to do it by hand.

There are many software's available you will be finding many Pascal code which does all this thing you just have to feed the data and but then you should know the theory behind. So, you know when β 0 β 1 β 2 values are given then what is my logic this is my logit, this is my logit expression, 2 explanatory variables. So, β 0 + 1 x 1 + 2 x 2 this is my logit once I know my logit what is my p p of x p of x is nothing but what is p of x.

We have seen what is p of x this is nothing but e of x + 1 + e of x is not it x sorry not x t logit, p of x is $e^{t/1+e^t}$ we have already seen that. So, here what is we have considered what is t t is nothing but the β x this is our t. So, and we have considered base = 10, so this is the formula this is how we got it. Now, we have this now for any value given any given any value of x will be able to find out what is the probability that y = 1.

We have this this is our model now this is the model from this model if we have our value x because there are 2 parameters, we will we need to know the values of both the 2 parameters x 1 and x 2. And both the parameters are given then we will be able to; suppose in case of shooting the target the parameters may be distance from the target. This is one parameter and a way direction of the wind may be another one parameter.

So, accordingly both the parameters values when both the parameters values are given we will just put it here and we will get the probability. So, now see here just for this value what is given β 0 = - 3 what is β 0, β 0 is nothing but the intercept and what is β 1 and β 2 β 1 and β 2 basically it is it implies how the different parameters are affecting the response variable.
(Refer Slide Time: 16:17)

So, when we have β 1 = 1 means that increasing x 1 / 1 increases log odds by 1, how it is affecting the response variable. B 2 means increasing x 2 by 1 increases log odds by 2 when the parameter values is more does explanatory variable has more effect on the response variable.
(Refer Slide Time: 16:41)

So, now that is all from the binary logistic regressions, now we will consider multinomial logistic regression.

(Refer Slide Time: 16:51)

Multinomial logistic regressions I can tell it is a case tree where many explanatory variables and many categories. So, many categories like we can just take a simple example like based on some concentration of certain thing element in our body we can say whether there is a whether it will lead to a benign tumour cancerous tumour or it is not a tumour at all basically I can say normal.

So, there may be 3 categories, so here my explanatory variable may be there may be 2 maybe 1 explanatory variable it has 1 or many basically, 1 explanatory variable that is basically the concentration or maybe 2 explanatory variable concentration of a particular element and maybe the a's, a's and concentration will have an effect. So, these are the 2 explanatory variable.

And the different categories maybe it is a benign tumour or it is a benign or cancerous or normal means it is maybe it is not a tumour at all. So, there are three different categories.
(Refer Slide Time: 17:58)

So, now and so when the above cases of two categories the categories were indexed at 0 and 1 and we had 2 probability distribution. We had just 2 probabilities 1 probability distribution for 0 another probability distribution for 1. By now I am sure all of you know what is probably distribution if at this point if you are forgotten what is probabilistic, please go back to my lecture where I have explained very clearly what is probably distribution that is in the beginning.

Now, the probability that the outcome was in category 1 is given by p x and outcome is in category 0 is 1 - p x and the sum of the both the probabilities equal to unity. Of course, there are 2 classes and probability of 1 class is p x another class is 1 - p x and both has to equal to 1 and that was we have seen in case of binary logistic regression. In general, if we have M + 1 explanatory variable including x 0 we have seen why we have taken x 0.

So, as to make the form compact if we have M + 1 explanative variable and N + 1 categories the example recently what I have given that can benign cancel as a normal. So, there are N + 2 + 1 categories, so there are total 3 categories. We will needs N + 1 separate probability distribution. So, we will be needing if there are 3 categories we will be needing 3 probability distribution.

If there are 4 categories, we will be needing 4 probability distribution. One for each category indexed by M, which describe the; probability that the category categorical outcome y for

explanatory variable x will be in category y n. So, if there are 4 categorical variables that is four response variables and four classes response classes, then we will be having four probability distributions and what this probability distribution described.

It describes the probability that the categorical outcome y for the explanatory variable. If it is one explanatory variable 2 or 3 whatever it is explanative variable at one point will be in category y n.

(Refer Slide Time: 20:13)

So, and now if there are 3 categories the sum of all this probability has to be 1. If there are 4 categories sum of all these 4 categories has to be 1. So, it is also required that the sum of these probabilities over all categories be equal to unity that is obvious. So, now using the mathematical convenient base e, these probabilities are same format what we have used for binary logistic equation. What we have written for p and f?

So, what we have written for p x p x we have written is e^t e^t 1 + e^t, what is t? t is nothing but the logit what is the logit is t = β 0 t is nothing but β 0 + β 1 x. So, I have to make it compact let me write it β 0 x 0 β 1 x ,1 so t is nothing just simple we can compare form we can write as β x, t is β x, so this is p x. Similarly, I can write it for 1 - p x.

So, now here there is when there was only two category, we could write it this way. Now, here more than 2 categories, so what will be my p x p x again I will have to write it in this is my p x p and x just for suppose one category p and x. so, this is nothing but it will be in this format where for n = 1 to n so here I have taken 1 + 1 n + 1 / n + 1 you will see the difference here, now y instead I could have easily considered n just to make the things more understandable I have used n + 1.

So, now this is for p of n x here I have just considered p x and otherwise rest one is 1 - p x. Now I cannot do that there are an N + 1 category so it will be p 0 p 1 p 2 p 3 up to P n. So, now I am first let me consider from 1 to N so p of n x p of n x is nothing but $e^{\beta n x}$ + 1 + summation of all these values for all the different categories. So, this is the value for p n x and I have separately I have considered p 0 what is p 0, p 0 is nothing but 1 minus of this.

Just to give reflect this only I have separately used the p 0 I can use p 1 also p 2 also. Just why I have used means if I want to find out the probability of one class it is nothing but 1 - summation of all other probabilities it gives me p 0. So, that is why I have just written it in this way p 0 format and there is one more necessary I will come to that. Now, this is how I have defined p n x p 0 x.

(Refer Slide Time: 23:12)

So, now it can be seen that as required the sum of p n x over all categories is unity. Of course, of all categories it is the, note that the selection of p 0 x is defined in terms of other policies artificial. Here, we have used p 0 x you can use any one p 1 x in terms of the other probabilities. Why I have used p 0, when I use p 0 it becomes easier for me to explain it summation of because 1 minus it will be 1 - we did not back at p 1 + p 2 + p 3 up to p n.

So, I can write it in very compact from this way. So, I have taken p 0, now one more thing like when I use the odds there is one more requirement why I have separately used one classes where I am again, I am repeating you can consider any 1 p 0 p 1 p 0 any 1. So, when we have seen when we say when we calculate the odds of a particular class when we calculate the odds of a particular class how did you calculate the odds of the particular class. So, we will just see how we have calculated the log odds.
(Refer Slide Time: 24:22)

So, when we calculated the log odds so, log odd of say p 1 x p 1 x was what? p 1 x was number of occurrence of 1th class by number of occurrence of 0th class remember. So, that was my that is how I calculated my p 1 x. So, basically now if I should not write it this way, this way it is becoming very complicated. So, like my p x I will tell it. Here, this is p 1 x / p 0 x, so this was my log odds.

Log odds that it falls in the category 1 given a variable given the explanative variable x it falls in this category 1 this is my log odds. That is the number of occurrences that falls in the category in the one divided by the number of occurrences that falls in a category 0, this gives me log odds. Now, when I have more than two categories then what I will do is that I will use one of the classes one of the response classes as a Pivot class.

Or we can call it regress class basically other class other classes, I will regress it based on that class. The other class I will regress it based on that class. Now, suppose I have 3 categories what are the sub the example what I have given that based on the concentration to there are three different types of it can be 3 different type of cancers and 3 different type of classes maybe it is a benign tumour maybe it is a cancerous or maybe it is normal it is not a tumour at all so there are three different classes.

So, when there are 3 different classes then let me find out how will I find out the log odds of whether it is a log odds of benign, log odds of cancers, log odds of normal. So, 3 different log odds I have to find out. So, in that case how do I find out is like in the when there are 2

categories I found out the log odds of just one category that is p 1 and other, I did not find that is 1 minus of that and that only gives me that.

So, now when there is three categories I will find out the log what is up 2 categories, when there are n categories I will found out the logouts of n - 1 category. How do I find? For finding out the log odds from any category, I will be using a one reference class that I call it a p both class. So, based on that class I will be basically regressing my log odds whichever class I trying to find out I will regress that based on that reference class or Pivot class whatever it is.

Now, here let me con use that I am using p 0 as my Pivot class. So, if I use p 0 as my pivot class then what happens you see here.
(Refer Slide Time: 27:05)

The special value of n is termed as a pivot index and the log odds t n are expressed in terms of pivot probability and are again expressed as a linear combination of explanative variable. So, t n and for that it belongs to the log odds that it belongs to the nth category this I am writing it in this way. Log of p n f y / p 0 x where my p 0 x I am considering it as a pivot index, that I am considering that as a reference class my 0th class I am considering as a reference class.

Let me consider in here may be my benign class and maybe I am considering as a pivot class. Now, when similarly suppose let me see if I have such a way if I have three 3.
(Refer Slide Time: 27:47)

So, how will I do and let me again consider my 0th class as the 0 what to say 3rd class by normal class if I consider as a pivot class then my I will be having three differences. That is log of 1 is p 0 of x p 0 of x by if I am considering, sorry I am not considering p 0 this is p 1 p 2 p 3. So, p 1 x / p 3 x, so this is my logit of t 1 log it of t 1 is this. Now, my logit of t 2 = log of p 2 x / p 3 x, so I got log it of t 1 and t 2.

Here basically you see I have compared p 1 with p 3 and p 2 with p 3, see if compared p 1 with p 3 p 2 with p 3 which comparison is left here from here p 1 and p 2 combination is still not there. What I want to show here is that whichever whatever reference class you use either here I have used reference class as a p 3. Whichever reference class you use it is immaterial, it will be getting the same value.

Because, see here I have used the reference plus but I have done also using just 2 logits I will be able to get all the logics of all other classes because see here in t 1 I have compared p 1 with

p 3 and in t 2 I have compared p 2 with p 3. Now, what is remaining is my remaining is p 1 and p 2, now if I do t 1 – t 2 what would I get you see t 1 – t 2 what will I get des log of this - log of this what will I get log of p 1 x / p 2 x.

So, see using this I got this comparison as well. So, whichever may be the logit reference class whichever maybe the pivot class you consider that is immaterial in a way you will be basically trying to compare or while finding out the odds basically you will what to say. It will compare the odds for all the classes you will find out the odd odds for all the classes. So, if there are three the categories.

So, I will find log it is 2 logics if there are four categories I will find three logics so basically if there are k categories so we will total find k - 1 log it is. So, when I find k - 1 logit that means I will be getting k - 1 expression. So, now from the logit is from this logged in term is in terms of the parameter you can see β 1 β n x here is not it logit is always in terms of the different parameters.

Now, to finalize my model I will have to find out the best values of these betas. So, how do I find out the best values of betas. So, I will have to again the same technique I will have to basically find out what to say I will have to do the first order derivations of my problem likelihood function.
(Refer Slide Time: 31:14)

So, what is my likelihood function here the likelihood function is the joint probability distribution of here the likelihood function is the joint probability distribution of all the categories all the different stores in the since there are and probabilities and different categories. So, for n different categories I will be having and different probabilities, so it is the joint probability distribution of all categories. So, I will be finding my; that is my total function.

So, once my total function is there, I will have to find out the log likelihood function. Once you find out do the derivative of this function likelihood, log likelihood function mind it. We first find out the likelihood function, likelihood function is nothing but the probability distribution of all the different classes. So, then we will find out the log likelihood and then we will have to optimize this log likelihood function.

So, to optimize this log likelihood function how we will do will first do derivation first order derivation and first order derivation based on the different parameters.
(Refer Slide Time: 32:05)

And once we found out the first order derivation on different parameters then we will find out the and we will equalize it to 0 and find out the different probabilities, different values for the different parameter. Similarly, the same technique is applied here also, well so what is basically, we use a function we use a special term this is called an indicator function which is equals to unity if y k is = n and 0 otherwise, so now this is my log likelihood function.

So, this function I will use the derivative of that and I will try to find out the parameters of that. So, again for this it is better to use by software, software code is available pascal code is available.

(Refer Slide Time: 32:47)

So, basically you will have to derive this term using all the parameters. So, what does B nm see I am differentiating with respect to β nm what is β nm β nm is the mth coefficient of the m β and vector. See we saw here β is a vector because there are different categories so β is a vector and what is m even this is also a vector, we have m variable. So, β mm is the mth coefficient of the β n vector and x m k is the mth explanative variable of the kth measurement.

So, differentiate with respect to β and m for different values of n different values of m and then we can find out by using Newton Raphson method you can find out the values of the different betas. Once you find out the values of β then finding out the probability is nothing just put the values in the logic function and once you know the logic function, we can very easily find out the properties of the different categories.

(Refer Slide Time: 33:50)

So, now coming to the different examples of logistic regression and logistic regression it has applications in various domain. So, in medical domain it has tremendous applications, so once this application we can see in the tree school that is trauma and injury severity score which is widely used to predict the mortality in injured patients was originally developed by Boyd et al using this is used logistic regression, this is a very well-known process.

Then many other medical scales used to access severity of a patient have been developed using logistic regressions. Again, logistic regression may be used to predict the risks of developing a given disease like diabetes, coronary, heart disease based on different parameters what is the risks of the developing diabetes then it may fall into the different category. So, based on the observed characteristics of the patient like age, gender, body mass index results of various blood tests etcetera.

(Refer Slide Time: 34:46)

Then, in social sciences also there are lots of examples of logistic regressions. Then in engineering and engineering basically, this is I have specially kept an example from the reliability engineering prospect. This technique can be used in especially for predicting the probability of failure of a given process system or product given the values of the different variables.

Variables means different conditions, maybe for a system maybe for a system if we see the values of the different like the system may have capacitor inductor resistor many other components. So, it is different values at a certain point of time it is different values, how this will affect the you know what to say failure of the system, failure we can keep it in different categories prone to failure very much prone to failure that way we can keep in different categories.

And these different values of this will defined in which probability of system being in which state at that instant of time, similarly, it has application in marketing domain.
(Refer Slide Time: 35:47)

Then, it has also application in economics, in economies it can be used to predict the likelihood of a person ending them in a labour force and business application would be to predict the likelihood of a homeowner faulting on mortgage. So, and here I forgot to mention one thing do I have that. So, like in bank transaction whether it is a customer is a fraudulent customer, a genuine customer they say logic revision is used to find that out. So, in natural language processing also logistic regression has it is applications.
(Refer Slide Time: 36:20)

Then, coming to the conclusion we learned a multinomial logistic regression. So, now we have from in three classes we have learned logistic regression. In the next lecture we will cover a tutorial on the logistic regressions.
(Refer Slide Time: 36:34)

So, this is the reference and thank you guys.

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

Lecture - 48
Tutorial on Logistic Regression

(refer time: 00:30)

Hi guys, so today is the tutorial lecture. So, first as usual first will be starting with some objective type questions and then we will be doing some problem as well.
(refer time: 00:39)

Now first coming to the objective type equations. Objective type questions always I am repeating it again do not see the do not click the means do not see the what to say answer immediately first try to answer yourself, then you can understand your level of understanding. So, which of the following statements is or are true or false. So, logistic regression may be used for classification, so is it true or is it false?

Say logistical equation basically we have seen it is basically it is not a classifier but of course we can use this for classification. How? If it is a two cluster so we can have a threshold value say 0.5 if my value is probably is 0.5 or above then I can sell it is it in a class one if my value is less than 0.5 probability then I can say it is in class 0, if it is a two class. Similarly, if there is three class accordingly, I can set two threshold for the demarcation between these three classes.

So, it is not a classifier per se but we can definitely use it as a classification as also so logistic may be used for classification this is true. Next is the logit in the logistic regression is a linear function in terms of independent variables. This question I have a objective behind keeping this because whatever I have discussed till now all you have seen the logit is a linear function of the independent variable.

But that is actually it is not always the case that logit is a linear function sometimes there is a non-linear logic maybe a non-linear function in terms of the independent variable as well. So, in the example what I have taken is just linear function. So, that is why logit in a linear function a logic in a logistic regression is a linear function in terms of independent variable that is not true it may be linear it may be non-linear, so this is false.
(refer time: 02:41)

Which of the following is necessary for auto regression analysis? Remember auto regression, so the dependent variables should be normally distributed, applications to both periodic and a predict time series data. Time series data should satisfy non-stationary property, time series data should satisfy both stationary and uniform property, which one is following is necessary for auto regression analysis.

Auto regression analysis we have seen for auto regression analysis our data is not a simple data it has to be time series data. And we have also seen and for auto regression we have also assumed some we have made certain assumption about the time series data that is. What is the assumption? It should satisfy both stationary and uniform property remember what is stationary property stationary property means?

First let us talk of uniform property what is uniform property? Uniform property means suppose we were considering from say 2000 to 2022 we are considering monthly data so that means so when we are considering from 2000 to 2020 monthly data that means we need 12 months 12 data for each year. So, there should not be any missing months data or in some cases there will it is given instead of months it is given days data no we do not need that, it should be any form that is the uniform property.

And stationary means the probability distribution should be stationary it should not vary with (04:13) so that is the stationary property. Probability distribution of means probability for each time each data for each data it should have the same probability distribution. So, probability distribution for all the data of the time series of different years maybe if we consider years or months or whatever it is, it should be it has a uniform probability distribution.

So, that is the stationary property. So, time series data should satisfy both stationary and uniform property. Now if you treat to see the other options the dependent variables should not be normally distributed the dependent variable should not be normally distributed. We did not in time series data we did not talk about the dependent variable distribution of the dependent variable because that would have come.

When we would have specifically talked about the distribution of the dependent variable but that was we did not assume anything of that so, this is definitely this is not necessary, applicable both for periodic and a periodic time says that are definitely logistic auto regression is applicable only for periodic data it is not applicable for a periodic data. And time series data should satisfy non-stationary property that is definitely when it satisfies when our assumption is just non-stationary will be (05:31).
(refer time: 05:31)

The auto correlation is defined as correlation of a given time series with another time series data definitely not given time series it another time series data. How? Why? No, the correlation of a given time series with lag value of another time series data that is also definitely not it may be lagged value the same time series data. The correlation of a time series data with its own lagged value, this is correct. The correlation of one time interval data with another time interval data that is also not correct.
(refer time: 06:06)

Logistic regression transform the output probability to be in a range of 0 to 1 we have seen logistic regressions that is the our model we need such a model that if we get the output probability will be in the range of 0 to 1. Given an input our output would be a probability value it should be value within the range of 0 and 1. Which of the following function is used by logistic regression to convert the probability in the range of 0 and 1?

So, which of the following function we use? We use sigmoid function, hyperbolic sine, logarithmic, hyperbolic tan, which function we use? Definitely we use sigmoid function because sigmoid function f x f x outputs a value between 0 and 1. If we consider hyperbolic sign, hyperbolic sign what is the how is the hyperbolic sign it is something this way this is the hyperbolic sine curve.

So, and similarly if we consider hyperbolic 10 that will also that is also something this way so definitely it will not be in the range of 0 and 1. Logarithm to definitely not base log what is log of zero all of you know that so logarithm function is also definitely not so it has to be when we since we in logistic regression we want our output to be in the range of 0 and 1. So, sigmoid is the function for that.
(refer time: 07:39)

Which are the following options are true? For linear regression error values have to be normally distributed but in case of logistic regression it is not the case. This for logistic regression error values have to be normally distribution but in case of linear regression it is not the case. For linear and logistic relation error values have to be normally distributed. Both linear and logistic relation error value have not to be normally distributed. Which is the correct answer?

See here for linear regression error values have to be normally distributed that is very true for linear regression we have assumed that a error value is normally distribution it has a constant variance with σ^2 is not it but its logistic regressions we are not considering the error values here we are just giving the probability we are not considering the error values and we are not considering that it is normally distributed or whatever for linear regressions we are considering that so, answer is the correct answer is a.
(refer time: 08:39)

Then which of the following is a sigmoid function? this very easy which of the following is a sigmoid function? This is the sigmoid function, is not it? This also we can write it in this form e^x / $1 + e^x$ same.
(refer time: 08:58)

So, what is the value of the sigmoid function already we have seen so value of the sigmoid function as x equals to ∞ as x goes to ∞ sigmoid function it gives 1 as x equals to basically minus ∞ so it gives 0, so 1 and 0 is respectively.
(refer time: 09:19)

In which of the following case of logistic regression the target variable can have three or more possible values without any order, the target variable can have three or more possible values but without any order there is no ordering but it can be more than two basically if more than two definitely it is not binary if model 2 definitely it has to be multinomial but if it maintains a order then we would have called it ordinal logistic regression if it maintains order. But here it is selling without any order so it is the multinomial logistic regression.
(refer time: 09:57)

The ratio of the probability of event occurring to the probability of the event not occurring event occurring to the problem not occurring not that is odds and log of odds its logit.
(refer time: 10:12)

Which of the following is a time series problem? Estimated number of guest room booking in next 6 months estimating total, next six months what will be the number of guest rooms booking we want to estimate that is it a time series problem? Definitely yes because to estimate that we need the value for previous years data or previous months data based on that we will find out it is in all this periodic data will need based on that we will be able to estimate that this is basically very much a time series problem.

Estimating total sales in next two years of an insurance company this is also very much a time series problem. Estimating the number of calls for next one week that also we will estimate it based on the regress data based on the lag data is not it. So, this is also a time series problem, so which is the answer all three other time series problem so 1, 2 and 3 this is the answer.
(refer time: 11:13)

So, now we will see certain problems. See here the question looks a bit big but nothing to worry very simple. In a study of urban planning in a country a survey was taken of 50 cities 24 use tax base funding and 26 did not out of 50 cities have surveyed 50 cities out of 50 cities 24 cities have used tax-based funding and so there is in software writing is TF and 26 did not use tax base funding.

One part of the study was to investigate a relationship between the presence or absence of TF and the median family income of the city. So, is this tax-based funding as it is depend on the medium and median income of the family. So, there are total 50 cities out of 50 24 use tax-based funding for different work maybe different what to say different type of work for setting up of some public organizations, different utility anything it used tax-based funding and for 26 other cities it did not use.

So, now the study is to investigate under which sort of city are using tax based funding, where the median income is greater there we are using tax base funding is. Basically, is there any relationship between the tax base funding and the family income of the particular city. So, data is given in table one with median income in order of 1000 dollar so here it is in order of 1000 dollar so here you see the income here it is given access 9 point it is in the order of 1000 dollar, so 9.2, 9.2, 9.3.

So, when this sort of data's are given so you see it is very difficult to find out for the 9.2 there are two there are 9.3 9.4 9.5 very small intervals it is very difficult to find data for each and every data likewise. So, in if when the data in such a way in statistical method we usually group the data. So, first thing is that we will group the data we will make certain group maybe from 9 to 10, 10 to 11, 11 to 12 like this whatever group if you find out whichever group is best.

So, accordingly we will we have to find a grouping in such a way so that our group also group should not be too many at the same time groups should not be too less also. If the group is two less then too many members will the frequency of the group will be will increase. So, that way we will have to make and make the grouping, so first for this we will make a grouping.
(refer time: 13:45)

So, here we have income category from 9 to 10, 10 to 11, 11 to 12, 12 to 13, 13 to 14, this is how we have made the income category and then midpoint of this income category 9 to 10 midpoint is 9.5, 10 to 11 is 10.5. So, now in this category how many is are number tax based funding zero we are indicating for one of the classes zero here 0 is for tax base 0 is where city which does not have tax base funding and one for tax base funding.

So, in this category there are total 13 is non-tax based funding and one is tax-based funding. So, we are trying to find out log of 1, log of 1 what it will be 1 / 13. So here similarly here for this case 10 to 11 category number of zeros at three that means total three cities which are in this income range in this income range three cities which has which does not have tax base funding four cities have tax base funding.

So, what is log of text based when we find out odds of tax based funding so it will be 4 / 3.

Similarly, for this 11 to 16 in this range odds of tax base funding will be 6 / 6 this is also 6 this is 6. And log of tax base funding for this range it will be 10 / 4, similarly this is, here you see in this range from 13 to 14. So, log of tax based funding will be 3 / 0, when there is a zero term in any of the frequencies see here it is given.

One of the cases has zero occurrence creating an undefined ratio odd ratio that is 3 / 0 it becomes undefined. Since a log of odd is undefined, we followed a common practice of adding one to both numerator and denominator counts in a calculation of all the odds. In such case when there is zero occurrence what we do is that we add one both to numerator and denominator. So, that and it will not make a difference if we add 1 to both the numerator and denominator.

So, that way we can avoid the undefined results. So, this is similarly we found the logouts of this so we have for now we have the logouts for different values then now the question is that we have the logouts. Now what we need we need to find out the parameters value? How will you find out the parameters value? Parameters value will be using maximum MLE method. So, one other way also we can use parameters but that is we will we may not get the best parameter it is just if we plot the value.
(refer time: 16:30)

See here we have plot the values midpoint of income because what is that independent variable here what is the explanatory variable here explanative variable is the income category. Is not it? So, here we have taken the midpoint so this is my explanatory variable. So, I will plot it based on this is my explanatory variable and this is my log odd so I will plot the data. So, by plotting the data if we could fit a straight line then it is very good.

Then from there or if the maximum points if I can find out the line which crossed or the maximum point then from the line itself, I can find out the intercept I can find out the slope and then that will give me the values of the different parameter that is one way. But of course, with that it is we cannot be sure that we got the best value of the parameters. Another wave is that we have the log odds value we have the since independent variable.

So, we can have different expression, we can have different equations, two unknowns two equation we can find out a value. But then from this many number of logouts we can have if I take two to each so I will be having different combination of equation. From different combination of equation, I will see I will get different β 0 β 1 values now which β 0 vitamin I will consider which one is best I need to consider which one is this.

So, best method is go for MLE and then you find out the β 0 and β 1. Now suppose here we

found out the β 0 and β 1 value we found this is my β 0 value and this is my β 1 value. So, how we have found out by using your software to use you definitely use MLE what are your software and you found out this is your β 0 value decision β 1 value. Now see β 1 value is 0.8529, what does this indicate? It indicates a strong correlation.
(refer time: 18:23)

So, does the logistic model chosen is in this form where x is the city's median income odds and odds is the probability that a city will have tax base funding divided there probably that it will not have tax base funding.
(refer time: 18:34)

So, we got this is my β 1 value, so we find that there is strong evidence of a relationship between the tax based funding and the median income because this is β on value is quite high value it is not close to zero it is close to one. So, when the β 1 value is the high value then we can see that a strong evidence of relationship between the tax base funding and the median income.

So, more value cities have a higher probability when there is a evidence that means when the median income is more that means cities have a higher probability of adopting tax base funding. So, now if the city has a median income of 12,500 the probability of adopting TF is just put it in the value this is how we got the T value once we know the T value, I can find out the $e^{t} / 1 + e^{t}$.

So, this is the probability so there is 0.704% probability that the city with 12,500 median income will adopt tax based funding. So, once we have basically logistic regression once we find out any other regression model basically once you find out the parameters then there is nothing. Just find out the parameter, once you find out the parameter then given the predictive variable find out a response variable that is all.

Any of the regression analysis whether you call simple linear regression, auto regression, logistic regression whatever it is just that logistic regression gives a probability and other regressions gives us a value, that is all.
(refer time: 20:05)

Now next question Time Magazine used data from USA to compare whites and blacks opinions of the death penalty. The data consists of response from 32937 participants collected from 1972 to 1996. The outcome variable is whether the respondent did or did not support a death penalty. The survey provided a table of the percentage of whites and black each year that supported the death penalty.

So, in different year what is the percentage of people who have supported the death penalty is

given white who have supported the death penalty and number of blacks who has supported the death penalty percentage basically percentage is given, so what we need to find out?
(refer time: 20:54)

It is given comment on any if there is a trend in time does it appear linear or quadratic, so convert the percent is given into table so log odds within each race and year and plots log on versus year. Comment on a pattern you see is there a trend in time does it appear linear or quantity. So, we have to see whether how this what to say how does opinions with respect to year it is here is the independent variable their opinion is the dependent variable is there any relationship between that.

So, and is if there is any relationship is it linear or is it quadratic? So, what we will do that? We will first find out the log odds of white and log odds of black we will do it separately.
(refer time: 21:51)

First, we will find out this is given white so we will find out the odds which supporting the death penalty odds of white. How do we find out odds of white? Percentage of white person supporting percent of white person opposing, a 57.4 is supposing the 100 - 57.4 is the opposite. If you would have found out the probability what we had have done 57.4 /100. So, now it is odds 57.4 / 100 - 57 that is how we find I found out the odd of 1972.

Similarly, I found the log odds similarly for all the years we have found out so all the years we have found out the log odds.
(refer time: 22:27)

So, once we found out the log odds then basically once we found out the log odds then what happened we found a logit function then we will have to find out the MLE estimation MLE and then estimate that parameter this is we have estimated the parameter. This actually I will not go in details of this question because I have not explained here non-linear when the logit is non-linear we have only seen the logit is linear.

This example we have I have solved it some using the software. So, I could just show you the result. So, here once first you assumed as a linear logit as a linear on the function of all the parameters and explanatory variables linear function of all the explanative variable basically that is why we have used the parameters. So, then using MLE methods we found the parameters and um we found out the parameter that is one model we have seen.

Using the that model we got that model then we wanted. Here the question is that does it appear linear or quadratic. As I told you so sometimes just from the plot, we can we can say whether it is a linear the dependency is linear or non-linear. But sometimes it is very difficult to

tell just from the plot. So, what we have to do first we will definitely try linear and then we found out the R^2 value.

Remember we have talked of R^2 value; R^2 value is the quality of fit. What is R^2 value? 1 minus of if variable in error by the variability in the data. Remember we have done that 1 minus of sum of that is variability in the error divided by the variable in the data that we have variability and error and variables in data we have taken in terms of sums squares we have discussed that.

So, we found out the R^2 value more R^2 value it is a better is not it, so we found out for linear we found out R^2 value. Then we will try for non-linear, so non-linear is the say $\beta_0 + \beta_1 x_1 + \beta_2 x$ if I have just one parameter lamb $\beta_0 + \beta_1 x + \beta_2 x^2$ so it is non-linear of order two. So, again using MLE we found out the parameters of β_0 β_1 β_2 then we found out again within found out the R^2 value we got a particular R^2 value.
So, if we found out that the quantity R^2 value is better than the linear R^2 value then definitely it is not linear it is non-linear. Now non-linear means it is of degree 2 or degree 3 first we have tried with 2, then we will see with three non-linear with degree 3 we will some using degree 3 we will find out the parameters then we will find out the R^2 value. Suppose using degree 3 with our value of R^2 reduces then we understood that no.

It is that means it is non-linear and it is of degree 2. So, similarly for this example we have found out the R^2 value for quadratic is 0.842 weight and for linear is 0.6532. So, it is from this you can see that it is a logistic linear model whose logit function is quadratic because this R^2 value is greater than the linear value.
(refer time: 25:41)

So, similarly we will do for black for black also we will find out the odds of the black odds means basically black with supporting the death penalty. So, how it will be here if we see it is 28.8 odds will find out 28.8 / 100 - 28.8. So, then we will find out the log odds so once we find out the log odds. Then we will get the logit function now we have to find out the value of the parameter same way we will find out the value of the parameters.

And once we found below the parameters, we got the regression models then we try to find out the R^2 value.
(refer time: 26:21)

So, similarly for black also we will find out the R^2 value, here we have seen for our black also when we try to find out the opinions on the regarding the death penalty of the black. Here also we found out that R^2 quadratic as quadratic R^2 value is greater than linear R^2 value. So, this is also it is quadratic line fits the data more even from the plot also we can see that so, with that we come to the end of this lecture.

(refer time: 26:47)

So, then thank you guys.

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

Lecture - 49
Introduction

Hello guys, so we have come almost to the end of this lecture like this course basically, so this is the last module which we are going to discuss, this module is called classification. So, classification is something which all of us know it is something which we use in our day-to-day life in our day-to-day activities we get to see that we classify things we classify objects, even as a child also when initially when the style when the child just started and starts learning things.

So, basically at that time the child does not know what is if I talk of animals the child does not know whether it is a cat or a dog so maybe the mother teaches the child see this is a cat, this is a dog and so basically once he or she gets the when the child gets this knowledge that this particular animal we call it cat, this particular animal we call it dog and so next gradually when the child grows up the child can automatically whenever a child sees an animal which the child has already known the level of the animal.

The child can identify, this is a cat, this is a dog this is a cow whatever it is. So, desires is this classification it starts from the very beginning of her life. So, this classification we get to see in many aspects and coming to the technical aspect of it like when why we are going to learn classification, basically and in many of the disciplines I mean engineering disciplines many social science research many things in many domains medical domains we really need to classify some observations.

Classify objects; classify data we need to classify data based on some classified data into some basically relevant categories. So, now the thing is that so why do we classify data to some relevant character categories we can classify object to some relevant categories. So, we classify some observations different kind of some emails this may be various reasons maybe some for some medical diagnosis maybe some medical research from social science research and then maybe some business perspective for personal use.

And maybe for security reasons for different types of data we may have different security requirements like say mail we may need to classify a spam mail and a non-spam mail, so, this all this is what is called classification. So, we are going to learn definitely all the algorithms is not possible to learn in this course. So, we will be learning a couple of algorithms how we classify different object different observations different data.
(refer time: 03:17)

So, in this lecture first will be covering the introduction to classifications. Then there are two

broadly there are two different categories of classifications that is we call it supervised classification and unsupervised classification we will be discussing that in this lecture. And then finally I will give a formal statement of supervised classification technique. In this lecture I will restrict my description discussion to only supervised classification technique unsupervised classification techniques I will not be discussed.
(refer time: 03:51)

So, now coming to the; introduction of classifications. First now see there is a picture I am asking you to identify the objects. Now how you will identify these objects? Of course, as now you already know what are these objects so that this is an apple, this is an egg, this is an onion this is a potato all of us we know we can directly identify these objects. But the thing is that how did we identify this object.

This all these objects it has some characteristics it says some features or parameters or attributes whatever you say. It has some distinguished feature each type of classes when I call it a onion and when I call it a potato each has some distinguished characteristics that which we call it the features. Based on these features we try to classify the objects, this is an onion, this is an potato, this is an egg based on these features we do that.

So, now if we need to identify these objects since this and it is already we already know what is or when we call a thing as onion so when we see the peaches matches then this is an onion, that is how we identify the object.
(refer time: 05:01)

Now again there are some other way also where sometimes when we identify some data but some classify some data they may be it may be in and not I should not say this characteristics basically it is some sort of rule using some sort of rule sometimes we classify data. And rules means it is again specified some sort of features as well like here. If you see we are classifying the marks of the students we all we as two teachers we always do we give the grades to after exams we give the grades to the students how based on the marks.

When the marks are greater than equal to 90 we give a greater excellent great whatever it is. From 90 to 80 we give big grade so this is nothing this is again one sort of classification only here how we do the classification. It is directly by the rule base it is very much deterministic does mark above this is a great between this and this is B grade this and this is C grade it is a very deterministic rule.

So, are there based on the deterministic rule we specify these are the different classes of marks different classes of students grades. So, this is one way by which we can classify data. Once we have seen we can classify data based on some features different its distinguishing features, distinguishing attributes and another is that we can also classify this way by based on the some rule base.

If this how we are framing the rule basically again basing based on some features, features may

be this we can if we can consider this marks greater than 90 this is a feature. So, basic but there is nothing is just I have made a specific deterministic rule for that.
(refer time: 06:42)

So, now again some example if you see like from life science, I already mentioned in classification is very much to use in case of medical research medical diagnosis like predicting tumour cells as benign or malignant how do doctors how do Radiology does that. They realize we do MRI based on the MRI data this is what they basically do some sort of image processing. So, how do they do? So, there are different features.

So, they I did not they try to identify different features and they have the knowledge that these features meaning means these features implied this. So, based on the different features when an image is given so based on the different features they try to identify this is an benign tumour, this is a malignant tumour and then next coming to some example of security credit card transaction at the legitimate or fraudulent credit cards.

Sometimes what happened usually an average person and every person uses credit card every maybe once in a week or every two second day or every third day and even average person average middle class person takes out money. Say around every week if a person uses their uses maybe the card ATM card and then withdraws money say around 5000 or around 10,000 per week or every alternate day.

So, suddenly and this for the same card if the if it is founder it is withdrawing and lots of money in a very short interval so every half an hour it is at except hours suppose the person has withdrawn 1 lakh again after half an hour it is withdrawing another 2 lakh again after half an hour or one hour to withdrawing 3 lakhs very frequently it is we are drawing a huge chunk of money.

So, and this machine can understand that this is this may be in fraudulent transaction maybe some the credit card has got stolen and that is why the people the person who has who got hold of this character is using it this way. So, this is how this transaction can be classified, this is a legitimate transaction this is this may be a fraudulent transaction. Similarly, in case of it has many applications classification is something it has a varied application.

Like now again we can think of in the weather prediction like satellite what has the satellite collects data it collects the data of a humidity, wind speed, pressure, random and many other factors like it collects different data and then based on this data it can classify the next day it will it will fall in which category. It will be a cloudy day sunny day whatever it is. So, these are the different features they try to take the value of the different features based on different features it identifies to identify it.

Similarly, entertainment category new stories as finance, weather, entertainment, sports etcetera. You have might be you have noticed when you if you a user of YouTube If you use YouTube usually in YouTube a personal profile gets stored profile in the profile meaning the person's browsing history gets stored. And so, when a person if I load the YouTube and if some

other person load the YouTube they have informed the videos which I will get it will be different from what the other person will get it is.

Because on the things the staff I browse it basically tries to keep a history of that. And based on that it categories into my choices it has been categorized into a particular classes maybe I see lots of news stories. So, when I try to load YouTube, I will get to see a different types of news in my in the screen. Similarly, some other person maybe whenever he opens the YouTube, he usually watches movie so maybe we watch this new movie.

So, whenever he loads the YouTube whenever he opens the YouTube, he will get to see all the different new movies that has come that has been loaded. So, that is how it categorizes the persons based on the features of the person browsing history it categorizes the preferences of the person, so this here also we use classifications.
(refer time: 11:19)

So, now let me formulate we have seen lots of examples of classifications. So, let is formally define what is classification. So, classification is a form of data analysis to extract model, it is a form of data analysis to extract model describing important data classes. So, it is a form of data analysis to extract models will have will have to analyse the data and we will have to extract a model we will have to find a model which basically it is a black box a model.

What the model describe it will describe important data classes. So, given an input it will give the which data it will it belongs to which data class C 1 C 2 or C 3 which data class it belongs to which data class. So, that is the formal definition of classification. Essentially it involves dividing up objects whatever objects or observations or data whatever it is, so, that each is assigned to one of a number of mutually exclusive and exhaustive categories known as classes.

So, given an object it will be assigned to one of the mutually exhaustive and exclusive categories known as classes. Means when given an observation given an object it will be assigned to only one class it will not be assign to one object will not be assigned to more than one class and each object will be assigned to one of the classes at least it will assign to at one of the classes and one object which is assigned to one class it will never assign to some other classes.

So, it is mutually exhaustive and exclusive. The term usually exhaustive and exclusive simply means that each must be assigned to precisely one class that is exclusive each object; must be assigned to precisely one class. And that is never to more than one and never to no class at all never to no class at all, this exhaustive.
(refer time: 13:32)

So, now as I told before the classification consists of assigning a class level to the set of unclassified data and this classification broadly it can be classified into two different classifications. We can have broadly there are two different techniques for classification technique if you see broadly. So, one is called supervised classification and the other is called

unsupervised classification.

Now what is supervised classification? Supervised classification the data with possible class of each is known in advance. So, like as the example what I have given initially when I started with classification a child is taught this is this element and this is this animal this is cat this is dog. So, that is in supervised classifications. So, the child tries to put it in the child's in the brain however these are the features when these features are there and it is called a cat.

When these features are there and it is called a dog. So, this is the data features are known as well as its class level is known for a particular set of features this is the class level this then we call it a class supervised classification. Like your apple suppose it is round shaped it is red coloured and I mean what whatever may be the features whatever and so then this is called an apple. So, it is an again round shape it is yellow colour then maybe it is an orange.

So, it is based on this attribute is round shape red colour round shape yellow colour there are what these are attributes these are features and what is the label, label is apple orange these are labels. So, data with possible classes of each is known data attributes based on the attributes what is the class it is known. And then in such a case and then in next when we given data with an unknown level then it will be able to identify this belongs to this class.

This is called supervised classifications where the classes level are known in advance how many classes are there that are known in advance. The next is unsupervised classification. Unsupervised classification is something like class labelling of data is not known given a data set we first of all we do not know how many classes will be there. What will be the different class level? Nothing is known to us, the data are given.

Somehow when the data’s are given we try to group the data in such a way, the big data set is given we try to group the data in such a way the data belonging to a group are very much similar to each other compared to the of means all the object belonging to this group are quite similar to in each other in its features whatever may be in some of the features then compared to the object belonging to the other group.

So, like I can give you an example one unsupervised classification a very good example classic example actually it is used from long time back in data mining we basically use this customer segmentations like in what to say in a big retailed retail shop. So, daily data daily data is there or one monthly data we see monthly data used data many customers has come many customers has purchased many things.

So, this the used customer purchase data is with us so we do not know from this what sort of classes we can bring out and how many classes we can bring out nothing is that the data what is this data this customer purchase history. From this if we try to group data and group data and there can be different grouping now if as I told you it tries to form a group. So, that the characteristics of the object within a group are very much same because when we compared characteristics of the object belonging to some other group.

So, like here maybe one group may be people who are buying on electronic store, then there are some people who are only buying groceries there is some people who are only buying cloth material. So, this maybe we can somehow group this way this may be one can group or maybe some other grouping maybe people are buying when people are buying one x they usually they are buying y also so this can be one group.

So, again some other group members some customers they were quite active before now they have completely stopped. So, but on the contrary, there is some customer who is quite initially they were not there was no activity now they are very much active. So, these were the different groups there is different groups basically it is called clustering. So, unsupervised cluster and classification basically we call it clustering we form different clusters.

So, this is we will not be discussing an unsupervised classification in this like course basically we will restrict our discussion only to supervised classification. So, unsupervised classification the basic idea I have given I think you could understand that.
(refer time: 18:44)

So, now it is a good example of supervised classification here different emojis are there. So, it is different emojis according to its label of the emojis are also given. So, when we see the first emojis say this is this emojis indicates a good friend. So, when we get some this sort of icon similar kind so once we know this means a good friend so if I get that means this belongs to this class. So, this is supervised learning.
(refer time: 19:10)

Again, unsupervised classification you see here the different cats. So, now how will you all are cats only, so like as I told you in a retail shop there are all our data only purchase data all our purchase data how to classify this data how to think what can be the different classes of this data. Similarly, here all are cats only now how to classify this data, how to make different clusters of this data. Maybe we can have some maybe some cats which have say blue eyes.

So, this is a cat which a blue eyes this and cat we say blue eye so is there someone else no so maybe cats which are blue eyes we can have one cluster. So, then cats with a brown eye we can have another cluster so that way if we make the cluster based on the eyes only based on some other features maybe. So, that way but cluster label is not known how many classes will be there it is not known.

Then we just try to find out we just try to make the group with the characteristic that object within a group are very much similar and it is compared to the object belonging to the other group.
(refer time: 20:26)

So, supervised when you talk of supervised classification techniques. So, now any technique first we need to know what is the input of this check first any technique first what should be the goal of this thing why at all we are thinking of this technique. So, why at all we are thinking of

this technique that is the first thing we should know then only we will start preparing things, then that is goal that is known now then what will be the input to this technique.

And what is the input to this technique and then given this input and this is the goal this is the input this is the goal now what it will do to the input so that we get this goal. So, now for supervised classification technique, what is the input? Input is a collection of records we call we also call it a training set; it is nothing but a collection of records, collection of different observation of data. If I talk it in case of statistics it is this collection of different observation.

So, each record contains a set of attributes and among that one of the attribute is a class. So, it is a collection of observations and this observation each observation has some attributes and one of the attribute is the class label. And what is the task? Find a model for class attribute as a function of the values of the other attribute that means we have to find Y = f x I have to find these functions this x are the attributes for function of this attributes this x are the attribute.

For this attribute if these are the attributes then what is the class, basically I have to find out Y = f x that is my task for given attribute x what is the class. First initially for some so that I can understand before doing going to classified as things I should understand the stuff. These things belongs these things means it is belongs to this class. So, that for that is we have already modelled we have already trained it ourselves.

So, once we have trained ourselves like the child training so then we this is our task. What is the goal previously unseen records should be assigned as a class as accurately as possible. So, task is given a set of observations, this observation has also the class label. So, from this observation we have to find this function Y = f x for this class level what is this function somehow, we will have to find this function.

Once this function is known then what is the goal? Given any unseen record with a class level is not given just the attributes are given, we have to assigned a class as accurately as possible and how we will assign a class satisfying the property of mutually exclusive and exhaustive. A class has to be assigned to one observation has to be assigned to one class not more than an observation cannot be assigned to more than one class and each observation has to be assigned to one class.
(refer time: 23:35)

So, this is an example like see there are different we have total 15 observations. So, these are the different attributes we have total four attributes. So, for 15 observation we have 4 attributes, the first three I am calling it attribute 1 attribute 2 attribute 3 and a third one and the fourth one I am calling is a class. So, for these values this is the class is no let us consider that two classes is no and yes, maybe.

First, I have say 10 observations and the 10 observation all the attributes values and the classes are given using that I will try to understand if this is the features then this is the class. Once that is known then given it tests it given any attribute value, I should be applied to tell I

should be able to tell, this belongs to which class. So, this is induction learning model and this from this model I will apply the model and I will get a deduction for these classes.
(refer time: 24:40)

So, as now to define it formally the same thing given a database D it has some tuples what is this t 1 to t n different tuples and each tuple is nothing but a set of attribute each tuple consists of a set of attribute t 1 is a set of attribute A 1 to A n t 2 is a set of attribute another set of attributes A 1 to A n different attributes. So, given a database with a set of tuples and each tuple is defined by a set of attributes.

And a set of classes given a set of tuples and a set of classes C the classification problem is to define a mapping from D to C. But this is a classification problem classification problem is nothing just a mapping from the set of tuples to a set of classes where each t i is assigned to one class it is not that a t i will have no class assignment each t i has to be assigned to one class. So, note that the tuple t i belongs to is D is defined by a set of attributes.
(refer time: 25:54)

So, now the number of classification techniques I am talking supervise only. So, some supervised classification techniques are known which can be broadly classified into following categories. So, what are the different methods we know? One is statistical based method distance-based method decision tree-based method error-based matter. So, statistical based method basically we use statistical based method.

When we basically try to identify the level of a class when based on some known relevancy based on when we know this some classes with a known level some objects. We know some you know some object with no level and some other objects we try to put classified that object under some classes based on the relevancy of the known class level. So, here in this in statistical matter basically it is not very deterministic like.

If these attributes are there then definitely it will go to this class it is not that it is deterministically, we cannot say. Like for an apple we can deterministically say that if it is if the size is of this to this much means and this colour is this shape is this and top is this bottom is this then definitely it is an apple, like deterministically we can say. But in statistical base methods such deterministic rule does not apply.

Actually, did we cannot say if this is that then this it will belong to this class. So, in statistical based method we try to just find out a relevancy of what we have to predict and what already we know. From what we know we try to predict based on the relevancy of that so the that is the reason in statistical base method usually we use probabilistic terms. So, we will see and one of the most used statistical based method is called Bayesian classifier which we will be discussing in our in my next lecture.

Then next is the distance based method. So, and distance based method here also here how we try to classify the object we try to classify the objects based on the similarity index. Like if I

tell a cat and a dog, cat and a dog similarity index both has four legs one tail then two ears so the similarity index is quite similar does not it. So, as it and similarity index means both will belong to the same group.

But at the same time there are some features which will significantly differentiate between the dog and the cat. So, there it will the similar similarity index will be quite different for some other features which will differentiate between a cat and a dog. So, this similarity index we usually define in terms of the distance of one attribute distance between two attributes one attribute from one objects and one attribute from the another objects when we if we try to find out the distance.

The similarity index we try to find out the similarity index we usually consider the distance between distances of these two attributes. So, based on this distance based method one of the approaches K nearest neighbour which will be discussing in I think the next another after two three lectures. Then next is decision tree based method. Decision based method it is or it is very much like a rule base what I have told initially and the beginning of this lecture.

So, if this done this sort of that a rule base actually. So, here what we do is that from the attribute set we try to find out one attribute which we continuous when we go on continuously splitting those attributes. So, that it forms a sort of tree. Initially if there is a we will the for the best classification first is the good selection of a attribute which will act as a root node. So, first we will find out the attribute one attribute and then we will try to based on that attribute.

We will try to divide by using some rule we will try to branch it to different other nodes then again from that node we will further go to different other nodes. So, and gradually we will reach a leaf so that is basically the classes. So, distance that is the distance tree base method however in distance tree based method will not be discussing in this lecture. So, then the fourth is the error-based method, error based method I will be discussing support vector machine I will be it is a very important classifier I will be discussing the support vector machine in a couple of lectures.
(refer time: 30:47)

And so, coming to the conclusion in this lecture we learn about a basic introduction of the classification tasks the idea of supervised and unsupervised classifications also have put and have also put this taxonomy about some popular classification technique. In the next few lectures, we look into the working of some such classification matter.
(refer time: 31:07)

These are the references and thank you guys.

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

Lecture - 50
Bayes Classification (Part - I)

(refer time: 00:37)

Hi guys, so today we will talk of Bayesian classifications and we have last class I have already given a basic introduction of statistical based classification method. So, Bayesian classification falls under statistical based classification matter. So, here we will see we will first introduce you to Bayesian classification technique what is that principles and properties we will see our Bayesian classifier.

And further in Bayesian classifier less theory of probability is very much in use we used theory probability took classify using Bayesian technique. So, the already we do we have already learned probability theory but still it is a long time that is in the beginning of this class. So, I thought I will just recapitulate a portion of this theory of probability that will quickly see go into this. So, that is all we will be discussing in this lecture. So, now coming to the introduction.
(refer time: 01:27)

So, as I already this is this figure already, I have shown you in my last lecture. So, this different classification techniques so we will see this is the Bayesian classifier it falls under statistical base matter.
(refer time: 01:41)

So, Bayesian classifier is a concept if it works like a duck, walks like a duck then it is probably a duck. As I told you in my last lecture, we try to classify the dark classify any object based on the relevancy of the some known objects known of some known object means some object with known class level. We try to predict the class of an object based on the relevancy of some objects with known class level.

So, here it is if it works like a duck, quacks like a duck then it is probably a duck. So, this is the concept of a Bayesian classifier.
(refer time: 02:22)

So, just that is its principle it is a statistical classifier it performs probabilistic prediction. That is it

is predict class membership problem is a probabilistically it classifies probability of this objects this observation belonging to this classes 0.9. Probability of this object belonging to this classifier is 0.8 so that way it gives a probabilistic prediction of the class membership. Its foundation is on bayes theorem.

I do not know if you can remember bayes theorem it is a very important and interesting theorem in probability theory anyway I will be discussing I will be having a quick recap of that. So, the classes are mutually explosive and exhaustive that is obvious in all the classification techniques the classes are mutually exclusive and exhaustive. And it has this Bayesian classifier it has certain assumptions two assumption mainly.

What are these? The attributes first assumption is the attributes are independent given the classes. So, for a class the attributes that we will consider all the attributes are independent, independent means suppose for any class suppose we have attribute a b c and for that say suppose this class is say X so these are the attribute values I should not say here X anyway forget about this. These are the attributes a b and c are the attribute.
So, these attributes are independent meaning that it is not that if I get an attribute value a 1 that will somehow affect my getting an attribute values b 1, b 1 for attribute b or getting a b 1 value for attribute b it will has somehow have an effect on getting a c 1 value in attribute c it is no not that it is very much independent. That is an assumption in real life actually that is not very much what to say it is not applicable very much this (04:17) independence of attributes.

So, but then the base classifier it considers this as an assumptions the attributes are independent given the class. And this will be more clear when we see some examples and then the attributes have equal contribution to the outcome. So, each attributes these are the attributes a b and c this all these three attributes will have equal contributions. Suppose we have for this for a particular value of a b 1 c 1 we have classified it to be X 1 for a particular value of a b 2 c 2 is classified it to be X 2.

So, this all has equal contribution attribute a attribute b attribute c all three have equal contribution for identifying to a class. It is not that if attribute b will overpower attribute c not like that so, because of this assumption this is called a naive, naive classifier because this assumption are very much unrealistic actually. So, that is why it is called a naive classifier basically because of this assumption.

However, though these assumptions are very much unrealistic but empirically it is proven to be very useful and the empirical it is found that it really gives very good results. Though the assumptions are very naive and it scales very well. When I talk it scales very well means it gives it does not need very much computation it is not very computational intensive even if we use

very much a high dimensional data or high volume data.

And both the cases high volume higher dimensional it is not very computationally intensive that is why it is a bayes classifier has also advantage of the scalability.
(refer time: 06:03)

So, a consider an example. So, from this example you will be able to understand better what do I mean by independence what do I mean by equal contribution first example. So, let us consider a set of observations which are recorded in a database there are some observation are there which is recorded in a file whatever we say. So, what data we have in a file? Regarding the arrival of air planes in the route from any airport to New Delhi under certain conditions.

But we have in this file what data we have any arrival of flight to New Delhi airport from maybe from any airport and under certain conditions. Conditions may be weather condition may be day whether it is a weekday whether it is a Sunday holiday whatever it is then condition may be the weather condition, condition may be the season whatever it is under certain condition we have tried to keep we have maintained the data of all the flight arrivals to Delhi airport.
(refer time: 07:10)

So, see here these are the observation which I am talking about. So, like these are the different observations that we have noted that. If it is a weekday means Monday to Friday and the season is spring and when there is no fog and no rain then the flight arrives on time. Then similarly weekday winter no fog slight rain then flies arrives on time. Again, here if we say if it is a holiday winter and high fog slight rain the flight arrives late.

Again, here we see if it is a Saturday spring high fog heavy rain flight cancel. So, these other data which we have kept in a file. So, this is the class level these are the attributes day season for rain these are the attributes feature set whatever you call it attribute set or feature set this is. This is the class level. Now these attributes day season for grain all are independent does the what do I mean by independence being a weak day it does not under any way influence the season obviously.

And being a season it does not anyway influence whether they will be fog or not but that is not true actually that is why I told this assumption that (()) (08:28) assumption is not very realistic. So, season and fog and rain it will it usually there is it we cannot tell this is independent there are some dependency to it. But in Bayesian classifier we consider the (()) (08:38) independent. So, an equal contribution means that the flight arrives on time it has these weekdays this day season for rain all has equal contribution.

It is not that if the day is a weekday, then definitely or irrespective of this condition usually with the flight arrives on time there is no it is not that this weekday will overpower these attributes no all has equal contribution, that is the assumption. As I told you again this is not very realistic, this is how we start with Bayesian classified now. Let us see how we will classify this.
(refer time: 09:17)

So, in this database there are four attributes. What are these attribute? Day whether it is a weekday or it is a Saturday or Sunday or a holiday then season spring season winter season summer season but rainy season than fog with a (()) (09:33) very high fog less fog no fog than rain (09:36) rain heavy rain light rain. So, these are the different attributes and we have total 20 tuples of data 20 observations.

Now different categories are on time late very late cancel these are the different categories. Now we have to form a model were given these values given the values of a weekday. What is the day? It is a sorry given the value of the day given the value of the season given the level of the fog and amount of rain and we should be able to predict whether it is the flight is on time late very later cancel.

So, given this is the knowledge of the data and classes we have to find a most likely classification for any other unseen instance. For example, if it is a weekday winter high fog no rain then what will be its class level. Classification techniques eventually to map this tuple into an accurate class. So, that is the classification task. You will have to it has to map this technique to a accurate class. What are the different classes? Classes are on time late very late and cancelled.
(refer time: 10:56)

So, Bayesian classification methodology first is in many applications the relationship between the attribute set and the class level is non-deterministic which I already mentioned in statistical based method this is one of the characteristics. The relationship between the attribute set and the variable is non-deterministic. We cannot say this is there this attribute is there that means definitely it will fall in this category in this class.

So, that deterministic is not there it is non-deterministic like from this example this is a very good example you cannot say there is a weak day that means deterministically we cannot say since it is a weekday and since the spring season the updated class it will the flight will always be on time we cannot deterministically say. So, this is one of the features of the statistical classification technique.

In other words, a test cannot be classified to a class level with certainty. Since there is no deterministic it is not deterministic, we cannot certainly certain and we cannot predict it with

certainty. This will be this class we cannot predict with certainty in such a situation the classification can be achieved probabilistically. The Bayesian classifier is an approach for modelling probabilistic relationship between the attribute set and the class variables.

It is an approach for modelling probabilistic relationship there is a probabilistic relationship between the attribute set and the class variable. More precisely Bayesian classifier use base theorem of probability for classification. In remember bayes theorem do not worry if you do not remember I will have a quick recap. So, before going to discuss Bayesian classifier we should have a quick look at the probability theory and bayes theorem. Because without knowing bayes theorem we cannot proceed with Bayesian classifier.
(refer time: 12:50)

So, quick recap of probability, so if there are n elementary events associated with a random experiment n elementary event associated within random experiment and m of n of them are favourable to an event A, then the probability of happening or occurrence of A is m / n. The probability of happening or occurrence of A is m / n. Now it is irrelevant but do you remember now, what is odd?

We have discussed odd if I would have find out what is the odd of 8 and what would have been the odd of A? Odd of A is m / n - m anyway this is not needed here or this odd concept does not come here but still I have just I just wanted to know whether they forgotten or not anyway. So, this is the concept of probability. Probability of occurrence of A is if m of n are favourable to an event A is the probability of A is m / n.

Some small example what is the probability rolling an even number on a die? Yet a random X when E is rolling a die, A is getting an even number, favourable outcomes for the even number when we rolling A die how many will be even numbers what are the favourable outcomes rolling a die even number is 2, 4 and 6 so there are total three favourable outcomes. When we roll a die how many total outcomes six total outcomes.

So, probability have given getting an even number so favourable outcome is 3 total is 6 so probability of getting an even number is 3 / 6 that is equals to half simple concept of probability.
(refer time: 14:37)

So, we have also seen what is probability of A U B. Probability of A U B is probability of A + probability of B - probability of A ∩ B. Why we do that? Because this is probability of A this is probability B suppose this is A this is B and if we try to find out A U B then if we take probability of A means these whole things get covered and probability of B means these whole things get cover probability of A + probability of B this thing sets we have contacted twice.

So, since we have counted it twice, we will have to subtract it so that is minus of A ∩ B this portion is minus of A ∩ B and now suppose if A ∩ B is 0 so if my A is like this B is like this. This is A this is B then this A ∩ B is 0 then in that case probability have A U B is probability of A + probability of B and under if what case if this will be when both are mutually exclusive.

So, if A and B are mutually exclusive then it will be something of this case this sort. So, then probability if A ∩ B = 0 then probability of A U B will be simple probability of A + probability of B it is not it. So, now if A and B are independent events suppose this is not mutually exclusive suppose this is probability of A U B as this is the formula for us. Now we know that probability of A and probability B are independent events.

So, probability of A and B are independent events then what happens. Probability A ∩ B is nothing but probability of A into probability of B, what is the probability of raining today and probability of me coming to the class. So, this is an example of dependent probability because if it strains, I may not come to the class. So, there is a very bad example I have given. Probability of any independent events now I am not I cannot think of any so immediately anyway.

If A and B are independent events probability of A ∩ B is probability of A and probability of B. Probability of me going to market (16:46) and probability of someone acts also going to market both are independent event. So, in that case probability A ∩ B is probability of A into probability of B. In that case what will be a probability of A U B probability of A + value B - probability of A into probability of B.

We have not come up dependent event dependent will come. Thus, for mutually exclusive event probability of A U B will be probability of A + probability of B mutually exclusive events when this is the case when both they have events cannot happen at the same time.
(refer time: 17:28)

Now if the events are dependent, then the probability is expressed by conditional probability. Probability of A given B given that B has already happened what is the probability that A will happen given that B has already happened what is the probability that A will happen. Now if both A and B are independent if both A and B are independent the probability of X going to market and probability of Y going to market this is very much independent.

So, given that X has gone to market what is the probability that Y will go to market. These are very much independent. So, what is the probability of Y will go to market is probability of Y? So, here probability of A given B = probability of A if it is an independent event. Now if it is a dependent event, dependent event suppose I cannot stand Y at all. So, if Y goes to market if

somehow, I come to know that Y goes to market I will try not to go to market.

So, then it is a dependent, what is the probability of A given that probability of B as have happened. So, this is dependent event. So, the events are dependent and the probability is expressed by conditional probability. The probability that A occurs given that B is given that B has occurred is denoted by this. Suppose A and B are two events associated a random experiment the probability of A under a condition of B has already occurred and probability B is not equals to 0.

Conditional probability also we have to see the probability of B is not equals to zero that is given by this, what is this probability of A given B number of events in B which are favourable to A number of events in B which are favourable to A that is nothing but A ∩ B. Number of events B which are favourable to A this is B this is A number of events in B which are favourable to A is this portion, this B this events it is favourable to A as well.

So, this is A ∩ B so, that is probability of A ∩ divided by number of events in B so, P(A | B) is P(A ∩ B) / P(B) this is how we find when these events are dependent.
(refer time: 19:44)

Similarly, so this is we from this P of A given B is P of A ∩ B / P of B so, how we can write P of A ∩ B P of A ∩ B is nothing but P of A given B into probability of B so, that is what this I can write in this way given probability of A is not equals to 0. This I can also write it this way because probability of A ∩ B = probability of B ∩ A so, I can also write it this way.

Similarly for three events this already we have done all these things just a quick recap probability of A ∩ B ∩ C probability of A into probability of give B given that A is occurred into probability of C given that A and B has occurred. And if it is independent, it is nothing but the multiplication of all the probabilities. So, if events are mutually exclusive if one happens data cannot happen so probability of A given B is ill then it will be equal to 0.

Toss up a coin if head occur still cannot occurs what is the probability that a given a head occurs what is the probability that a tail will occur that is zero mutually exclusive. So, probability of A given B if A and B are independent probability have X going to market and probability Y going to market both does not know each other. So, it is independent so probability of A give probability that probability have Y going to market given that X is going to market is called your Y going to market.

It is there is no dependency on X so, similarly probability of A given B into probability of B I can

also write probability of B given A into probability of A from this to expression. And this already we have seen.
(refer time: 22:01)

So, basically probability of A given B this already we have seen and probability B ∩ A we can write it in this way. So, this probability A ∩ B I can write it in this way basically this is nothing but the bayes theorem only. We will see there is nothing but the bayes theorem so what we have just done some twisting if initially our probability of A given B is this and we know probability of B ∩ A is probability of A ∩ B is nothing but probability of B ∩ A.

And probability of B ∩ A I can write it in this term then probability of A ∩ A given B, I got this.
(refer time: 23:02)

So, this is remember what is a total probability there is more space here let me use this here so, total when I tell it total probability I have already explained so I am still I am repeating suppose I want to suppose this occurrence this occurrence is suppose this is A. I want to find out probability of this occurrence, this occurrence this round portion I do not know what is the probability of this occurrence.

But I know this occurrence has contribution of three different other events. So, this is on this is B 1 this is B 2 this is B 3 this B 1 has contribution to occurrence of A, B 2 has contribution to occurrence of A, B 3 has contribution to occurrence of A when I know this three then with this knowledge I can I am able to find out the probability of occurrence of A this is nothing but the total probability.

This total probability, probability of occurrence of this total A (()) (23:58) I find out with the knowledge of this disjoint events these are the disjoints events. How do I find it? This probability of B 1 ∩ A + probability of B 2 ∩ this is B 2 probability of B 2 ∩ A probability of B 3 ∩ A that gives me my total A so, when what when I know probability of B 1 ∩ A I know the formula for that I know probability B 2 ∩ A.

I know the formula for that B 3 ∩ A I know the formula of order when I know that then I will be then I found out basically the formula for the total theorem total probability.
(refer time: 24:39)

So, this A is formed by this the events B 1, B 2, B k constitutes a partition of the sample space S these are the partition of a sample this is whole is a sample space S such that probability of B i is not equals to zero then for any even A of S this whole is a sample space any event A I do not know what is the probability of occurrence of A but I know the occurrence of all these events all these are all disjoints events.

Then I can find out this probability of occurrence of A how with the ∑ all this disjoints events. So, that is probability of A is this C ∑ probability of B i ∩ A that means probability of B 1 ∩ A + probability of B 2 ∩ A + probability of B 3 ∩ A plus what how many other interesting disjoints events are there. So, this is my A and I know what is my probability of B 1 ∩ A.

I know what is that this already I have seen when I have sin what is probability of A ∩ B is nothing but problem A given B into probability of B I have seen this so, this simply we will use this, what is probability of B i ∩ A? Probability of A given B i into probability of B i this is nothing but my total probability this is A. These terms I have broken into this term the probability of A given B i into probability of B i.

Why I have written in distance? Because I know probability of B i I cannot write in terms of probability of A because I do not know probability of A I know only probability of A so, I will because if I take this form that B i ∩ A is same as probability of A ∩ B i then this would have been my different end, I would have needed a term here probability of A but I do not know what is probability of A so I will not be using this.

So, this is my this expression I have written it in this form probability of A given B i into probability of B i so this is my probability of A it is total probability. This is theorem total probability now from here comes the bayes theorem. What is bayes theorem here? Now given that I A has happened given that A is happen now I know what is A given that A is happen what is the probability that it is from B 3 what is the probability that it belongs to B 3 class.

Now I can see the class concept that is coming given that these are attributes what is that it belongs to a B 3 class. Given this these attributes what is the probability that it belongs to B 5 class. These are the attribute, what is the probability that it belongs to B 4 class my bayes what is my bayes theorem, bayes theorem is that given that this A has occurred given that my A has occurred what is the probability that is due to the event B 3.

Event that probability that event A has occurred what is the probability that it is due to event B 3 what is the probability that it is due to event B 1 that is bayes theorem and this bayes theorem we can very easily use in classification how given that this are my attribute because this is A as occurred means these are my attribute (28:36) basically. What is the probability that it belongs to this class because A is the combination of all these classes.

The whatever attributes we have all the classes hidden here so this particular set of attributes I

got what is the probability that it belongs to this class, what is the probability that it belongs to B 1 class. That is how we have used here bayes theorem.
(refer time: 29:02)

So, by the laws of total probability this is what already what we have discussed.
(refer time: 29:13)

Now this is the bayes theorem if E 1 to E n be n mutually exclusive and exhaustive event associated random exponent. If A is in any event which occurs with E 1 or E 2 or E 3 then what is the probability of E i given A has occurred. What is the probability that it belongs to class A I given that A has occurred given that these attributes are there given this is my observation. So, this is nothing but this is the denominator is nothing but the total probability total probability of A.

Denominator is the total probability of A, what is the numerator? Numerator is the probability of that class into probability of A given probability of E i. What is the probability of A given that E i has occurred into probability of E i divided by the total probability will give me what is the probability of E i given A. So, one more thing here I forgotten now this thing is the this thing is called a posterior probability.

What is the probability of A given that E i has occurred? Because I want to find out probability of E i given A. Now this probability of A given as I this is the posterior probability and this is called a prior probability.
(refer time: 30:59)

See here the table shows that event A has two outcomes so event Y has two outcomes A and B basically we can say that it is these are the different classes. Namely A and B which is dependent on another event X with various outcome x 1, x 2, x 3 maybe this may be the X is one attribute it has just one attribute set and this is the class level. Suppose we do not have any information of an event A then for a given sample space we can calculate probability of Y = A.

How many occurrence of A by total occurrence the five version this is the probability of Y = A this is called prior probability this is called probability of Y is equal to this is called prior probability. Here in this case this is called prior probability. So, now suppose we want to calculate what is the probability of X given Y = A this we will be able to calculate is not it what is the probability of X given Y = A given Y is what is the probability of X = x 2 from here we saw that from for Y = A x 2 how many is there this is here this is another one.

So, it is 2 / 5 total Y occurrence is 5 so this 2 / 5. So, this is called posterior probability we will be using this posterior probability and prior probability concept while we will be solving Bayesian classifier.

(refer time: 32:33)

In the next lecture we will continue our discussion on page and classify. So, in this lecture we learned a scope and application of agent classification the properties and its applicability we have also done a quick recapitulation of Bayes theorem and the next lecture we will go to the details of Bayesian classifications.
(refer time: 32:50)

Thank you, guys.

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

Lecture - 51
Bayesian Classification (Part-II)

Hi guys, so in continuation of our earlier lecture on Bayesian classification so we will be continuing from what way we have ended last lecture.
(refer time: 00:36)

So in today's lecture we will see the name based classification method we will see that within help of an example and base classification it has some problems as well as the solution also exists for that. So we will be discussing those problems as well as the solutions.
(refer time: 00:53)

So now so first is base classification classifier what is it that it is the what to say based on the base theorem we have seen how is it right. So now just a formal if I can write suppose Y is a class variable and X is equals to this x is the set of attributes, so if Y is a class variable means y 1, y 2, y 3 whatever now if there are four classes so y 1, y 2, y 3, y 4 the classes. So I am using variable Y for the classes and variable X for the set of attributes where X consists of;

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

Lecture - 52
k-Nearest Neighbor Classification

Hi everyone, so today we are going to start a different classification strategy. Now, that we have already seen base classifier that was based on statistical matter. Now we will be seeing distance-based classification that is a k nearest neighbour classification.
(refer time: 00:41)

So, in this lecture we will cover the concept of k nearest neighbour classifications, how does this work? How does this kNN works and why we need to use kNN algorithm? How to choose the optimal value of k? What are the advantage? What are the disadvantage of the kNN and does kNN and have the curse of dimensionality problem this will be discussing these things in this lecture. Now the k nearest neighbour classified to start with.
(refer time: 01:05)

So, first of all, as I have already mentioned that we are discussing only supervised classifier under this in this course. So, we have already seen under supervised classifier there are different classification techniques, one is the statistical based standardized distance-based error based and distantary base. So, kNN is also a; we call it a k nearest neighbour classification technique.

It is a type of supervised learning classification and it is very widely used to solve many classification problem. The kNN is a data classification method for estimating the likelihood that a data point will become a member of one group to another, what does that mean? See here estimating the likelihood that a data point will become a member of one group or another.

Meaning, we try to find out the similarity index as I have discussed in my last to last lecture if I remember correctly. So, when we suppose there are two different classes of data. And we want to given a set of attributes we have to find out each bill this setup attribute means this data basically this observation belong to which class, class A or class B. What we try to this? We try to find out how does set of attributes this data is similar to which set of which class?

Basically, it is similar to class A or it is similar to class B. We try to find out it is similarity the index, less similarity. One way of finding a similarity is index is by finding out the distance Euclidean distance of these attribute values of this my test data point with the different classes the object belonging to the different classes. So, with the distance where the distance is smaller, so definitely my test data will belong to that class.

So, that is what we estimate the likelihood that a data point will become a member of one group or another it will try to estimate the likelihood. How we will try to estimate the likelihood? By finding out the Euclidean distance of this between these two object between these two observations. And based on this Euclidean distance we will tell this data point this stage data belongs to which group? Based on what group of data point is nearest to it?

We will classify the data based on the group which to which this data point is nearest and it is a distance-based classification technique that we have already mentioned.
(refer time: 03:46)

So, when I tell it is distance, how do I find out the Euclidian distance between two points suppose my attribute sets are the there are two attributes in a tuple in one observation suppose there are two attributes suppose this is. So, suppose a point that is x 1 y 1 and another point is x 2 y 2, if I want to find out the distance between these two points is nothing this all of us you know this how to find out Euclidean distance of two points.

So, if there is a two it is if the data is two dimensional. That means, if it has two attributes this is how we find out the distance that is $(x 2 - x 1)^2 + (y 2 - y 1)^2$ root of this whole thing. So, this now suppose if my observations or if my object has three attributes if it is a set of three attributes then we will have to consider three dimensions, so this is how we find out the distance.

We find that is $(x 2 - x 1)^2 + (y 2 - y 1)^2 + (z 2 - z 1)^2$ square root of the whole thing. So, similarly if there is a two dimensional, we will consider it as a straight line if there is a three dimensional will consider it a 3 dimensional plane and if it is more than three dimensional then it will consider it as a hyperspace. So, Euclidean in distance in n dimensional hyperspace if the n dimensional data points are this x 1 x 2 up to n y 1.

And the distance between them is same thing whether it is 2 dimensional, 3 dimensional, 4 dimensional, 5 dimensional just the same thing just you have to find out the distance.
(refer time: 05:23)

Now, how do you do it? Suppose, for a given set of data there are two classes suppose we given observation are there is just two classes. Suppose, here I have seen this is some this one class is this blue square which we see this is one class and another class this red colour triangle which is suppose this is another class suppose we have two classes. Using this training data, the training data or this observation data this data which is already we could observe.

You can I am using the term training if I will talk in terms of machine learning. I will use the term observe data if I will talk in terms of statistics. So, let me tell it is an observed data, so using this observed data we want to know the class of an unknown point. Now, my unknown point here is this green circle. So, greens I want to find out what is the class of this green circle how do I find out.

So, we have plotted this suppose the data are plotted suppose since it is each attribute has it is observation has two attributes. So, we can very well point plot it in a 2 dimensional way. So, now this if I plotted the data this suppose that plot takes this form. Now, what is the first step calculate, the distance of all training point from the test point what I will do is that. I will calculate the distance of all these training points how many training points are there?

Here, we have 3 4 5 6 4 8 9 10 11 total 11 trading points are there. So, I will calculate the distance of this green point from all this training point what is the distance of this and days distance of descent days distance of descent is this everything distance this this all the distance. I will calculate the distance of all the training points that is my then that is my first step. What is my second step? Find out k training points which are nearest to the test point.

Basically, once I have calculated the distance then I will arrange it in the ascending order. I have calculated distance then I will arrange it in the ascending order and then I will find out k training points which are nearest to the test point. I will come to what is k, let us k is some number then, I will find out k training point which are nearest to the test point, nearest to the test point means distance will be less.
(refer time: 07:52)

The k in kNN is a parameter that determines the number of nearest neighbours to include in the voting process. So, to find out as I told you here find out the k training points which are nearest to the test point. So, I will find out this k training point, now this k determines which k training points which are nearest to the data test data point. This k will determine how many of this number training data will be involved in the voting process.

Basically, this is in k nearest neighbour we just it is a sort of voting process. How many of the point is suppose there are? We will see the example here let is not talk now the test sample should be classified either to blue squares or to the red triangle yes, it has just two classes blue square and red triangle. So, this test is either has to be either should belong to here or to here if k = 3 we have found out the distance we have sorted in an order.

And now suppose I have considered k = 3. Suppose, if I consider k = 3 then what will k = 3 that is the nearest point the 3 that means the three nearest point to the green circle. So, this k = 3 means this portion and if for if k = 3 it is assigned to red triangle. If k = 3 what happens? There are how many direct triangles that is nearer to the green point nearer to that green point there are two red triangles and however nearer to the green point is only 1 blue square.

So, that means what when we if we take the voting process red will win because two will both for green and the only 1 blue will vote for green, so to green red will win. So, now that means the green circle belongs to the class which are these red triangles. Now, if k = 5 if I take k 5, 5 nearest neighbour. The 5 nearest neighbour means this dotted circle you see 5 nearest the neighbours in this case how many points are how many red points are closer to green.

There are 2 red points which are closer to green and how many blue points are closer to green,

there are 3 blue points that are closer to green. So, in voting who will win the blue points will win. So, that means the green point green circle belongs to the blue square cases. So, if k = 5, it is assigned to the blue squares, 3 square versus 2 triangle inside the outer circle. It does not consider the distance here if you see the distance red is very near the rates are very near it is not that.

We find out the nearest first nearest k and then out of this k how many the class which has more representation the test data will belong to that class. So, when we have considered k = 3, which class had more representation rate class at more representation that is it has 2 and blue class at 1 and when we considered k = 5 our blue had 3 representation and red has 2 representation. So, our test data will go to blue.
(refer time: 11:15)

So, if I now that is the step basically so, now if I write down the steps so first is load the data choose the k value. How to choose the k value? I will come. Now, for a test data point find the Euclidean distance of all training data samples. We have done that, store the distance on an ordered list and sort them in descending order sorry it is a missionary it is not descending order it should be ascending order.

We will store them in ascending order, so and then what happens; choose the top k entries from the sorted list. We will choose the top k address from this the sorted list top k means I should not say top k but the nearest the value which is less value. So, level the test point based on the majority of the classes present in the selected point. So, first what we will do we will find a equilateral distance of all the training points then we will sort the data.

How will you sort the distance? We will sort the distance and the ascending order and then we will consider the top k points, top k points means, the points which are the k points which are nearest to the; which has the smallest value basically and then among this k point which has the major representation our test data will go to that class.
(refer time: 12:47)

Now, this kNN is also called a lazy learner, why it is called a lazy learner? Because you see in all of the classifier or regression analysis whatever we have done regression and all those are also prediction classify also prediction all this is prediction only. Everything in all these cases you see we have developed a model, for all this we have once we get a training data then we develop the model.

So, based on that model given an input we try to find out what is the output? If it is a regression analysis given x what is the value of y that is what we try to find out if it is a classifier given x what is the class of this x that we try to find out. Whatever it is either it is regression analysis or classification whatever it is, so all clustering. So, now there we form a model first, but in case of this kNN we somehow do not form the model.

So, we whatever test data is given we just store the test data. So, from the statistical

perspective when the test data is there it is just there that is that we do not do any sort of calculation that is data is there. This data belongs to this class, this data belongs to this class. So, this data are there, so that is why this process is called lazy learning algorithm it does nothing when it does that when it gets the observation when it gets the initial observation is does nothing.

It only starts walking when it has to test a data when given an observation we have to find out the corresponding class then only it will start it is action before that it will not do any action. So, that means it does not learn any mathematical equation or model, it does not form any model, instead it just stores the training data. and starts computation when the test data is presented. So, that is why it is called lazy learning algorithm,
(refer time: 14:41)

Now, kNN is also a non-parametric method because, it does not make any assumption about the underlying data distribution. It is considered a non-parametric method because why not when we use consider parametric method, we have certain assumption about the underlying data this data has to belong to these distributions. Now, here in kNN is very much a non-parametric you do not have to be you do not have to bother about the distribution of the data.

And that is the input data that is the attributes of this different tuple. Simply, speaking kNN tries to determine what group a data point belongs to by looking at the data points around it. Given it the observation we have to find out it belongs to which class this class or this class I will just have to observe this training the different classes of data of the test data for test data whichever we find as test data is nearer to whichever class, we that our test data belongs to that class.
(refer time: 15:47)

So, now the condition equation is how do we find out the k value, I have already specified that we have sort the data then we take the first k value first k points to find out it is nearest neighbour for scale nearest neighbour. Now, this question is how to find out this k value there is not any specific way to determine the base k value. So, there are different empirically you can do find out take different k and find out the values.

So, on one way how is you can take it you might have to expand with a few values before deciding which one to go forward with. Means, you will try and deal with many values many k values and then you find okay this k value gives the better result, then we can go with this k. Now, how do we find out how will I experiment that how many k values will take and then which case giving good result how do we know.

Because, in the observations as does the supervised learning technique and the observation the class level is given. Now, and for the different observation for which we have to find out the class that is our test data basically it does not have the class level. So, if we are taking a whatever k value, we are taking 40 test data we are getting the correct value or we are not getting the correct value, we would not be able to say.

Because the states values data and class categories is not known to us. So, only way is that if we can find out the base k value if we have some chess data for which our class data is known to us. We have certain such data for which the class data is known to us and then we fix a particular k and then we try to for this data, we try to see it belongs to which category and whether it is correct because already we know the class level.

Now our exam this result is giving the result what we expected if the results matches then, this is good if the result does not matches then something is problem then that maybe we will have to change the k that is how we do it. So, that is what I have written here so, one way to do is by considering or pretending that the part of the draining sample is unknown. So, the observation what is given to us.

Or let us say if it is a training data or whatever it is doing that portion of it, we will consider as if it is not known to us, means we have not stored those data. We will take it a portion of this training data we will not store to a store we will just use a portion of the observation to make different classes. Now, this other portion which observation for which the class level is known to us, that we will try to put it into different categories taking different values of k.

Using this subset of training data states that are the true class levels are known in this state this data here the true class levels are known. Perform classification using kNN algorithm and for different values of k. Check the classification accuracies for different k will because, this class levels are known to us. This suppose we know this data x 1 x 2 x 3 it should fall in class A which is known to us.

Then again x 4 x 5 x 6 it should fall in class B it is known to us likewise we have many data we know which it is falling which class it is known to us. Now, we try to find out the in this x 1 x 2 x 3, we assume a particular k value and we see what which class we get. If we get class A then it is good whatever is whatever we expected we got that. So, likewise suppose we got we took 15 such suppose, we took 15 such data for which class level is known to us.

We took a particular k and accordingly we predicted the class then, we out of this 15, how many of us gave the correct result. We will see that, then again, we increase the value of k and we see out of this 15, how many of that gave the correct result? That way we will do different experimentation on k and will find now to choose that value of k which gives the highest accuracy this is one technique of taking the determining the best value of k.
(refer time: 20:08)

So, when dealing with a 2 class problem it is better to choose an odd value of k when there are 2 classes always it is advisable to choose an odd value for k. The value of k must not be multiple of the number of classes present these are some of the constraints. Choose the optimal value of k is by calculating root and where and denotes the number of samples in the training data set. One way of selecting k already have mentioned.

These are some other techniques you can one way is that definitely does the best method

which I have just mentioned. And another if you do not want to do so many because training data maybe you have around 220 training data which we want to use and try and find out the value of k. So, if many times you have to do it fine take one values of k, then okay you try it second value is, so that is definitely the best matter.

Otherwise, if you do not if you directly go for taking a value of k. So, these are the different steps at the; choose an odd value if it is 2 class A should not be a multiple of the number of classes. And k it is advisable to select k as a root of N where N denotes the number of samples in a training data set k with lower value such as k is 1 or 2 can be noisy and subject to the effect of outliers.

If we take a very less k value it can be noise and subject to the effect of outliers there are some variables like yesterday, I mean my last class I have given you an example suppose I have one class of cat another class of dog. So, cat also has but to say 4 legs 1 tail dog also have 4 legs 1 tail. And some dogs again there is some dogs whose height is almost same as the cat but usually dog's height is more than a cat.

So, this is not an out layer but there are some dogs with height is less but I am just to give you an example. So, likewise there will be some sometimes some attributes which basically we get some value of the attributes because of some out layers which is usually not normal because of some abnormal situation or maybe because of some very rare situation we get those value. And if we select k value very small suppose, 1 or 2 we tend to we may select a round classes because of this outlier values.

We will find a very good match with the first data because of this out layer attributes maybe actually that is not the class of that. So, it is always advisable to take a higher value of k. On the other hand, came with larger values in most cases will give rise to smoother decent boundary but it should not be too large as well if you take k value too large the whole sample size generally the whole training data sets.

Then, what will happen lots of calculation which will take a long because; calculating Euclidean distance for each and every point it will be very computationally intensive. So, you should not be very large also but then a large value of k is advisable it gives a very smoother boundary decision boundary. Otherwise, groups with fewer number of data points will always be out voted by other groups.
(refer time: 23:28)

So, some pros and cons of kNN pros it is easy to understand and simple to implement we have seen how easy to understand it just find out the training data sort it find out the k nearest number, which is the how many representation the class which has the highest representation test belongs to that class. It is so easy implementation. It is ideal for non-linear data since there is no assumption of down drilling data, data can be any format.

Since, there is no underlying assumption these classifications that work very well for non-linear

data as well. It can naturally handle multi-class cases; it can perform well with enough representative data. Then, what is cons associate com associated computation cost is high as it stores all the training data it has to store all the training data it has to find out a Euclidean distance of all the training data sources computation cost is very high.

Requires high memory storage need to determine the optimal value of k. Prediction is slow if the value of N is very high and it is sensitive to irrelevant features. Sometimes some classes have some irrelevant features, some very uncommon features and it becomes what a bad thing is that it becomes very sensitive to those irrelevant features also. We consider those irrelevant pieces also important.
(refer time: 24:53)

And the term curse of dimensionality means, if the data is of high dimension. So, if the data is of high demands high dimensional data high dimensional data means many attributes for it a tuple has many attributes suppose a tuple has 100 attributes then what happens? We have to find out the distance you can imagine the time it will take to find out the distance of these 100 attributes.

So, this kNN and suppose from the curse of dimensionality. With the increasing number of features time complexity of the distance computation largely increases. Obvious, does not it we have to find out a distance of such high dimensionality our disc competitor it is very computationally expensive. kNN does not work well if there are too many features. Hence, diamond is dimensional reduction technique like principal component analysis, feature selection must be performed during the pre-processing.

So, if you are interested in using kNN classification algorithms if the number of features are used. Usually, that is the case usually it is same, there are a number of features a huge then what will happen we will have to reduce the features there are different methods of reducing differences. We will have to reduce the features during data pre-processing, principal component analysis which I will not be discussing in this lecture.

So, way of written feature reduction is that one way is principal component analysis basically from the different suppose there are 100 features. Out of 100 features there is some data some features which are redundant. Again, there are; some dependency in some feature set. So, suppose let me tell you suppose there are feature set suppose say X Y Z A B suppose these are the different features and for one observation we have total 5 features for one observation.

In this feature suppose if we have seen if X increases Z also increases or if X increases Z decreases. So, there is a dependency between these 2 features, when there is a dependency between these two features. Then, why we will consider these 2 features, that means this feature will have an effect on the class that we understood. This feature means to find out the class to find out the class which class it belongs to we will need to consider this attribute.

We have seen we will need to consider this attribute but if we consider only X that is sufficient because, X and Z are very much correlated either positively correlated or negatively correlated.

So, we can just ignore one though this we can find out given the features that we can find out which of the attributes are correlated. If the features it attributes are correlated, we can ignore taking all the attributes that is one way of reducing the features.

So, that is how basically if there are too many features again there are some features which have which does not have any effect on the class level. So, why unnecessarily drag that features unnecessarily find the distance these features. So, there are different ways how do we find out how do we reduce the different features different attributes it. So, once we reduce the attribution then kNN becomes quite good. Then, it does not it is main problem is the cost of dimensionality.
(refer time: 28:16)

So, now coming to the end of this kNN classification we learned about a distance based classifier which we call it a k nearest neighbour method. This is a very simplest classification technique in fact from all the classification this is the simplest one. In fact, for base theorem also I will not say it is simple as we will have to calculate different probabilities. So, and it is widely used in many applications.

So, we learn the kNN classification method the salient properties of kNN it is advantage and it is disadvantage. So, now is the time for tutorial in the next class we will have a tutorial before going to other classification strategies.
(refer time: 28:55)

So, these are the references, thank you guys.

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

Lecture - 53
Tutorial on Classification Techniques

(refer time: 00:33)

Hello guys, so today is the tutorial class, tutorial class means interesting. Anyway, so now coming to our tradition of doing a tutorial class first we solve some objective equations then we solve some problems so today also we will be following the same trend.
(refer time: 00:47)

So, first the objective type questions see here. Which of the following statement is true or data which of the following statements are true? In supervised learning all input training data are level with some class names as it is true or false and supervised learning all input training data are level with some class name that is definitely true. Then what is what about the next statement in supervised learning all input training data are not necessarily level with class names.

That is definitely false. Supervised training means in the training data definitely the observation along with the observation falls in which class it is specified. Clustering is a type of supervised learning technique. Is it true? No, clustering is a type of unsupervised learning technique an unsupervised learning technique where technical class levels are not there. So, we try to group the objects based on the similar characteristics so that is that we call it clustering.

So, in unsupervised learning all input training data are level with some class name that is definitely false. So, our true is the only statement a.
(refer time: 01:57)

So, next, you have a data set of different flowers containing their petal lengths and colour. Different flowers there are sets of different flowers it contains their petal lengths and their colour. A model has to predict the type of flower for a given petal length and colour which type of model it would be? So, we have a data set that contains the petal length and a colour and this it corresponds to which flower that is specified.

Now we have to develop a model which will predict the type of flower given a petal length as well as the colour. So, which type of model is it, is it a Ch-square test definitely not Ch-square test we use for what we use we use Ch-square trace for correlation analysis and this is not a correlation. Then a classification model, is it a classification model, definitely it is a classification model we will have to predict to which class it falls.

Clustering model definitely not because the class levels are already specified into different flowers containing it is specified for different petal lengths and colour it belongs to which class it is already in a training that it is specified. So, it is not a clustering model regression analysis, definitely not, I think you can remember what is regression analysis what we do in regression analysis given x what will be the value of y that is what we find in regression analysis. So, it is a classification model.
(refer time: 03:22)

Next consider a table given below the table shows that even y has two outcomes namely A and B which is dependent on another event x with various outcome like x 1, x 2 and x 3. So, even y has two outcomes simply A and B and this is dependent on another even X that means Y we can say it is a class variable sort of it is just there are two events A and B we can say consider it is a different A is a class B is a class.

And it depends on the variable x and x has three variables outcome x 1 x 2 x 3. So, depending on the values of x 1 x 2 x 3 will be able to predict whether it falls in A or it falls in B. Now it is asking what is the question what is the probability of P X = x 1 given Y = B, given Y = B what is the probability that X will take a value x 1? So, this is an example of but there is an example of yesterday we have done posterior probability.

Remember we have done what is prior probability posterior probability if I would have asked what is the probability of Y. Then it is it would have been a prior probability probably Y = A it is a prior probability of Y = B that is a prior probability. Now what is the probability of X = x 1 given Y = B that is a positive probability. So, how do we find this out so what is the probability of X = x 1 given Y = B, how many occurrence of B are here?

There are 5 occurrence of B are here 1 2 3 4 5 and for given that Y = B what is the probability that X = x 1. X = x 1 how much we have just one just one so it is 1 / 5 that is 1 / 5 = 0.2 is the answer. We have done this type of problem and last two last lectures when we have tried to calculate the prostate probability and prior probability in fact for Bayesian analysis also, we have used a similar kind of example.
(refer time: 05:36)

Now for two events A and B the base theorem states that so, this is the formula which is a formula for base theorem. What is the probability of a given B how do we find out probability of a given B probability of A given B? So, what is this? This is probability (B | A) x probability of A divided by the total problem total probability B is not it. So, which is here given this is the correct answer.
(refer time: 06:14)

Naive Bayesian classifier cannot be applied to training data with, what it is given cannot be applied to training data with categorical attribute. Naive Bayesian classifier; remember naive Bayesian classifier cannot be applied to training data with categorical attribute. Which one is

correct? It cannot be applied to name base we cannot apply to which sort of things we can apply to category attribute definitely as we can apply to category.

Numerical attribute with continuous range of values even that also we can apply. There we have seen there are two approach out so if it is a continuous range of values and it is a numerical attribute is a continuous range of values how we can solve it. One we will either discretize it or we will keep use as a probability distribution function try to find out the mu n sigma and then we can find out the posterior probability.

So, this is also you can do it and null occurrence of some attribute values for some classes if there is a null occurrence of some attribute value still, we can use it M estimate. Remember the M estimate approach if there are null occurrence that we could solve it by using the m estimate approach M estimate approach remember the formula. We will see some examples as well. So, this is also a mixed set of attributes even that is, so which the answer is none of the above. So, all these things it can be applied to Bayesian classifier, so answer is none of the above.
(refer time: 07:49)

Time complexity of the name base classified for n features k class data. What is the time complexity for name-based classifier? I will not go how to find out the time complexity because that is a part of algorithm course and I think all of you know that also. So, if you are interested in finding out the time complexity of name-based classifier which is n features and k class data. So, n features this side we have features decide we have class we have seen the diagram. So, how many total what will be the complexity it is nothing but order of n cross k.
(refer time: 08:28)

So, which of the following is true about kNN algorithm? kNN algorithm does more computation on testing time rather than training time, is it true? Definitely not true kNN is called a lazy algorithm why because it does not work in training time at all it just stores the data during training time and during testing time only it does what work it has to be done. Again, an algorithm does more computation on training time rather than testing time this is correct.

kNN performs much better if all of the data are of different scales, is it true or false? See kNN performs much better if all of the data are of different scale see if all of the data are of different scale what happens. In kNN we try to find out the difference between the; we try to find out a distance among the attributes. Suppose in the two points it has two attributes say one we have a attribute say a particular attribute say age and second attribute say we have a salary.

Two attributes some sort of classification we are doing and two attributes age and salary and suppose a is range from say 30 to 40. Age could take the value from 30 to 40 and salary it can take the value from say 70000 maybe 70000 to maybe 10 lakhs. So, salary from 70000 to 1 lakh you can imagine the brains so when we try to find out. So, this each attribute will have will have this value each observation each observation will have the value say age and salary.

So, and then when we try to find out the find the difference between the given any testing data,

we try to find out the difference of this testing data with all of the training data. This salary because of its range of data it will overpower the age what to say the contribution the age attribute will make. So, that is that way what happened only salary will be the defining factor which will this deciding factor which will decide it which is the which value is nearest.

Though age is also one of the important equally important factor but because of its very lower range compared to the salary it becomes overpowered by the other attributes. So, that is why kNN and it does it works very badly when the data are of different scale. So, kNN performs much better it is definitely wrong kNN does not performs better kNN performs very bad when the data are of the different scale data should be usually another uniform skill.

And not necessarily in the same scale but it will be it should not be there should not be too much a variance it too much of variance then one overpowers the other and the overprint powering attribute becomes the defining factor deciding factor to decide object or observation will belong to which class. When kNN derives a mathematical relationship between the feature set and a class level during training, definitely not during training it does not do anything.

So, only a is correct it does more computation on testing time rather than training time, in training time it does not do anything, so only a is correct.
(refer time: 11:52)

Which of the following is true about kNN algorithm? It is a parametric and eager learner algorithm. See here I have in a note I have given what is eager learning and what is lazy learning, lazy learning I have already mentioned but still you can see. Lazy learning is a learning method in which generalization of the training data is delayed until a query is made to the system. Lazy learning means it does not do any work during the training time.

It only starts working when the train testing data has come so that is we call it lazy learning. That is what kNN does then and there is another kind of learning which you call eager learning and in the case of eager learning the system tries to generalize the training data before receiving queries. Before receiving any queries only, it tries to form the model. So, that once it receives the query just by utilizing the model it can find out the class.

So, that is we call it eager learning so kNN is definitely not a eager learning kNN is very much a lazy learning. So, which one is correct here it is a parametric and eager learner algorithm the kNN is not a parametric which I have already mentioned here the data as we are not bothered about data belongs to each distribution on and it is definitely not a eager learning algorithm also. It appearance it is a parametric and lazy learner algorithm it is a lazy learner but it is not parametric. So, answer is d, it is a non-parametric and lazy learner algorithm.
(refer time: 13:24)

Consider the following figure. The black triangles are the data points belonging to class 1 and the grey rectangles are the data points these are the grey rectangles data points belonging to class two, the test point is marked as green this is the test point green pentagon. What is the

class level of the test point as per kNN and classified which k = 1 and k = 3 if k = 1 then what will be the class level of this place point?

If k is only one which one is nearer it is only nearer to this green rectangle. So, then its if k = 1 its class level is class 2. So, the answer will be between these two definitely this and this will not be the answer. Now we will see what happens when k = 3, when k = 3 which are the nearest point, these are the nearest point then it belongs to if when k = 3 more number of class 1 points are nearer to the test point. So, that means when k = 3 that this test point belongs to class 1. So, my answer will be this.
(refer time: 14:45)

Which of the following statement is true about kNN classification algorithm? See if you see the objective questions in fact in all of my tutorial if you go through the objective question then if you forget anything you can just go back to the lecture and everything will be clear actually. I am trying to touch each and every topic what I am teaching in the lecture classes I am touching each and everything in my tutorial classes. It is sort of a quick recapitulation actually.

This which is not discussed only. Which of the following statement is true about kNN classification algorithm? Which is true? kNN plus it is easy to implant implement and robust to noisy and outlier this already we have seen kNN I already yesterday I mean last lecture we have already mentioned that it is very easy to implement implementation is very easy. You just have to find out the distance of the train test data the distance from the test data to all the training data.

Then sort the data in ascending order take the top k data and in the top k how many of them the in the top key data the class which has a major representation the test data will belong to that class. It is that easy implementation these are these are the three four steps. But then it has one disadvantage it has that if there is some noisy data if there is an outlier that is creates gets affected when we try to find out a neighbour.

Some of the funds some insignificant feature some very really occurrence features which actually sometimes this is different becomes the defining factor. So, it gets affected by this sort of outlier. So, that is one of the disadvantage of the kNN algorithm. So, that way a is definitely the right answer. Next let us see the other choices it is easy to implement and affect it largely due to noisy data and outlier.

There is some mistake here it is easy to implement an affect it largely due to noisy data and outlier. It affects but it does not it is robust to noisy data it is not robust so this is not a correct answer I have given here wrong this is not the correct answer. It is not robust to noisy data and outlier it affects it is but it at the same time it is not largely affected also it gets affected. So, I will be the answer will be it is easy to implement.

And it gets affected due to noisy data and outlier it is not largely. It is tough to implement and robust to noisy data and outlier definitely this is wrong. It is tough to implement it is not tough to

implement and affected largely due to noisy this is wrong. So, this three are wrong so it is easy to implement and affected largely this largely is wrong actually but this is the right answer. And a slide I will be changing it so it is easy to implement and its affected due to noisy data and outlier.

But it is not largely affected. If we take a very high k value for low value at low k value of course it is largely affected but if you take a very high k value it is not affected largely, I will be sending changing this in the slide.
(refer time: 18:07)

Then we will be solving some problem and the given data set provides the information of climate condition and rainfall of total 14 days. So, a test data of today is shown below, so what data it is given it is given the climatic condition based on a climatic condition whether there is a rainfall or not rainfall that is given data for 14 such days are given. And then we are given a test data of the climatic condition of today.

Then based on this climatic condition we have to predict whether there will be rainfall or there will not be rainfall. Basically, whether this data falls in the category rainfall or it falls in a category no rainfall that is what we need to find out. Using Bayesian classifier comment about if it will rain today or not, we will be using Bayesian classifier. Bayesians classifier already we have seen it is when the things are not deterministic and certainly, we cannot say with certainty because of this it will happen.

So, because in certain climatic condition we cannot say for certainty that climb it will rain or it will not rain. So, this is sort of in under this sort of situation when this sort of examples comes even here it is mentioned using Bayesian classifier. Even it is not mentioned you can directly go for a Bayesian classifier. When you see there is a when you can solve it using probabilistic value. The type of example where probabilistic value will give us a better solution than product immediately you can you should go for Bayesian classifier.

So, here it is already mentioned Bayesian and this example also shows for Bayesian classifier. Now so this is given, so Bayesian classifier we have already done some example in our class so I think you will be able to do it basically what we have to find out we will have to find out the posterior probability we will have to find out the prior probability. And then definitely Bayesian means will be used base theorem we know the formula for the base theorem.

And then when we know this two term posterior probability and prior probability we will be able to find out the probability of rainfall or probability of no rainfall. We will find out the probability of rainfall, we will find out for this particular data we will find out probability of what is the probability of rainfall, what is the probability of no rainfall. And the probability which is bigger we will consider that as the class for the given data.
(refer time: 20:26)

So, here these are the data given. So, first we will find out the prior probability how will you find out the probability, I think you can remember. Prior probability is because probably we only find

for the classes is not it there are how many classes there are just two classes. So, yesterday remember we found out we were talking about the air traffic data in our last lecture and we have given an example of air traffic data.

Air traffic data based on that there were many classes late very late on time cancel there were four classes. So, we found out the prior probability of those four classes. Always we remember we founded the prior probability of the classes. So, here how many classes two classes, no rainfall, yes rainfall. So, we will find out the prior probability of that so, what is the prior probability of yes first how many data's are here tier total there are 14 data's.

Out on 14 data's how many occurrences of yes if you count you will see there are nine occurrences of yes. So, if there are nine occurrences of yes so prior probability will be 9 / 14, prior probability will be 9 / 14. Similarly, prior probability of rain is equals to no and say out of 49 is yes then definitely 5 will be no so prior probability of rain is equals to noise 5 / 14 so we got the prior probabilities both the class.

Now once we got the prior probability of both the class then next what you have to find out is. Next, we will find out the posterior probability. What is the positive quality how will I find out? So, there are how many attributes climate, humidity, outlook, windy. So, we took the different attributes together. So, first let us consider for climate. When the climate attribute is hot what is the probability that essentially what we are trying to find out what is the probability that attribute climate attribute is hot.

Given that my classes rain what is the probability that my prior probability is my at climate attribute is hot. So, for rain how many hot a day you can just find out. So, here this is hot this line and it is rain and then this is another one only two we have. And total how many range occurrences are there? Total nine already we have seen. So, what will be the posterior probability 2 / 9, similarly we will find out for mild.

Mild how many occurrences before you rain, for this is rain and for hot no rain so no rain hot no rain this is one, this is one there is no more. So, it is how many occurrences of no rain was there it is 5 so it is 2 / 9. Similarly, for mind mild if you see how many occurrences given that it is class is rain how many what is the probability that attribute climate has mild climate. So, mild and yes this is mild and yes this is one and this is one, this is one, this is one, this is one, this is one, total how many 1 2 3 4 we got 4 so 4 / 9.

Similarly, you find out the posterior probability of all these attributes for both the classes rain is equals to yes, rain is equals to no, so we found out the posterior probably for both the classes then we found that the prior probability.
(refer time: 24:03)

Now we will have to find out what is the probability of this class this condition climate is equals to hot, humidity is equals to normal, outlook is equals to sunny, wind is equals to false. So, climate hot humidity in normal so climate hot means basically this one humidity normal then

what was that outlook, sunny and windy falls out looks, sunny, windy is false basically deserted probably 2 / 9 x 6 / 9 x 3 / 9 x 6 / 9 x 9 / 14.

That is for probability that it will rain probably that it will not rain 2 / 5 x 1 / 5 x 3 / 2 / 5 x 2 / 5 x 5 / 14 that is what.
(refer time: 24:55)

So, the class probably have yes given today's attribute we got this no today's attribute we got this, now which is bigger definitely this is bigger. So, this today's attribute we can say that it falls in the class rain means with these attributes it will rain probably it will rain maximum probability that it will rain. That is how we do Bayesian analysis Bayesian classification.
(refer time: 25:22)

So, as per Bayesian classification rain for today is higher compared to the class probability of no rain for today.
(refer time: 25:32)

So, on there is another question for page why I have kept this here if we will see how we have used M estimation technique. We have learned a technique remember M estimation when we have used M estimation when some of the posterior probabilities or some of the class occurrence itself is zero. Some of in our testing data in our observation if some of the class occurrence is zero if the class occurrence is zero definitely the corresponding posture probably will be zero.

And sometimes class occurrences are there but the positive probability is zero. Then how do we calculate that because the whole multiplication of that will give us zero one class totally becomes unrepresented. So, that then we use M estimation approach so this is the problem given where you can use M estimation approach. So, we will quickly go through this because it is the same type of example.

A food delivery application determines the overall quality of a particular delivery depending upon three user ratings delivery time, food quality and delivery person's behaviour. Based on this we give the from food overall quality. These three parameters are rated by as a user after each delivery on a scale of 1 to 4. Depending upon this user ratings level quality of the delivery is class ABC.

Depending upon this rating the overall quality we fall in it falls in three classes that is A B and C, depending on these three variables. The data from 15 such deliveries are given in the table, now suppose for a new delivery the received user ratings are. For new delivery disorder user ratings, now what is the overall quality whether it is A B or C can you predict overall quality of this delivery using naive Bayesian. So, similarly name based on how we have done I will not explain it since the same thing.
(refer time: 27:18)

So, you see using the naive Bayesian what we got here so, when we try to find out probably that class is equals to A for this attributes we got this 0.017 probability that class is equals to B for given this we got value 0 because this one occurrence zero occurrence you see when the for Z = 2 class = B it is nowhere it is given here it is B here z 4 here B it is z 4 here also B z is 3 B z 1 for B there is no occurrence of z 2.

If B has occurred there is no z = 2 that such occurrence is not there at all so zero occurrence what are zero occurrence I got here postal probability probably that behaviour of z = 2 given that B is that I got 0 making that so my class probability has become 0.
(refer time: 28:22)

Similarly, if I try for C also class C also, I will get 0 because of this posterior some of the posterior probability. In other words, there is not enough training data in favour of this cases Bayesian that is one limitation Bayesian we need lots of training data, training data should be used and there is one that is a limitation. The there is not enough training data in favour of this case. This is an issue of name-based classification that can happen due to small size of training data.
(refer time: 28:53)

Now such issue how we can solve it we can solve it using the M estimation. Yesterday we have already discussed what is M estimation so this is the formula for the M estimation, this is the formula for the M estimation. Usually without if we do without M estimation what is our formula P of A is equals to particular variable given a particular class, we just write see basically this is the n c i / n.

We usually do this what is n c i? n c i is the number of training examples from class here that take the value A j = x. And n is the total number of instance from class C i when we consider calculate positive probability we just use this function this expression and C i / n but in under such cases there we may get some zero values as well. So, we have modified it and we have used two different terms n c i + mp / m yesterday we have discussed.

What is m? m is a equivalence m is called the equivalent sample size and p is a user specified parameter, p is a parameter user specified parameter that totally depends on the applications that totally depends upon the number of total sample size class representation there are different various factor based on what we can give the p. If it is a very sensitive, we take a very small p if it is not that sensitive you can take a very bigger p, p is definitely but value will be between 0 to 1.

Now what is the equivalent sample size? Already I have mentioned equivalent sample sizes so, here how many classes we have it is given total we have 15 sample size and we have 3 classes. So, if it would have been a stable data 15 sample size 3 classes if you if it would have been a stable data maybe we would have got 5 samples for each class is not it. If it would have been a sample data that may not be true but when we try to find out the equivalent sample size that way we will talk.

So, here since there are 15 sample size and 3 classes for a for stable data it would have been 5 samples for each class. So, I will consider here my equivalent sample size as 5 so my m = 5.
(refer time: 31:09)

So, in this example p is called the price estimate of the probability in the absence of other information and uniform prior is used. So, the uniform this is one of the heuristic and uniform I can also take as p = 1 / k where is where k is the number of values that attributes can take. So, this is one of the heuristic how we can set k value, k is the number of values that the attributes can take this attributes can take how many values.

So, suppose X can takes 4 values, Y can take suppose 6 values, Z can take suppose say 3 values. Then we will consider a maximum maybe so we can take the k value as 6. So, that way this is one of the heuristic different methods to take k value so k value here we have taken 4 so my p =1 / 4 and M as I already told is a constant equivalent sample size we consider it a 5. This is also arbitrary it is not necessary that you can consider it this way. You can consider equivalent sample size you can that is also very much user dependent.
(refer time: 32:22)

So, now using this value m = 5 p = 1 / 4 we calculated again and we found for class is equals to A we found. So, for class is equals to A we got this is the probability I am not showing the calculation same calculation or we got this probability.
(refer time: 32:40)

For class is equals to B we got this probability.
(refer time: 32:44)

For class is equals to C we got this probability and which one is bigger class A so the given test that is classified as Class A as per naive Bayesian classifier. So, that is how we use M estimation approach.
(refer time: 32:58)

So, now next question the speed and agility ratings for 20 college students and whether they are selected in the college football team is given in the table. Based on the; speed and agility there whether they are selected in the college football team. So, if the speed is high as it is high maybe they are selected or maybe not so this data's are given. Using the kNN technique you can predict that a student with speed 6.75 and agility 3.0 will be selected in the team or not.

So, we will have to find out using the kNN technique. So, kNN technique means our this is our case data, speed = 6.75 as well agility = 3.00 this is our training data. So, that means it has two attributes and this is our test data 6.75 and 3.00 this is our test data does two attributes and this and the table what we have given it is this is our training data it has two attributes. So, in kNN what we try to do? We try to find out the distance of each training data from the test data.

So, we will find the distance of each training data from the test data then we will sort the data in ascending order.
(refer time: 34:09)

So, that is all this is how we found out the distance. Simple Euclidean distance $(x_1 - x_2)^2 + (y_1 - y_2)^2$ root that is a two attributes. So, we found out the Euclidean distance this is the value of the equilibrium distance then we have sorted it. Now we will have to take the first k top k. Now what is k here now as the number of neighbours that is value of k to be considered is not given it a problem, we should check the classification output for different values of k.

As it is values for k is not given it is just ask for kNN algorithm. So, we can check for different values of k as it is in my last lecture, I have already mentioned values how to find out the values of k different heuristic I have given. And the best part is that to find out the values of k considering a portion of the training data as if that they are not training data their test data with a class level is not known.

Assuming that assuming a portion of the training data the class level is not known and then we take different k values and we try to find out which k value gives the better choice better classification for this data. Because this data we know the classes we know the classes and we take different k values and we find out what class this data's gives. So, based on the accuracy the for k value which we get highest accuracy.

We can consider that as a k value for the given problem given application that is one heuristic and there are other heuristic as well which I have already mentioned in my last lecture.
(refer time: 35:40)

So, here just for an example maybe different values we have seen different values if I consider different values of k if I can consider value of k as 1, so this is in the sorted order. If I consider value of k as 1 then which one is the; this is which only one is only one means the it is only one neighbour so first neighbour is what first number is that it is in the yes. That means the first neighbour a first number is yes only one number that means my output will also be yes.

If I consider 3, 3 means this 3 out of 3, 2 no 1 yes then my output level will be no 5 then 1 2 3 4 5 out of this 5 how many 3 no 2 yes, so it will be no again likewise.
(refer time: 36:32)

So, that is all it is shown in the problem.
(refer time: 36:35)

Is a same this is I will just see what this Manhattan distance is so considered in following data sets the x and y coordinates of the one of the training points and the corresponding class levels are provided. The test point is given as 4.5 predict the class level of the test point using 3 NN classifier here it is given k = 3 means it is kNN and k instead of k it is given 3, hence k = 3. Using 3 NN classify considering Manhattan distance as the proximity measure.

So, what we have seen in the for kNN we have considered Euclidean distance as a proximity measure this there can be different heuristic to consider the proximately measure, one heuristic to consider proximity measure is Manhattan distance.
(refer time: 37:20)

Manhattan distance is that is how we calculate this is what Manhattan distance. Absolute value of x - x t that is the training data + y - y t x and y are all the training data all the training data and x t and y t is the test data. When so we; find out the absolute values of the difference of the attributes for all the training data. Other previously Euclidean distance what we do we take the square(x 1 - x 2)^2 + (y 1 - y 2)^2 + y 3 - y x 1 - x t test data (x 1 - x 3) ^{2} + (y 1 - y t)^2 plus.

Whatever how many attributes are there accordingly and we do square root, here in case of Manhattan this is the formula we do not do square root which and we do not do square also just absolute value this is how. Based on this stress procedure is same nothing to explain here.
(refer time: 38:14)

So, it is given will be able to see it.
(refer time: 38:16)

So, that is all for today, thank you guys.

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

Lecture - 54
Support Vector Machine (Part - I)

Hello guys, today we will start a very interesting and important classifier and it is something like if you this classifier I am talking about support vector machine. If you do not know support vector machines means like you do not know anything it is sort of this support vector machine is like it has become sort of, I can say it has become sort of status symbol. You do not know status back as SVM so that is it so SVM is such an important classifier.

So, and but the problem with this is just SVM you will hardly find any good material in the net even in a book also it is not written many of the textbook it is not written very properly. So, I some try to get information's from all this and try to put it in as simple words as possible because it is a bit difficult to understand. And initially it is very easy but as I proceed well it is a bit difficult to understand.

So, please pay attention and that is all it is a gun let us learn this. Now let us start this support vector machine.
(refer time: 01:35)

So, first we will in this class first we will introduce this support vector machine and then in support vector machine there is a concept called maximum margin hyperplane. So, what is this maximum margin hyperplane that is also we will see here.
(refer time: 01:51)

So, this support vector machine it is fondly called by people as SVM that is not mostly see people call it as SVM. So, it has that has received considerable attention this classification as it is considerable attentions and this technique as one of the co-inventor of this technique is Vapnik, Vapnik is not the only invented is one more person which at present I cannot remember his name and but the Vapnik why he is important he is the person who has used this technique in some in statistical learning theory.

As the first person to have used this technique in the statistical learning theory and made this popular. So, he is one of the co-inventor of this. So, in SVM what it does is that it is as a tax of classification it searches for the optimal hyperplane now also called decision boundary. So, when I talk of hyperplane what is this called also called the distance boundary, separating the tuples of one class from another. Here you can see the figure.

So, suppose this is one class of data this is one class of data these are two different classes of

data. So, to separate how it searches for a hyperplane this is the hyperplane, it searches for some sort of boundary line or something certain which differentiate between these two classes. Suppose a wall it is a wall the data behind and wall is behind this wall is belongs to some other category data in this side of wall belongs to some other category.

So, that is a hyper plane basically and hyperplane has this is a hyperplane and it has a boundary, each hyper level has a boundary. Now we will see what is this boundary. This hyperplane is also called decision boundary means that is give making us that hyperplane is helping us in taking the distance given and it is data whether the data falls in which category, data falls in this and it falls in this category or it falls in this category.
(refer time: 04:02)

So, SVM what are the advantage of SVM, why people when we talk of classification techniques mainly SVM people think of SVM, why it has many advantages. First is SVM works well with higher dimensional data and thus about its dimensionality problem. With higher dimensional data it can work very well like kNN and we have seen kNN and it is a curse of dimensionality it is very difficult to work with higher dimensional problem when if we use kNN algorithm.

And again, the if for different data scales we cannot use kNN algorithm. So, if the different data skills of the attributes. Likewise, this SVM does not have all those disadvantages SVM, it works well with higher dimensional data and thus it does not have the dimensionality problem cards of dimensionality does not have. Although the SVM is based on classification that is the training time is extremely high.

SVM the training time is extremely high, the result is however highly accurate, further testing and unknown data is very fast. Here we can say SVM is the eager learner remember what is in kNN is a lazy learner it does not think during training. So, SVM does the whole work during the training time only it does the whole work during the training time, the whole workman is basically finding out the hyper plane which differentiates between different classes.

So, and then it takes a long time for that of course but then it is very accurate and once that is done for and testing time is very fast, SVM is less prone to overfitting than other methods. So, what is overheating? I will see it in the next slide. It also facilitates compact model for classifications. See here you see and here we have four classes. So, these are the different hyperplane, this is one hyperplane, this is one hyperplane, this is one hyperplane, how it differences between the classes.
(refer time: 06:07)

So, this is the over under fitting, over fitting, optimum fitting what is that? You see here and this the first line, suppose these are the data points this blue colour dots are the data points. The blue colour dots are the data points and suppose I found a function my function what I found is just a straight line. Can you see the straight line? The straight line is my function and classification in all this classification and all what we preservation, (06:39) whatever it is we just try to find that function, is not it?

So, here I found that the function is my straight line. But if this is the straight line and what happens many data speed does not fall into straight line or this figure is not very accurate. Anyway the some lines some points maybe it falls into straight line some points it falls in a straight line there are but many data does not falls in a straight line. In such case we call as under fitting. When in such cases what happens?

For such under fitting curve then what happens our training error is also quite high and our testing error will definitely high. When our training errors are high definitely our testing error will also be quite high in case of under fitting. So, this is an very good example of an under fitting. So, actually this; my true function would have been this but I got this as my function. So, this is there are lots of training error.

Training error means when I the data which I use for training and there is data also when the function what is given even that function did this data is not satisfying the function itself, the training data also. This training data means whose class level is known. When I am using this function so when I am using this function to classify it into two classes suppose I have two classes classified into two classes from my training data only I am you know suppose for some data I need to get it in B class but I am getting it in A class.

For some data I need to get it in A class but I am getting in B class so that is called training error. So, in under fitting case there is lots of training error when there is training error definitely there will be testing error also. So, in case boat and training and testing errors are high in case of under fitting. Now you see over fitting, over fitting means you try to you have found out the hyperplane in such a way that it caters to all the points.

Model means we try to based on the observation based on different trial and error. We are trying to find out the best model is not it A base function that fits the data. So, if we draw if we find such an model which fitted all the points almost all definitely not all the points if it is almost all the points then what happens? In this case our training error will be very low and negligible when we get such a curve.

Training error will be very negligible but when we find this sort of curve because there may be some data's which are outliers. There may be some data which are very rare occurrences we are trying to what to say or trying to address those data also. When we do that then what happens? Our model is not the true it model does not mean the true model. In that case our testing error increases and see in over fitting our training error is very less because we have tried to fit almost all the points almost all the training points.

So, but then training error is very less but our testing error is high. This is the case of overfitting and this is the example of optimum fitting. Optimum fit is that where my training error is also low and testing error is also low that is we call it optimum fit. So, here but SVM is less prone to overfitting than other methods. As we missed definitely it definitely does not do under fitting but it is less prone to over fitting as well.

When I am talking of overfitting let me also tell you one thing. This kNN algorithm what we have seen again and if we take a very less value of k then chances are there that we do overfitting. For less value of k where the success of overfitting is there overfitting means training error is very less but testing error increases.
(refer time: 10:35)

So, now support vector machines the types of SVM linear SVM and non-linear SVM. Now we will be discussing linear SVM definitely will go to non-linear SVM also later. So, linear SVM a classification technique when training data is a linearly separable. This is an example of linear SVM by a straight line by a plane by a hyper plane we can separate the different datas. And this is a very good example of linearly non separable.

Here also we have two classes, one is this red circle another is the magenta blue squares and yellow curves yellow circles. But these are not linearly separable so we call that as non-linear SVM. So, now let us see what is maximum margin hyperplane.
(refer time: 11:27)

See we have talked about the distance boundary. When a data are described in terms of two attributes so two attributes means in distance ax + by the in when we describe a data in terms of two attributes then our distance boundary may will be a straight line. What will be the equation of the straight line? That is ax + by + c = 0 where c is the intercept a and b are the slope. So, ax + by + c = 0 and it has two attributes.

So, this is where distance decision boundaries may be a straight line. If by a straight line I can differentiate between the two classes. Now when there is data's are in three dimensional x 1, x 2 and x 3 or x y z when data is under three dimensional then we will use a plane to differentiate between the two categories. Then my differential boundary will be a plane entry dimensional similarly.

If my data's are in 4d, 5d, 6d so it is very difficult to just what to say picture that it is it then decision boundary will be hyper plane. So, for two attributes it will be a simple straight line, for three attribute it will be a plane, for more than three attribute will be a hyper plane. So, in general I will tell it is a hyper plane only hyperplane which is differentiating between the classes.
(refer time: 13:00)

So, now formulation of the classification problem. So, formally let us formulate it. So, in our subsequent differential assume a simplistic situation that given a training data with a set of n tuples, which belongs to two classes either plus or minus and each tuple is described by two attributes say A 1 and A 2. So, what is given to us some given a training data this is my training data t 1 to t n these are a different observation, t 1 is one observation, t 2 is another observation.

That means total I have an observation our n tuples which belongs to two classes this

observation. Suppose let us first simply stay for simple of discussion I am just considering that it belongs to two classes. The classes name I am giving is plus and minus and each tuple is described by two attributes A 1 and A 2. So, each has two attributes so two attributes we will be using a straight line to find out the two different classes to differentiate between the two classes.

So, for this current example we are considering that data is linearly separable so the current model is linear SVM model.
(refer time: 14:17)

So, now so here you see the figure, the figure three shows a plot of data in 2D. So, it is that we have considered the data has two attributes plus and each data has two attribute and there are two classes, classes are positive and negative. So, suppose these are the classes. Now these data are linearly separable so what happens? So, if this data is linearly separable so how linear is separable means we can separate it using a straight line.

So, now see if these are the data set actually then we can have different type of straight line to different to mark as a boundary decision boundary between these two point. This also can be one of my line, this line can also be one of the distance boundary, this can also be one of the distance boundaries, this can also be one of the distance boundary this can also be any number I cannot find indefinite in finite lines which can act as a decision boundary between these two categories.
(refer time: 15:20)

We can there are an infinite number of separating lines that can be drawn. So, distance boundary there can be infinite number of distance boundaries in the input and here are my distant boundaries and straight line. So, a straight line I can draw it just because to show that the; what is a set of data. Suppose this is my straight line, the set of data falls in different category and this set of data falls in different classes; that is all.

So, now since this is the case in this figure what you see then this figure there can be infinite lines. Now there are two questions whether all hyper planes are equivalent so, for the classification of data is concerned so there are infinite hyperplanes. Let me call it instead of line we are using a general term because data may be 2D, 3D, 4D, 5D anything, is not it? So, we are using a general term hyper plane.

So, if it is an considered and using this term hyperplane so now all this hyper planes are equivalent that means I can consider any of the hyperplane to the but to say to classify it into two classes if not which hyperplane is the best. If all the hyper planes are equivalent if not then which one is best which one should we take that is the two questions that arises.
(refer time: 16:38)

So, we may note that so for the classification error is concerned with training data all of them are with zero error. See if you take any of this line in this figure if you consider this line also this line this straight line if you consider this straight line also training error if I consider this error all

my pluses in this side all my minus in this, this side of this there is no training error. There is no minus here no plus this side so there is no training error.

So, if I consider this line also there is still no training error. If I consider this line also still not any of the line if I consider there is no training error. So, that means one from the training perspective any airline is equivalent. But let me see what happens when I talk of the test data. Suppose this is my test data. However, there is no guarantee that all hyper planes perform equally well on unseen data.

All these hyperplanes perform good in case of training data, I have seen here from this figure but all this hyperplane does it performs well in terms of the test data that we need to find out. Our requirement is since the training data also should be training error also should be low, testing error also should be low. Note that only training errors should be low testing error can be high or training error can be high testing or can below no both we need both low optimum.

We say I have seen optimum free it is that training error is also low testing error is also low. So, for this hyperplane training error is slow for all the cases. Now what about the testing error? You understand by meaning by what I mean by error, error is that training error means a data, a observation which is labelled as plus. When I use the function when I use this prediction model then I have given set of attributes are there and its class level is also given.

I know this class level is plus but I did not specify the class level. What I have done? I have just given the attributes and I have used this model to predict to which class it belong. When I have use this model to predict the switch classes belong for my data suppose I got class minus then that is a training error because I know this should be plus but it has gone to minus, this is in training error. So, similarly testing error is what?

Testing error is also same thing, testing error how about the class levels are not given there. So, based on the model you will have to find out the class to which class it belongs. Thus, for a good classifier it must choose one of the infinite number of hyperplane so that it performs better not only on training data but as well as on test data. So, we have to choose that hyperplane which performs well both on training data as well as test data.
(refer time: 19:42)

To illustrate how the different size of hyperplane influence the classification error consider any two hyperplane. In this figure you see to we will consider any two hyper plane, one there were many hyperplane in the last figure we have seen. Now we will just consider two hyperplane just one this is hyperplane one and another this is hyperplane two. Just two we have considered just two hyperplane to just to illustrate which hyper plane is better.

This to hyperplane H 1 and H 2 have their own boundaries say this hyperplane have their own boundaries. Now how do we find the boundaries? Now this H 1 which is the boundaries of H 1? H 1 boundary is this one b 11 and b 12. This dotted line you can see this dotted red dotted line; this is the boundary of H 1. Similarly, what is the boundary of H 2? Boundary of H 2 is this

dashed line black colour dashed line b 21 and b 22. This is the boundary of H 2.
(refer time: 20:48)

Decision boundary, how do you find out? It is a decision boundary is a boundary which is parallel to hyperplane and touches the closest class in one side of the hyperplane. So, I told for the hyperplane one H 1 my decision boundary is b 11 and b 12. For the hyperplane two my distance boundary is b 21 and b 22. Now how do I find the decision boundary? So, here it is. A decision boundary is a boundary which is parallel to the hyper plane and touches the closest class in one side of the hyper plane.

It is parallel to the hyperplane. So, H 1, this is H 1 so these are the two parallel lines and it touches the closest class. So, my decision boundary suppose this is my hyperplane. So, I will go on increasing my distance boundary till it touches one of the classes. Suppose here I got this I am growing in equal thing both the side. Suppose this side is starches class one plus that is class A. When it touches the class the distance between the hyperplane suppose in the middle it is the hyperplane distance between the hyperplane.

And my decision boundary suppose it is x then he decides also other side also it will be x it is the same distance. So, a descent boundary is a boundary which is parallel to hyperplane and touches the closest class in one side of the hyper plane. The distance between the two distant decision boundaries of a hyper plane is called a margin, this distance between two. So, you can see this m 1, m 1 is the margin of the hyperplane H 1 similarly m 2 is the margin of the hyperplane H 2, see in the figure.

So, if the data is classified using hyperplane H 1 then it is with larger margin than using hyperplane H 2. If I classify data with hyperplane H 1 and I have a larger margin and however if I classify with H 2, I have a lower margin m 2. The margin of the hyperplane implies the error in classifier. In other words, the larger the margin lower is the classification error because I have a great margin here.

So, senses of becoming an error is very less, this is my decision boundary, this is my hyperplane my class B starting from this point, you see here. This is my distance boundary my class B is starting from this point, class A starting from this point. So, I have a lot of gap in between so senses of getting an error given a test data senses of getting an error is very less. On the other hand, suppose this is my hyperplane.

My B started from here my A started from here my chances of becoming an error for a very small variation my B will become A, A will become B, senses of a classification error is very high. So, bigger the margin lesser is the classification error. So, that means they have a different hyper planes we have seen their infinite hyperplane I will look for that I will go for that hyperplane which has the highest margin. So, that my classification error is less.
(refer time: 24:16)

Intuitively the classifier that contains hyper planes with a small margin is more susceptible to

model overfitting and tend to classify with weak confidence on unseen data. Thus, during the training or learning phase the approach should be to search for the hyperplane with maximum margin that is and we will try to select that hyperplane which has the maximum margin. Then what happens? It is maximum margin it is very less susceptible to error.

Margin is lesser more susceptible to error. Our confidence in classification reduces if the margin is less. So, such a hyperplane with maximum margin, such a hyperplane is called maximum margin hyperplane. So, that is the concept of maximum margin hyperplane. Support vector machine means you have to you need to know what is maximum margin hyperplane. We may note that the shortest distance from a hyperplane to one of its decision boundary is equal to the shortest distance from the hyperplane to the decision boundary at its other side.

I told you it is equal both the sides. Alternatively, hyperplane is at the middle of its decision boundaries this is obvious I told you. This whatever is the distance from the hyperplane to the decision boundary whatever is the distance this distance is suppose this distance is X similarly hyperplane to decide also this distance will be x, hyperplane is in the middle of the decision boundaries. So, now linear SVM model. We will not be completing linear as we will model in today's class we will just start.
(refer time: 25:56)

So, SVM which is used to classify data which are linearly separable is called linear SVM data which we are linearly separable. Linearly separable means we can separate the data by using an hyperplane. But if it is too attribute by using a straight line, three attributes by using plane, more than three attribute by using a hyperplane then we call it a linearly separable. In other words, a linear SVM searches for a hyperplane with maximum margin obviously.

SVM means will always try to look for a hyperplane with a maximum margin. This is why linear SVM if often terms as maximal margin classifier. Linear SVM is also called maximum maximal margin classifier because we are looking for the maximum margin hyperplane. So, this is also called maximal margin classifier.
(refer time: 26:53)

So, now finding the maximum margin hyperplane MMH maximum margin hyperplane for linear SVM for a given trading data how do we do that. So, basically formal notation for putting it. Consider a binary classification problem consists of n training data now we are considering binary classification only for case of discussion. Consists of n training data notation of the training the how do we note it each tuple is denoted by X i Y i where x i's are different attribute corresponding to the attribute set of the ith tuple.

It is considering data is in a m dimensional and space that is why we have taken x i 1 x i 2 x i m there are total what to say how many observations we have taken. Suppose we are taken t observations where i = 1 to t and each observation has total m attributes. So, it is an m distance base and Y i is the class so Y i belongs to the; suppose there are two class plus and minus. It denotes its class level.

The choice of which class should be level as plus or minus is arbitrary which one is class plus which one is two classes minus is arbitrary. Given X i Y i for i = 1 to n we are to obtain a hyper plane which separates all the two sides of it which separates all into two sides of it. So, our goal is to find the hyperplane with maximum margin such that it separates all this data all this we have total n data, all this n data into both the sides of the hyper plane.

Some of the data will fall into one side, some of the data will fall into the other side of the hyper plane and this hyperplane should be with maximum margin that is our goal.
(refer time: 28:47)

So, before going to the general equation of a plane in n dimensional let us consider first a hyperplane in 2D plane. In 2D plane what will be a hyper plane? In 2D plane our hyper plane will be nothing just the equation of a straight line ax + by + c = 0 I am writing as w 0 + w 1 x 1 + w 2 x 2 where my attributes are x 1 and x 2, w 0 is the intercept where it is w 0 w 1, w 2 are nothing but the coefficients defining the slope and the intercept of the line.

So, my that means I have to find this line, my task in SVM is to finding this line what is this line w 0 + w 1 x 1 + w 2 x 2 = 0, I have to find this line. Finding this line means essentially, I have to find out the values of w 0, w 1, w 2. This line such that it has the maximum margin.
(refer time: 29:49)

So, this is my; the equation of the line. Now any point lying above such a hyper plane satisfies any point which is lying above this hyper plane point lying in this hyperplane is equals to zero. Any point lying above this hyperplane will be greater than zero any point which is lying below this hyper plane will be less than zero.
(refer time: 30:18)

So, an SVM hyperplane is an n dimensional generation of a straight line in 2D. It can be visualized as a plane surface in 3D but it is not easy to visualize when the dimensionality is greater than three of course. In fact, Euclidean equation of a hyperplane, we can write it in this way where i's are real numbers and b is a real constant called the intercept. Instead of w 0 I have written b you can write it in this way also which can be positive or negative.
(refer time: 30:44)

So, why I bought to this form basically. This particular equation I can represent it in matrix form. How can given a system of equation? How can we solve it is in matrix form? Same nothing so this is my equations w 1 x 1 + w 2 x 2 = b. So, if I write it in matrix form this will be my matrix from W X + b = 0. If you are getting confused you can also write W X + w 0 = 0. So, this is my matrix form where W is a vector having always value w 1 to w m different coefficients.

And axis again different attribute values and b is a real constant. So, here W and b are the parameters of the classifier model to be learned given a training data. Now does W and b we have to find out the parameter value of this W and b given as setup data given a set of

observations.
(refer time: 31:41)

So, consider 2D training data consists of two classes here the plus and minus. So, this is the line W X + b = 0, any point which is above the line is W X + b = 1. Since we have considered as plus one and minus one see here. Suppose b 1 and b 2 are two distant boundaries above and below the hyper plane respectively. So, it should be basically any points above this is W X + b greater than zero below this is W X + b less than zero.

So, consider any point x plus and x minus shown, suppose I have considered this point x plus and x minus here. Suppose this is my decision boundary, this dotted line. This dotted line is decision boundary suppose this dotted line the equation of this line is suppose W X + b = 1. And this distance boundary suppose W X + b = - 1 because this side it will be less than zero. Since W X + b = 0 and below this it will be less than zero. So, W X + b is to - 1 top side is W X + b = + 1, I am just assuming it is.
(refer time: 32:52)

So, now for x plus for any point x plus located above the decision boundary the equation can be written as W X + b = K where K is greater than zero. Any point which is above the hyperplane it will be greater than zero, is not it? So, W X + b I am writing a square where K is greater than zero. Any point which is below this point that is X bar I can write is W dot X bar + b it is K dash where K dash is less than zero.
(refer time: 33:21)

So, now basically thus if we label all plus as class label plus and all minus is class label minus then we can predict the class label y for any test data. How we can predict? Y is equal to plus if W X + b is greater than zero W will be minus if W X + b is less than zero. So, given any value of X what will I do? I will try to find out what is the value of W X + b. If the W X p+ b if I get greater than 0 then it will fall in plus class. If W X + b i find it is less than 0 then it will fall in the minus class.

So, that is how we will do the classification. Now but that will happen once I know what is W what is b once I know what is W, what is b that is the equation for my hyper plane with the maximum margin. Once I know the value of W and b then I can easily classify my data whether it falls in plus class or it falls in minus class. If it if W X + b is greater than 0 it falls in above the highway plane if W X + b is less than 0 it falls below the hyper plane.

Now how to find the margin of W X + b? How to find the margin of W X + b how to find out the value of W and b that will be discussing in the next lecture.
(refer time: 34:40)

So, in this lecture what we learned? We learned the basic concept of SVM and its advantage linear and non-linear SVM, the concept of maximum merge in hyperplane then the margin of maximum margin happened for linearly separable data in 2D. In the next lecture we will learn

more about MMH of a linear SVM.
(refer time: 35:04)

Thank you, guys.

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

Lecture - 55
Support Vector Machine (Part - Ⅱ)

(refer time: 00:34)

Hi guys, in continuation of our lecture on SVM so today is the second lecture on that. So, and we have in our last lecture we have seen that for a linear SVM our objective is to find out the what to say margin basically for the maximum margin. So, as like what we can separate the data into the both sides of it, is not it. So, nowadays minute of a margin our objective is to find out the maximum margin this is also called maximum margin hyperplane.

And then until now we were just discussing about for binary classification on (01:07) classification of two classes. So, in this context now in this class will be covering how to compute the margin we have seen that maximum margin what is maximum margin hyperplane that we have seen. Now in this lecture we will see how to compute the maximum module hyperplane MMH.

And then this computing this maximum margin hyperplane disk can basically be formulated as an optimization problem. So, we will see how to formulate this as an optimization problem and this to solve this optimization problem we use Lagrangian multiplier method which will be discussing in this lecture. So, now first is the computing margin of maximum margin hyperplane. (refer time: 01:47)

So, in the last lecture we have already learned that a equation of the hyperplane is in the matrix form that is W X + b = 0 this already we have seen is not it. So, what is W is? W is equals to different w 1 w 2 depend on how many parameters access how many features how many parameters. So, accordingly that we have the different coefficient w 1 to w n if access m features then X = x 1 to x n and b is a real constant.

We have seen that is not it and we have also seen that we have level plus as one plus and mind this is the plus class this is the minus class and we can predict the class level Y for any test data given any test data X we can predict a class level Y this is the way Y is equals to plus X if W X + b is greater than 0 what happens. If this is the hyperplane anything which falls in this range above the hyperplane it is above the hyper plane.

It will be W X + b it will be greater than zero and below the hyper plane this side it is W X + b is less than 0. So, we will get a X class class if W X + b is greater than 0 we will get a minus class if W X + b is less than zero, this already we have seen.

(refer time: 03:07)

Now in the equation W X + b = 0, what does W represents, W represents the orientation business is not it and b represents the intercept. Now this any straight line if Y = m x + C what happens if we just scale up or down by dividing a non-zero constant what happens we get the same hyperplane. If we scale it if we multiply it or if we divide it by a non-zero constant, we get the same hyper plane. So, there can be infinite number of solution using various scaling factors.

All of them geometrically representing the same hyperplane, if you just scale up by multiplying it by some numbers or by dividing by some numbers both W and b will essentially get the same hyperplane.
(refer time: 03:54)

So, this there can be infinite number of such solutions, now so this leads to a confusion, so avoid this confusion what we do we can make W and b unique by adding a constraint. So, what we do we want to make this unique so that if we add a constant so this unnecessary confusion will not be there. So, what did you do basically the in other words this is the hyper plane and we draw two parallel lines which we already have seen in our last lecture is not it.

So, and this helps us also to avoid this confusion of infinite number of hyper planes. So, what we do is that I am just writing it here so basically W X + b is equals to if I write k plus minus k plus minus k because k plus is above the hyperplane minus is below the hyper plane, so plus minus K if I divide it by k both sides so this, I can write it in this form more generalized form. So, W X dot X + b dash is equals to plus minus 1.

So, now it may be noted that this W dot W X dot X + b dash plus minus 1 it represents two hyperplanes parallel to each other which you have already seen. So, now this W X b dash is unnecessary it is it will create confusion just for clarity instead of W X this is nothing this the parameters only is not it, so instead of that we can just write W X + b = plus minus 1, just two parallel lines on the both the side of the hyper plane, now we got this.

Now having this understanding now, we are in a position to calculate the margin of the hyperplane. Our objective is to find out the maximum hyperplane maximum the hyperplane which will have a maximum distance. So, now we are in a position to calculate the margin. So, how we will calculate.
(refer time: 05:45)

So, here we have some we will see two different method first method say suppose X 1 and X 2 are two points on the distant boundaries. These are the two distant boundaries say this is b 1 say this is b 2 or whatever it is ordinary person. So, X 1 and X 2 are the two points on a decent boundaries b 1 and b 2 respectively does what happens I can write W X 1 + b is equals to if it is in this boundary.

So, I can write this and if it is in this boundary then I can write W X 2 + b = - 1 or if I take both

together I can I will get this is not it w dot X 1 - X 2 = 2. In fact, if you see here in this figure see this is one vector this X 2 X 1 what are these are vectors is not it this X 2 is one vector this is one vector. So, now then what is this is nothing but X 1 - X 2 vector is not it. So, if I want to find out the magnitude of this vectors what how will I write.

Say d is nothing but this vector X 1 minus here what I have written d is nothing but this vector. So, what is the magnitude of this vector? Basically, why I want to find out this is basically the distance between two vectors is not it this is because this is one vector this is one vector and so this is the distance let me rub it and show it again. So, this is one vector this is another one vector so this will basically give the distance between these two vectors.

So, now distance between these two vectors if I write it d, I will get this I can write it in this way because this is nothing but a dot product W dot X 1 minus it is nothing but a dot product. When we write the since this is a dot product if I take the magnitude of this vectors can I not write it in this way 2 by what is this and what is these equals to this whenever when I want to find out the magnitude of W it is nothing but square root of w m square + w square + up to w m square.

It is in a m dimensional space since we are considering X as m attribute so w will be also in the m dimensional space. So, the magnitude of w will be this so what we got does distance between two vector is 2 by magnitude of w, this is how this is one method of calculating the margin this.
(refer time: 08:13)

Let us see the second method of calculating the margin. Second method of calculating the margin so now see consider the two parallel hyper plane say this and this it is the same hyperplane what we have usually considered. Initially what we have considered let me write it here w 1 x 1 + w 2 x 2 + b = 1 and another one is w 1 x 1 + w 2 x 2 + b = - 1 is not it, it is a plus minus 1. So, if I take b decide what will happen so it will be w 1 x 1 + w 2 x 2 = 1 - b.

And similarly, this I will get a - 1 - b, so 1 - b if I write b 1 and - 1 - b, if I write b 2, if I this I write b 1 and this I write b 2 then I can this two equation I can write it in this format. So, do not get confused why suddenly I got b 1 and b 2 and this is nothing b 1 is 1 – b, b 2 is - 1 - b. So, these are the two hyper planes now these are the two hyper planes because we have already seen. Now if this is so how distance between, we want to find out the maximum distance between these two hyperplane.

So, max distance between these two hyper plane is this distance is not is this d, so how do I find d. So, this d is nothing but this AC this AC. So, if I was if somehow, I can find out the value of AC that AC is nothing but d, so how we will use some simple simplification to find out the value of AC. Now from this two line H 1 and H 2 whichever line whichever hyper plane you consider.

So, what will be the slope, slope is nothing but the - w 1 by w 2 from this line you can find out the slope. So, what is the slope? Slope is nothing but tan theta and what is that this is equal to minus w 1 by w 2.
(refer time: 10:11)

So, now here if you see what is this angle is theta, then what will be this angle, this angle will be 180 - theta. So, now if I find out the; sign of sine of this angle and now this is what this is the hypotenuse this is the perpendicular this is the base is not it. So, in triangle ABC AB is the hypotenuse AC is the perpendicular and the distance between this H 1 and H 2 I this AC is the distance between about to say H 1 and H 2.

So, now if I try to find out the sign of this angle what is this 180 - theta, 180 - theta is nothing but this AC by AB perpendicular hypotenuse sin theta. So, what we got this is one expression we got, so here AC = AB dot sin theta so if somehow if I can find out this value of sin theta and value of AB then I am done I got the value of AC. So, now here how do I find out the value of AB?

AB is easy to find out what is the distance from here this whole distance minus this distance is not it. This is AB distance from 0 to B. If this distance minus this distance will give me AB distance, so this is how I got AB and I know tan theta is also – w 1 by w 2. If I know tan theta, I can easily find out what is sin theta, sin theta I found out this simple calculation. So, I found AB, I found sine theta, so I found AC.

So, this is my AC, AC is equals to magnitude of b 2 - b 1 by root about w 1 square + w 2 square so what is b 2 - b 1 what is my b 2, b 2 is - 1 – b 1 is 1 - b. If I see so this is basically b 2 - b 1 is nothing but this is nothing but actually 2 by is not it this is not nothing but that we got the same thing here also we got. The first method also we got the distance between the two back to hyperplane here also we have got this and using the next method also.

This is what I got in SVM literature this margin is famously written as mu of W b, this margin this is called mu of W b, so this is my margin now.
(refer time: 12:43)

So, now I know this is the margin now what is that actually basically we will have to find out that finding the maximum margin hyperplane so that training phase of SVM involves estimating the parameters W and b for a hyperplane from agreement training data. So, what is the hyperplane our hyper plane is W x + b = 1 is not it so plus minus 1. So, now basically our hyper plane is W x + b = 0 the W x + b + - 1 is the parallel lines which is parallel to the hyper plane in equal distance.

So, now our in what we need to find out essentially in all or all the analysis what we have seen regression analysis and classification and whatever we just need to find out the parameters, here what is the parameters, the W and b is the parameters. So, now we have to choose the parameters in such a way that the following two inequalities are satisfied these are the two inequalities, one is W dot x i + b is greater equals to 1 and W dot x i + b less equals to - 1.

If this is for the plus class and this is for a minus class mentioning we are doing for binary classification so one is below the hyper plane. So, below the hyperplane it should be - 1 above

the hyper plane it should be plus. So, we have to choose the, what to say parameter in such a way so that these inequalities are satisfied. So, this condition impose the requirement that all the training tuples from class Y is equals to plus must be located on or above the hyperplane W x + b = 1.

All the tuples belonging to this class where it will be located on or above it will be located on this line or above. Similarly, any tuples belonging to two this class this class will be located in this hyperplane or below.
(refer time: 14:37)

So, now so these are the two inequalities which are always seen if this two in equal this if we summarize together. So, these two increases we can write it in this form by combining these two together is not it can we not write it yes say this is if W y = 1 for some simplification we are writing is this plus and minus basically plus class and minus class. So, I am writing W i = 1 another one is W i = - 1 there is nothing but the one is above the hyper plane class and below the hyperplane class.

So, if this is the case if I put these two together I can this is my inequality now joining these two together. So, W i this is my inequality now. Note that any tuples that lie on the hyperplane H 1 and H 2 are called as support vectors. That tuples that lie on the hyperplane those are those we call it support vectors this is very important like here we see this is one support vector. This is one support vector which lie on the hyper plane that is called the support vector.

Essentially the support vectors are the most difficult that to classify because why it is most as something which is away from the hyperplane which is much above the hyper plane definitely it will be in a plus class, I know something which is a below much below the hyperplane it will wind a minus class I know. But something which is on the support vector means on the what to say on this plane this is most difficult to classify is not it that is another thing.
So, because the distance is very less the support vectors are the most difficult apples to classify and I give the most important information regarding the classification most important reform information regarding the classification means we will see later. The support vector actually defines the complexity of the SVM and so that is why we call it; it gives the most important information regarding classification.

And it has some in support vectors also gives up will visually see that. So, in the following we just discussed the approach of finding this minimum maximum margin hyperplane and the support vectors, how do we find the support vectors and what are the disks maximum margin hyperplane.
(refer time: 16:53)

So, now we have seen the; what is the hyperplane how we have to estimate we have seen this are the two techniques. One is the first method this is the first method this is our margin is not it, this is our margin. The second method also we got the same thing this is our margin this is nothing but the same thing two by magnitude of W the same thing. So, this is the margin so now

and but at the same time it should satisfy the constant we have also seen that it should satisfy this constraint is not it should satisfy this constant.

So, now what we need? We want that our margin should be this margin what we have seen a margin is nothing by 2 by magnitude of W is not it that is the margin and this is what is them this is the constraint this should satisfy. So, now if cannot we; not put it in as an optimization problem. So, what is the optimization problem? Basically, first we need to maximize the margin this maximum margin hyperplane means our margin should be as high as possible we need to maximize this margin.

So, now we need to maximize this if I write it in the opposite way numerator becomes denominator if I put it in numerator then this is equivalent to minimizing the following. This is maximum this is maximizing then I can write as mu dash W of b is equals to this is if minimizing this then instead of if I write it in this pattern then I need to maximize this value but I am doing it opposite.

So, I am telling I need to minimize this and under the constraint subject to constraint this. So, in nutshell the learning test of SVM can be formulated as the following constant optimization problem what is this minimize this subject to this. So, this is how I formulate this as an optimization problem I need to minimize this, minimize what W by 2 and subject to W i y i into W dot x i + b and it should be greater equals to 1 subject to this.
(refer time: 19:02)

So, now so this is our optimization problem. So, the above constraint the above stated constant observation problem is popularly known as convex optimization problem. This optimization this type of optimization problem is also known as convex optimization problem. Where the object why what is the specialty about this objective function is quadratic in constraint here the objective function is quadratic is not it and the constraints are linear in the parameters W and b.

So, the specialty is the objective function is quadratic and the constraints are linear. So, this type of things we call it a convex optimization problem and for convex optimization problem the well-known technique is the Lagrangian multiplier method is used to find out the solution for such optimization problem. So, now we will learn this Lagrangian multiply method, how once we know this Lagrangian multiplier method then we will use this to find out our objective here is to find out the SVM.

SVM means our objective here is to find out the maximum margin hyperplane. Finding out the maximum margin hyperplane, essentially what we need to find out, essentially, we need to find out W we need to find out b is not it; we need to find out the support vectors those are the things we need to find out. So, for that we have formulated it as an optimization problem, what is the optimization minimize this subject to this.
(refer time: 20:32)

So, Lagrangian multiplier method follows two different space depending two different steps

depending on the type of the constraint. So, does what is Lagrangian multiply method? It has two different technique one is we minimize this function for us also we have a minimization function subject to it has a equality condition subject to g i x = 0 means it has some constraint which is linear constant and this is equals to zero that is my constraint here so, i = 1 to p means there can be p constraint any number of constraints.

So, another is inequality constraint minimize this function subject to let us constrain h i x less than equals to 0 and similarly there may be p number upon strain. So, this is equality constraint optimization problem this is inequality constraint optimization problem. See inequality constraint my constraint is equals to 0 and inequality constraint my constant is less than equals to zero and what in both the case we have to minimize the function.

So, our case is also our case is what we have to minimize a function but our case is subject to h i x if I consider is h i x is greater equal to 0 is not it greater equals to 1. It is an inequality constraint.
(refer time: 21:58)

So, now how we do the how we define the Lagrangian multiply in this way. So, this is does we introduce a new parameter which you call it is a dummy parameter that is lambda so Lagrangian of this how we write it as f X that is the function plus all this constraint summation of lambda i g i summation of all the constraint multiplied by Lambda. For each constant we will use one lambda 1 dummy variable.

If there is one constraint, so lambda one g i if there is two constant lambda 1 g i + lambda to g 2, 3 constant lambda 1 g 1 + lambda 2 g 2 + lambda 3 g 3 likewise. So, it is summation of all the constraint multiplied by the dummy variable and plus f X this is my Lagrangian this is how I write it. Now what I will do is that I will do take the first order derivative of this with respect to each parameter of x each attribute of X first order derivative of this equation with respect to each attribute of X.

So, if x has d attributes so I will be del L by del x i del L by del x 2 del L by del x 3 del L by del x 4 up to up to del L by del x d. So, this is the first thing and since we are trying to find out optimum, so definitely does whatever will do the differentiation first order differentiation will equalize it to zero so if there are d attributes so from this del L by del x i total, we will be getting how many equation total will be getting d equations.

Next again this will also differentiate it with respect to lambda i so if there are p lambdas so we will differentiate all lambda del L with respect to all the lambdas lambda 1 lambda 2 lambda 3. And similarly, we will be getting total p equations. So, here how many equations will get we will get d + p equations. And so, from this d + p equations we will be able to find out the value and equalizing all to 0 and from this will be able to find out the values of this all x 1 to x d and all the lambda i's lambda i lambda 1 lambda 2 up to lambda p.

From this g + p equations will be able to get this value. This is the first one equality constraint

optimization problem how to get the values of the different X how to get a what to say we have to minimize the function. So, what is the value of the minimize function we can know the value of the minimize function for that we will have to know the value different parameters of the X for what parameters of the X we get the minimum value is not it. So, this gives us the different values of the attributes X which will give us the minimum value.
(refer time: 24:48)

So, here is an example also suppose we need to minimize this x + 2 y subject to this constraint x square + y square - 4 = 0. So, Lagrangian how we write so see it was a function of x y and we have used, another one dummy so x y lambda. So, what is this is our function is not it plus lambda into the constraint since there is only one constant so this is just one lambda see is not it, the same thing how we have written f X how we have written f X.

This is how we have written is not it f X plus summation of all the; constant multiplied by lambda I so here there is only one constraint so just one lambda so this is what my L. Now how what will I do I will differentiate this with respect to x, I will just differentiate this with respect to y because there are two attribute x and y. So, first I will differentiate with respect to x, then I will differentiate with respect to y that is one.

Next, I will again differentiate this equation with respect to lambda just raise 1 lambda so 1. So, difference in this with respect to s I got this with respect to y I got this, then with respect to lambda if I differentiate, I got this. So, I got three equations three equation how many unknown three unknown x y and lambda. So, solving this I got this value these are the different x values y values and lambda values.
(refer time: 26:14)

Now what will I do so I got different x value and y value, now my con thing is that I need to find out for which x value and y value there are different x so it is like simple optimization problem you know optimization problem you know genetic algorithm how do we do we take out the different values of x and y so, similarly we got the different values of x and y now which values of x and y actually satisfies our function we need a value which is minimize that x + 2 y should be minimized and at the subject to a particular constraint so from this.

We got different values we for x and y we for x y lambda we got two, two values so from this which value will actually we will give us the minimum value that we have to find out. We will have to solve each and find out so total there are how many choice the total there are 2 to the power 3 choices each value has three each for each variable we got two, two values and there are three parameters so total there are two to the battery choice.

For Lambda is equals to this we take x we take different x and y values definitely lambda will not be used in the equation when we will try to solve because lambda was a dummy parameter which we have just introduced. So, we will take different x and y values we try to substitute it and what we got if we use first this index function we got - 10 by root 5. Similarly, if I take lambda is equals to minus and this x and y this values, I got this lambda is not necessary you

just take the different values of x and y.
Different combination of x and y with a different combination of x and y which gives the minimum at the same time this also said to be satisfied x y + y square - 4 = 0 this should be sanctified and this should be the minimum value. Here we have seen in this two you can calculate all the values and see we have seen has its minimum value at this, this value. You calculate it and see we got it for x = - 2 by root 5 and y = - 4 by root 5 we got the minimum value.

So, now what is the value of x and y x value of x and y is this is the value of x and y. Essentially will try to do will try to solve it you taking different combination of x and y. Take one x another one y again keeping that x change the y, now again the change keep the x constant and change both the x likewise you will be getting all the combinations put it in the function see whichever gives us the lowest minimum value.

But at the same time whether a constraint is satisfied or not it is in the minimum and the constant is satisfied that is my solution that is for equal constraint.
(refer time: 28:52)

Now for inequality constraint inequality constraint is a bit complexity is a bit more than the equality how it is not difficult. So, the method of solving this problem is quite similar to Lagrange multiplier method described previously here it starts with the same with the Lagrangian same the earlier how we have used Lagrangian here also we are using the Lagrangian same thing nothing till now nothing is different.

But in addition, we use some additional constraint which we call KKT constraint. So, this is in Karush- Kuhn Tucker constraint in short it is called KKT constraint. So, for inequality constraint in addition to days this Lagrangian this is we call it Lagrangian when we introduce the dummy variable lambda, we call it Lagrangian and in addition to this Lagrangian we introduce some other constrain.
(refer time: 29:46)

So, what are these constraints say one is first of all the one constraint is that this Lagrangian will have to derive it develop to the first order derivation for all each for each attribute of x i which also we have done for equality constraint. Same wise we will have to do it for each parameters of x i = 0 and for the equally constraint we have also done first order derivation with respect to lambda here we will not do that.

Instead, we have this tree constant what is one is lambda I should be greater or equal to zero that is one constraint my lambda I value should be greater or equal to zero it should not be less than zero my x i value will be less or equal to zero, excited a constraint my constraint value will be x less or equals to 0 and my lambda i x i should be equals to 0. See these are the three constant. Lambda I greater equals to zero please remember when we will solve it.
We will be needing this lambda I beta = 0 h i that is the constraint less equals to 0 and lambda i h i will be equal to zero. Lambda is h i = 0 means either lambda i = 0 or h i = 0. Solving the

above equation, we can find the optimum value of f x now if we solve this, we can find out the optimum value of f x.
(refer time: 31:12)

So, let us see an example suppose we need to minimize this is subject to this constraint. Subjects to this constraint we need to measure here in the inequality constraint our constraint is in less than sign so less than equals to zero. So, accordingly we will just do a bit of manipulation and will bring so this will become x + y - 2 less than equal to 0 is not it and this will become x - y from this we can write x - y less than equals to 0.

If you bring x this side of sine gets changed so x - y less than equals to 0 so these are by two constraints so there are two constraints so I will be using lambda 1 and lambda. So, first is this my function what is my Lagrange? Lagrangian is my function this is my function plus I have two constraint so summation of both the constraint multiplied after multiplication with the dummy parameter so this is one this is the second.

So, what is my KKT constraint, KKT constraint is first del L by del x here there are two parameters x and y so del L by del x whatever I get that s = 0 then del L by del y = 0. So, del L by del x i got this is equals to zero first order derivative simple del L by del y i got this is equals to 0. So, remember what was the what was my KKT constraint one is del L by all the attributes of x they will x i so here there are two attributes x and y that we have done.
(refer time: 32:46)

And what is the second constraint second constraint is my this all by what to say if this was my second constraint, I have written it in the second constraint is all my del lambda 1 lambda 2 is greater equals to 0, is not it. Lambda 1 lambda 2 is greater equals to 0 and what is my the third constraint third constraint is my all my h i h i is less equals to zero is not it whatever my constant h i less equals to 0 so this is the diff I can write it into x less equals to 0 format also.

That means x + y - 2 less equals to 0 then x - y less equals to 0. So, lambda 1 lambda 2 greater than zero dash less equals to 0 then what is the next constant lambda into h i = 0 so this is lambda into h i = 0 lambda into h i = 0 this is one constraint this is one this is one. So, these are my KKT constant. So, this many equations I got. Now with these equations I got all this equation so with this equation I will be able to find out the value of x y and a lambda 1 and lambda 2.

So, now I will have to find out the minimum value of f x y subject to this constraint subject to this KKT constant all my KKT constraints which has been satisfied. So, based on that we will have to find out the value.
(refer time: 34:05)

Now to solve KKT constant we have to check the following test. So, first if I take lambda 1 = 0 lambda 2 = 0 my what is my lambda 1 what is my constraint, my constraint is this is the constant lambda 1 lambda 2 greater equal to zero it cannot be less than zero. So, if I take lambda 1 and lambda 2 = 0 if I take lambda 1 and lambda 2 = 0 in this equation what happens lambda 1 = 0

lambda 2 = 0.

Then what happens, I got this is 0 this is 0 so I get 2 into x - 1 = 0 I get this expression similarly from this expression I get this 2 into y - 3 = 0. From this two equation I get x = 1 y = 3 if I solve it simple equation simple equation you solve it so but however when I get x = 1 y = 3 what is my constraint my constraint is x + y 1 of the constant is x + y is less than equals to - 2. This is one of my constraint another constraint is y is greater equals to x.

The first constraint only plus y less than equals to less than equals to two it is not satisfied with this value x 1 = y x = 1 y = 3 this is not satisfied so this is not a feasible solution. So, if I took lambda 1 is also zero, lambda 2 = 0 the value of x and I what I got that is not a feasible solution now let me take another one combination. So, another one combination let me take lambda 1=0 lambda 2 not equals to 0.

If my lambda 2 is not equals to 0 first look here if lambda 2 is not equals to zero, you see this expression equation if lambda 2 is not equals to 0 then definitely x - y = 0 again is not it. So, if x - y = 0 so from this what will I get x = y I go to another one if lambda 1 = 0 lambda 2 not equals to 0 putting these values here in this two expression. I got this is one equation this is another equation putting value lambda 1 = 0 and lambda 2 is not equals to lambda 1= 0 it is gone.

So, these are the three equation I got from this I found out this value x = 2 y = 2, lambda 2 = - 2 but what is given my lambda 1 lambda 2 should be greater equals to 0. So, here I got lambda 2 = - 2 so that means this is it violates my assumption so this is also not a feasible solution. So, I took lambda 1 0 lambda 2 zero so if I see here lambda 1 lambda 2 different combination I can take this zero both zero this I got this is not a feasible solution.

Then I take this is zero and this is not zero this also I found not a feasible solution then I take this is not zero and this is zero, I will find out the solution another is both not zero both lambda 1 lambda 2 not 0. These are a different solution I can get.
(refer time: 37:13)

So, next I took lambda 1 not zero lambda 2 = 0. Accordingly, I found these are the equations and this value I got this is lambda 1 = 2 this is also feasible because it is given lambda 1 lambda to beta equals to zero and I got x = 0 y = 2 and it satisfies the constraint also, what is my constant my constant is x + y should be less equals to 2 x + y less is equals to 2. So, it is satisfying the constrain also.

So, this is a feasible solution and a final solution what do you find lambda 1 is not equals to zero both lambda 1 and lambda to not equals to zero. This what I got this also if I solve it, I got x = 1 y = 1 but lambda 2 = - 2 my lambda so lambda should be greater equals to that is also one of the constant KKT constant. So, this is also not a feasible solution so just this is the feasible solution. So, what is the feasible solution of this optimization problem x = 0 y = 2.

That is my feasible solution that is how we find it Lagrange in multiplier using the Lagrangian

multiply method we how we find out the values of the values of the f x and y values of the function which will give us the minimum value of the function. So, that we have seen in for some example.
(refer time: 38:29)

Now basically we will be using this method definitely our is inequality constrained and we will be using the second method that is the Lagrangian multiplier for inequality constant. We will be using that method to solve our hyperplane problem that was our problem, for that problem we have used we have formulated as the optimization problem. Now we will be solving this optimization problem using the Lagragian and multiplier which we have just now learned to solve that.

So, now coming to the conclusion in this lecture we learn about the computation of the margin of the hyperplane separating the class in the class of linear SVM. We have seen how to compute the margin we have seen two different two different methods so what is the margin is 2 by w so, how we have computed we have seen then we formulated the optimization problem to obtain the optimal hyperplane with a maximum margin.

To solve the optimization problem Lagrangian multipliers were used we got a basic introduction of how the Lagrangian multiplies are equalized we have seen that. So, in this lecture we will attempt to solve the optimization problem for linear SVM.
(refer time: 39:33)

So, with that I end this class thank you guys.

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

Lecture - 56
SVM (Part - Ⅲ)

(refer time: 00:31)

Hi guys, so in continuation of our lecture on SVM, so this is the third lecture. So, in our last lecture we have seen lagrangian multiplayer method how lagrangian multiply method we can use to find out an optimization problem. So, this will be using now to solve the maximum margin hyper plane. Then finally we will see multi-class classification using linear SVM. Till now we were just discussing about the binary classification.

Where there is two classes, we have taken one is plus one is minus this is just an entry of use. So, now we will also see how we can do multi-class classification using linear SVM having said that as SVM actually it works best when we work with binary classifications.
(refer time: 01:11)

So now linear SVM. So, basically, we need to solve the maximum margin hyper plane to solve the linear SVM. So, what was our now we will we have learned lagrangian method how will solve lagrangian multiply methods. So, what will be done now lagrangian will have to now write the lagrangian. So, our constraint was; remember our constraint >= 1 this is our constraint remember this was our constraint.

Our problem was minimize this subject to this constraint. So, this was our problem we have formulated our problem in the last lecture. Now in the lagrangian inequality constraint we have seen that was not >= 1 but it is <= 0. So, when it is the same just the inequality constant here, we have used <= 0 here since β =1, so what it will be we will just change the sign so I can form this is this will be my lagrangian.

Instead of minus instead of plus here I will use minus because of this constraint. Since my constraint was greater so I have changed the sign to make it less so this is my lagrangian now. So, there is my only but so this is my lagrangian now I will have to find out the KKT constraint.

So, where λ i's are lagrangian multipliers, we have already seen W and b are the model parameters.

Now one more thing to be noted here so, my minimize the function was this was the function here I have used square W^2. So, we use this instead of this for the sake of simplification in conclusion and it does not alter the programs called programs goal adversely. So, even if we use W^2 it does not affect the program call it remains the same our value will not be affected. So, you will see if I use square my simplification becomes easier.

So, I have used instead of W / 2 that is the magnitude of W / 2 I have used the magnitude square by 2.
(refer time: 03:16)

So, now what will mean by my KKT constraint? Remember the KKT constraint the lagrangian first order differentiation of lagrangian with respect to all the attributes. So, whatever one is W, one is b that I have done and then equalize it to zero of course then del L / del W I got this so, from this expression I got W = ∑ λ i y i x i. Then when I do del L / del b this is what I got this expression you can see.

If I do del L / del b what I get this is only from here I will be getting is not it. So, I will be getting this from λ i y i and then what are my other three constraints in KKT constraints there are other three constant that is λ I should be >= 0, this is one constraint. Another is λ into x i = λ into h i that means the constraint so this is λ into this is my h i this is equals to 0.

This is another constraint third constraint is my all my h i's, h i should be <= 0. So, this I can write it into less equals to format, this is the constraint basically at a constraint whatever format you write. So, these are my three KKT constraint along with this derivation. So, now I will have to solve this to find out the value of the W b and x and y. So, solving KKT constraints are computationally intensive.

So, it is like the example what we have seen that was very easy so we could easily solve it. But then for this it is not very easy to solve it directly by hand you will have to use one of the numerical technique methods I am these numerical techniques I have mentioned in many places. Of course, now nowadays like you do not have to know a numerical techniques the softwares are available you just have to give the data and you will get the solution.

But I my suggestion is that of course let the computer do it let we have the ready-made code

and we will run the code, fine I am not doing it but I still I should still know it is not it, you should know the walking of the matter. So, please this numerical technique methods not only for discourse for many things you will be needing in many of your studies you will be needing it. So, please learn in different numerical techniques even you have not learned it till even if you have not learned it till now.

So, using any of the numerical techniques you see which numerical technique will fit and then we will solve the we will get the values of this.
(refer time: 05:44)

So, we first solve the above setup to find all the feasible solution, then like we have seen we will find all the physical solution feasible solution based on the different constraint. And then once we get all the feasible solution then we will among the physical solution which will give us the minimum velocity because there is a minimization problem is not it. So, which will give us the; minimum value that is my optimum value.

Then we can determine the optimum value for this, optimum value for the different parameters. So, lagrangian multiply λ i must be 0 so here we have got this λ i h i = 0 is not it. So, when h i = 0 λ h i = 0 either λ is 0 or h i = 2 when h i = 0 that means my when my constraint is equals to 0 when what will my constraint will is equals to 0 what when that will happen, when any points that lies on the line then it will be equal to 0 is not it.

Any points that lie above the line it will be positive any points that line below the line will be negative, so any point that line on the line it will be 0. So, as I told you the point that lines on the parallel what to say the parallel lines of the hyper plane that we call it the support vectors. So, here that is written lagrangian multiplier λ I must be 0 unless the training is an x i satisfies the equation this.

If it satisfies the equation if it satisfies this equation that means h i = 0 is not it - 1 = theta I can write this y i W dot x i + b - 1 = 0 if this is satisfied then λ is not equals to 0, if this is satisfied then what happens then λ i is 0. So, here it is written lagrangian multiply λ i must be 0 unless the training instance satisfies this and when the training instance satisfies this one when it happens.

When the training instant will satisfy this equation when the training instance are support vectors. When it is support vectors then only it will fall in the line it will fall in the line that means then we will get this is equals to 0 is not it. Does the training tuples with λ i greater than 0 lie on the hyper plane margin enhancer support vectors. When this is 0 λ i's not equals to 0.

So, λ i's λ i value already we have seen λ i greater or equals to 0 either equals to 0 greater than 0 when λ is not equals to 0 means λ when λ is not equals to 0 my h i = 0 when will by my h i will equal 0 under my points lie on the line on the hyper plane is not it. So, points that lie on the hyper plane what is the support vectors. So, with λ i greater than 0 are support vectors.

The training instance that do not lie on the hyper plane margin half λ is equals to 0 that is why I remember when I told you this support vectors give and last like the support vector gives us lots of informations. Support vectors gives us the information when λ i this is with λ i when λ i is greater than 0 then for all those instance λ i is greater than 0 those are support vectors.
(refer time: 09:11)

So, for a given training data using SVM principle we obtain a maximum margin hyper plane in the form of W b and λ i's is not it. We form a what is that all those KKT constraint and all this is nothing but our machine is not it that is nothing but our black box. Basically, from whim that is our function let me tell you that is our function, that is a function given any output it will use that function inside a black book.

And we get the output raise output it files in this class or that class, remember we are discussing binary classification two classes. So, this is called this black box this function we call it SVM support vector machine. Now let us see how this MMH can be used to classify test tuple say X given any test tuple X given any observation X how we will classify whether it falls in plus plus or it falls in plus minus.

So, we will find out this delta of X basically this is the representation we find the delta of X. So, it gives a magnitude and S sign and sine if it is lie above the hyper plane then it is under positive class if it is like below the hyper plane it is a negative class. And magnitude is how far it is from the hyper plane, lesser magnitude is lesser nearer to the hyper plane more the magnitude is far down further from the hyper plane.

So, how we can find? So, λ of X we can write as w dot X + b this is the dot product mind data W dot X + b and what is W we already found out from this lagrangian multiplier this is my W so dot X + b so this is W I have found out from the lagrangian multiplier.
(refer time: 10:54)

So, this is famously known as representer theorem this is known as representer theorem which state does a solution W always is represent as a linear combination of the training data. The solution W is always represent a linear combination of the training data you see linear

combination is not it, y i is the class x i is the input tuple it is the input basically and λ i we already know. So, W is a combination of this is a linear combination of this is called the representer theorem.
(refer time: 11:39)

So, the above involves a dot product of x i dot X and this say when you want to find out the class of X so I use and this is a dot product you see. Earlier there was a dot product of W dot X that we have from brought down to there is a dot product of W i dot X so x what is x i? x i all are all the initial observation that is we can say is the training observation the trading training data. So, we will have to find out the dot product of x i dot x and x i but this is from i = 1 to n, so this is a dot product.

So, in this above involves a dot product of x i dot X where x i is a support vector this is so because see here where the x i is a support vector why, if when λ i is not equals to 0 then all the for all those tuple λ is not equals to 0 those are support vectors in it. If λ i is equals to 0 then this equation will be λ is equal to 0 and this one becomes 0 only is not it. So, when I have X of when this is when I have this that means my λ i is not equals to zero.

For all those λ is not equals to 0, my x i are the support vectors. So, where x i is a support vector we now we find out the sign of λ of X sign as well as the magnitude of the λ of X. If it is positive then x falls above the MMH is and so the SVM predictor class belongs to class level plus if it is if the sign is negative then X falls on or below MMH and the class prediction is minus.

Now once the SVM is trained with training data the complexity of the classifier is characterized by number of support vectors. So, once we have trained a model, we have trainer model means basically we found the function, I found a function by which we call from which we found out the different parameters. Once we have done all those now the when we are into class given any tests tuple given any observation, we have to find out it belongs to each class.

Then what is the correct what is the complexity of that is this basically is not it; this calculation will give us the complexity is not it. And what is this how it is totally determined by how many number of support vectors are there because only for those support vectors will be getting this expression. If these are not support vectors then λ i will be equal to 0, then this expression will not be there.

So, we will have to do this calculation only for those X which are support vector. So, complexity is given by the support vector that is what I remember last lecture there was one point I told will come later discount support vector gives lots of information the support vector gives the

complexity of the classifier. So, the complexity of the classifier is characterized by the number of support vectors clear this is.

So, dimensional data is not an issue in SVM unlike in other classifiers, in (14:49) you have seen dimensionalities that we have faced the cards of dimensionality problem is not it. When we try to find out the distance heater if it is dimension is quite high then computation it becomes computational very intensive. But in SVM dimensionality is not an issue we will be seeing that in one of the lecture later.

But now for you to remember this dimensionality of data is not an issue in SVM. Basically, in SVM what we are doing we are just trying we are just finding a dot product this dot product. So, when you are finding the dot product here even if it is if it has a very high dimensional it will not have an a computational will not be not too much.
(refer time: 15:28)

So, now let us see an example. So, consider the case of a binary classification starting with a training date of eight tuples. Suppose these are the training data and it has two attributes A 1 and A 2. Now these values are given λ is values are not given suppose these values are given and for each values this is the training. So, training means the class level is mentioned for which we are getting plus class for which we are getting minus class it is given.

So, and then we will be using lagrangian multiplier using lagrangian multiplied and we will we will solve the values of the we will find out the different values. So, first we found out the λ values. So, using we can solve the KKT constant to obtain the lagrangian multiplier. So, if we found out the λ i for each tuple when we calculated we found a λ i for this double is this then right for this tuple is this for other λ i we found it 0.

That means what from this what you can conclude that means this two are these two tuples are the support vectors. These two all these values these values this value this value and this value it falls above the hyper plane and this and this it falls below the hyper plane that it belongs to minus plus below the hyper plane this plus it belongs to last class it falls above the hyper plane. And this two one belongs to plus and one belongs to my minus these are the support vectors that is why value of λ i is not equals to 0.
(refer time: 16:55)

Now what happens, once we know once we know the λ i, we this given the X and the x 1 and x 2 are given. So, what will be will be able to find out that w 1 w 2 what we have to find out we have to find out w 1 w 2 basically we have to find out W total what we need to find out we need

to find out λ we need to find a W, we need to find out the b is not it. So, we found out λ now we will be able to find out w 1.

So, what is w 1? w 1 we have already seen here this is λ w 1 W is not it. So, this is W, so W means since there are two parameters so there will have w 1 and w 2 dimensional. So, that what will be w 1? w 1 will be this λ i y i x i x i 1 w 1 will be this parameter for the to w m parameter will be taking this w 1 x 1 is not and w 1 x 1 w 2 x 2 for w 2, I will be considering this for w 1 I will be considering this.

So, see here λ 1 65.52 and y i class is plus so I am taking is one for plus class I told you already I am taking 1 for minus plus I am taking minus 1. So, then y of i 1 y of i 1 is 0.38 then again, I will be using this 65.52 into minus 1 this is minus class into 0.49 then I will be using this but λ i 0 the whole value will become 0 for all this value will become this will become 0, so only just to this to tuple we will consider, so this is what I got w 1 value.

Similarly, I will find w 2 value for w 2 value I will be using these two w 2 x 2, so I will be using these two values and I got this is my w 2 value.
(refer time: 18:48)

Now so once I know the w value, I can find out my b value. So, what is b 1? b 1 is 1- W dot x 1 and b 2 is 1 minus W x 2, so I found this is my b 1 this is my b 2, so both b does one only one b is not it. So, I will take the average of that since both are same values so I got b = 7.93, so I got W value, I got b value, I have all the values I got.
(refer time: 19:10)

Now I can find out my MMH the whole function, what is my function, my function is W x 1 + W x 2 + b = 0, so this is my W x 1 this is my W x 2 + b = 0, so this is my MMH. So, when once my MMH is there nothing is left just if you want to given any test data, I will be able to tell it which class does it belong just I will be using the MMH. So, how we have already seen how to find out the class of any tuple testable does the delta of X is W of X W dot X + b.

So, what is my X value, X value is point this is my x value is not it so what is my W value this is 0.64 9.32 - 0.61 so - 6.64 into 0.5 - 9.32 into 0.5 plus this b value 7.93, so I got this minus 0.05. So, that means it is a negative value, negative value means it falls below the hyper plane, it falls in a negative class. This implies that the test data falls on or below the MMH and the SVM classifies that X belongs to class level minus.

This is how we do, how we have found out once we found out the lagrangian maximum margin hyper plane just maximum margin hyper plane is known then nothing is left, like then we will have to given any testable just W X + b just find out W X + b but sign it gives if it is positive above, if it is negative below.
(refer time: 20:55)

So, now this is till now we have done for binary classification, let us see for multi class classifications. In the discussion of linear SVM we have limited our discussion to binary classification of two classes. The discuss linear SVM can handle any dimension for n greater equals to zero what we have discussed that the technique the matter what we have is discuss for linear SVM that can also handle for n greater equals to 0.

Till now our n = 2 binary classification. Now, we are this will discuss a more generalized linear SVM to classify n dimensional data belonging to two or more classes. It may be two classes or more class. So, there are two different ways how we will see it one is of course that will come different there are two when there is classes there are more than two classes there are two different possibilities.

One is all classes were pairwise linearly separable fear wise linearly all the classes. Suppose there are two class three classes one class I am writing is plus minus and into say these are the three classes. All these classes are linearly pairwise linearly separable, plus and minus a linear pairwise linearly separable, plus and into is pairwise linear separable, minus and into is pairwise linearly separable, plus and into is pairwise linearly separable, all this class are pair wise linearly separable linearly.

What is linearly separable? What is non-linear separable? I have discussed in my first class on SVM. So, there one is all are pairwise linearly separable and another one is there is non linearly separable, there are overlapping. Some I cannot separate it linearly using a hyper plane when I call it a linear event using a hyper plane, I can just separate it into two classes when I cannot separate it this way that is called non-linearly separable.
(refer time: 22:40)

So, there are two different possible days if the classes are pairwise linearly separable then we extend the principle of linear SVM. If the if it is pairwise linearly separable then we can use whatever technique we have learned for linear SVM binary classification same technique we can use here. If it is not linearly separable definitely, we cannot use that that we will see later what we will do in that case.

So, there are two set resists if there are pairwise linearly separable one strategy is called one versus one such as that is OVO strategy another is one versus all statistic is OVA strategy, OVO and OVA.
(refer time: 23:21)

And over strategy we have to find MMH for each pair of classes very easy. So, there are three class plus minus into I will have to find out MMH for each class. So, class minus into I will find a MMH for plus and minus plus and into then minus and into, all the classes I happen all different combination. So, if there are n classes then n C 2 classifier is possible, so total here if there are 3 classes I found how many classifieds three classifiers.

So, if there are n classes here there are 3 classes. So, if the n classes to totally how many classified there will be n C 2 classifiers. For an example in the figure if you see plus minus n into so here there are three classifiers for plus n minus n into plus and into total there are three classifiers. So, let H x y denotes MMH between class level x and y in linear linearization means we have to find out the MMH.

So, there are total three classifiers means we will have to find out total three different MMH. And another example you can think say for four classes if there are four classes so that how many classifier there will be four C 2 classifiers.
(refer time: 24:44)

So, with over strategy we test each of the classifier in turn and obtain del j i that is the sign of the MMH between the jth and the ith class for test data X given any test data suppose this is my test data. What I will do? Some suppose consider this example here example only where there are three classes three classes means I got three C 2 so I got total three classifiers. So, in over strategy what I will do? I will be testing my test data with all the three classifiers.

So, with all the three classifiers and I will find a sign of the when I will find out how this is remember this is how I find out the class this is del of X and I found find a sign it belongs to which class, similarly I will find a del of X for all the classifiers using all the classifiers. So, I will find del of X using all the classifiers this I am instead of writing this del of X I am giving a different thing means which two classes I am comparing del of j and i.

So, based on this I will find as sign which is belongs to which class. So, now if there is a class i for which del of i j for all j gives the same sign then unambiguously, we can say X is in class i. So, what happens suppose this is my test data I checked with this classifier first this classifier suppose I found that with this class where I found with some magnitude, I found that it falls

belongs to the minus category minus class.

Then again what I do again I will test this with this classifier, again I will get some values either this or that. There has to fall either this or this but the classification will not be proper because this the item this data actually does not fall in this class. So, classification I will not get a very good magnitude but it will fall in one of this class I will get a sign accordingly. Then again what I will do again I will check with this classifier minus and into here again I got minus class.

I will see that it will fall below this and it below I will see that I got is a sign as minus. So, here this is if this is a class i for which del of j i for all J that J not equals to i, not equals to i means for all plus and into gives the same sign I tried with plus I tried it into I got the same sign when I tried with i n plus I got minus when I tried i then into I got minus I got gives the same sign then unambiguously we can say x is in class I, this is the OVO strategy but it has OVO strategy has some cones.
(refer time: 27:38)

OVO strategy is not useful for data with a very large number of classes. Why large number of classes? As the computational complexity increases exponentially with the number of classes because if the number of classes are more see how many classes how many classifier will be n C 2 classifiers, when n becomes more you can see how many classifiers will be there, more number of classifiers that many number of times we will have to check if for each test data.

As an alternative to OVO strategies is OVA one versus all strategy has been proposed. So, what is OVA strategies? In this approach we choose any class say C 1 and consider that all other tuples classes belong to a single class. What we do says, suppose there are 3 plus minus into also is just one this and this I will consider another class. So, simple binary classification this is one and rest is other is one class.

Now again I will say is minus and plus and into is another driven class, now again also into and plus and minus is another one class, got it. So, how many classifiers I needed? So, if there are 3 classes I needed 3 classifiers, if there are 4 classes I will be needing 4 classifiers instead of n C 2 in the OVO strategy.
(refer time: 28:55)

So, this is therefore transformed into a binary classification problem using the linear as we have discussable and we can find a hyper plane. Let the hyper plane between C i and the remaining classes be MMH i. The process is repeated for each C i as I told for each plus minus into and getting the different MMH, in other words over strategies we get total k classifiers if the k

classes will get k classifiers unlike the other one where we get k classes means k C 2 classifiers which is a huge number.
(refer time: 29:33)

So, the unseen data X is then tested with each classifier. So, this unseen data so we have total one is minus and a plus and into another is plus minus and into another is into minus n plus so here is one classifier here is one classifier and raise one classifier. So, we will test it for on using all these three classifiers, the unseen attacks is then tested which each classifier so, obtained. Now this del of j X be the test result with from for each value for each classifier I will be getting this del of X is not it.

For each classifier I will be getting a del of X which has the now among this which has the maximum magnitude for this I will be getting a del of X with a particular magnitude, for this I will be getting a del of X with a magnitude, for this I will be getting a del of X with a particular magnitude the one which will give me the maximum magnitude that will be the class for the test tuple.

The X i will be the test result for MMH i which has the magnitude maximum magnitude of the test values. That is del of j is the maximum of this for all i, thus X is classified into class C j whichever gives us the maximum value magnitude for this del of j that is the class of the test number. So, note that the linear SVM that is used to classify multi-class data fails if all the classes are not linearly separable.

This we can do either OVO strategy or OVA strategy, both we can do if the datas are linearly separable, mostly uses OVA strategies OVA not OVO but then we can do only the datas are linearly separable. If one class is linearly separable to remaining other class and a test data belongs to that particulars then only it classes accurately. If we have many classes out of them one class is linearly separable to remaining other classes.

And now if we are given a test that as belongs to this class which is linear separable then we can accurately classify but there is some another class say X which is not linearly separable to other classes and my test data belongs to that group then I will not be able to classify it accurately, then my this strategy will not work properly it will give a poor result.
(refer time: 31:57)

See you can see it here, see the figure if it is possible to have some tuples which cannot be classified any of the linear SVMs like these things it cannot be classified to any of the classes because they are not linearly separable. There are some tuples which cannot be classified

unambiguously by neither of the hyper planes, unambiguously we cannot do. When we are trying with different classifiers sometimes it is getting this sometimes.

It is giving this anonymously we cannot find out this class this tuple belongs to this class only. When it will happen? When there are some tuples which cannot be linearly separable all this tuple may be due to noise errors or data and are not linearly separable, sometimes this table may be due to noise or some errors and sometimes because the data is are not linearly separable. If the datas are not linearly separable then we will have to think of some other way. So, we will be discussing next in our next lecture.
(refer time: 32:52)

So, in this lecture we derived the equation of optimal hyper plane using the lagrangian multiplier, we have seen that so, we also learned to classify given a test data we will element method through an example is not it. Then also we have seen how to extended these two class linear SVM to multi class classification problem we have seen that and we have seen that if their datas are not linearly separable then linear SVM will not work for multi-class.

Then in that case what we will do for multi glass or two class whatever it is if the data's are not linearly separable it will not give us good result. So, in that place we will have to learn we will have to use non-linear SVM, in our next lecture we will learn this non-linear SVM.
(refer time: 33:33)

So, these are the references and thank you guys.

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

Lecture - 57
SVM (Part - Ⅳ)

(refer time: 00:30)

Hi guys, so, in continuation our lecture on SVM today what we will discuss? We have discussed about linear SVM binary classification, is not it? We have seen only binary classifications on for two classes we have seen linear SVMs linear means where data are can be linearly separable. In that also we have seen for case of multi-class also in multi-class if the data is a linearly separable then what we can do how we can use linear SVM that also we have seen there are two strategies OVO and OVA strategy.

We have seen and we have seen the pros and cons of motor strategies. Today we will be thing for non-linear data, non-linear data what should we do? We should use linear SVM or non-linear SVM or if you use linear SVM then it is the same way will be used or will be changing something that will be discussing. Basically, we will and one more thing which we will discuss is that when for non-linear data now we try to make it linear.

So, for that we do a transformation that is called phi transformation that also will be discussing in this lecture.
(refer time: 01:39)

So, talking of non-linear support vector machine basically support vector machine is classification for linearly non separable data, what how we do? For linearly non-separable data so what sort of support vector machine do we use. So, here you can see in the figure there is a this so now this is here you see this this figure this is a linearly separable data. So, there are two class plus class and minus class which we can easily separate by using a hyper plane.

Now you see in this figure B you see the two data's there we can we cannot use a hyperplane to classify this to data because they are not linearly separable. So, in general if data are linearly separable then there is a hyperplane else if it is not linearly separable, we then we basically call it a hyperspace.
(refer time: 02:39)

So, such a linearly not separable data can be classified using two approaches. So, now we want to classify data which cannot be linearly separable. So, how we can do there are two approaches one is using linear SVM only but somehow twisting it twisting a bit in this or the other way. So, we will be using linear SVM only and another is we will use non-linear SVM that is now we how can how we classify data, we classify data using a hyper plane.

So, if in non-linear hyperplane we will not be using a hyper plane so we will see how we do that. And next the most comfortable technique for us will be because we are compatible with linear SVM if somehow, we can use linear SVM for separating this non-linear data.
(refer time: 03:29)

So, first we will see that then we will go to non-linear SVM. So, here in the example you see if these two sets of data you see. This is one class this is the other class so this is very much linearly separable bearing few data few data. So, here just we have two data these two data's are somehow if we use the hyper plane test data we cannot separate classify it into the two categories. So, this will be basically we can say classification error.

So, in this sort of if this is one point and this is the other point if the points are in this way, then basically, we cannot say that there is a linearly separable data. But there are very few cases in this example there are some such types where the very few data's which are which cannot be linearly separable. In such cases what we can do is that we can do the classification giving ignoring this data which cannot be ignoring just few data's which cannot be classified.

So, this is basically we are accepting some classification error because of this data and this data it cannot be if you use a suppose we have used this hyperplane red colour line if you see. If you have used this hyperplane along with this boundary this is one boundary and this is another boundary. If we use this boundary then what happens? So, this data will be misclassified. So, accepting misclassification for some data we have we have gone with the linear SVM.

So, linear SVM means we have found out that many maximum margin hyperplane. Now we will come back to this. Basically, these misclassifications how much misclassification we can afford that we will specify for that there is a parameter how we can specify, how much misclassification we can afford. So, essentially what I am going to say is that when we try to construct a SVM for linearly separable data we what we do is that we have a minimization function and subject to some constraint.

So, what was the minimization function our minimization function is magnitude of w / 2 that was

the minimization function. Now along with this minimization function what we will do? We will add another one component, another one constraint that is basically the penalty factor for the data's which cannot be classified if you use a hyperplane. So, when we use this penalty factor what happens? Our hyperplane becomes narrower and narrower.

Our margin becomes narrower and margin may be very less. It depends on how much classification error we can bear as well as how much we can what is reduce the MMH. So, it is basically a trader between the MMH and the classification error. So, we introduce another one factor that is we call it the penalty factor and this sort of changing the margin we call it a soft margin. The previous one what we do in the general case that is called hard margin.

And now when we change it this is called a soft margin. So, this I will not go into the details of soft margin because soft margin method is not very much applicable. Because when we see where some data's are not linearly separable. So, these are not a few data's and usually there are too many such data. So, when there are too many such data this soft margin does not work well, sub margin would have worked well if there are only few number of data which are not linearly separable.

Then we can reduce the margin by using the penalty factor and then we can classify data with good results. But since in reality that is not the case so soft margin is not it has not much applications. So, we will not go to the details of soft margin that is what this can be done by reducing the margin, soft margin is this concept. So, we will not go to the details of that.
(refer time: 07:14)

Now we will try to add how to say classify the non-separable data by some other means we will see that so, what linear SVM undoubtedly β to classify data if it is trained by linearly separable data obviously. So, if we have developed an SVM if you have developed a machine by using our training data is linearly separable data. For developing the machine, we have our we have used training data is not it.

Basically, if I say in other terms basically when I have computed the function that is the SVM that is y = f x this function when I have computed a function, I have used some observation. Now if these observations are linearly separable data, then in that case linear SVM is better to classify, it is better to classify any sort of training data which can be what to say given any training data. Now we can use this SVM to classify it efficiently.

Now linear as SVM we can also be used for non-linearly separable data provided that number of such instances very less that we have seen by using soft margin provided that such instance is very less. This is important. Linear SVM can also be used for non-linearly separable data

provided at number of such instances less meaning that in that case we can go for soft margin by using a penalty factor basically by trade-off between the margin and the misclassification error.

However, in real life application number of data overlapping is so high that the approach cannot cope to accurate classifier. So, soft margin does not give us an accurate classifier. As an alternative to this there is a need to compute a decision boundary which is not linear. So, if that is not possible. So, soft team soft margin concept if that is not possible that we do not get a good classifier our classification visual is poor then so to compute a decent boundary.

So, we need a non-linear distance boundary in that case because this one a linear decision boundary only can classify between linearly separable data. If the data are not linearly separable then we will need a non-linear boundary. So, non-linear boundary we also call it a hyper surface.
(refer time: 09:37)

So, here is an example this is a linear hyperplane this you can say this is a high non-linear hyperspace, this is a non-linear hyperspace. So, note that a linear hyperplane is expressed as a linear equation in terms of n dimensional component, is not it? A linear hyperplane we use we express in times in terms of a linear equation w x 1 + w x 2 + w x 3 + b = 0. So, it is very much a linear expression, is not it?

A linear equation in terms of the n dimensional component. How what is the attributes of the each data point? So, if the if a data point is two dimensional so it will be a linear equation and two dimensional and if it is a three dimensional then will be having the linear equation in a three dimensional that way. So, now if it is a non-linear hyperspace then it will be our equation will also be non-linear.
(refer time: 10:32)

So, we can express a hyperplane in this way we have already seen that. Whereas a non-linear hyperplane may be a non-linear hyperplane we can express in this way maybe. So, w 1 x 1 ^{2} w 2 x 2 ^{2} so it is basic suppose it has a dimensional is three, x 1, x 2, x 3 x is a three dimensional data. So, that means each observation has three attribute x 1, x 2, x 3. So, and if it is non-linear suppose the non-linear expression is this in this form nonlinear it can be any form.

So, w 1 x 1 ^{2}, w 2 x 2 ^{2} plus this is anything it can mean any form non-linear basically. So, suppose let us consider this is a non-linear hyperspace. Now what is your task now? It is to find a non-linear decision boundary. So, we have to find a non-linear decent boundary so non-linear descent boundary means we need to find a non-linear hyperspace. You know that is a non-linear hyperspace in input space comprising with linearly non separable data.

So, you see here this is this data's are not linearly separable so we need a hyperspace. So, as to separate this linearly not separable data. So, now these tasks the stocks of using if the data's are non-linearly separable now the as I told you if the data is a nonlinear separable we cannot use a linear hyperplane. Now that means we need to find a non-linear hyperspace. Now with how to solve this for this non-linear data?

This is something not very hard and also not very complex because there are we can basically use the technique whatever we have used for linear SVM we can extend that technique. Of course, there is one way of using non-linear that is different way but another way is that we can also extend that method what we have learned for linear SVM we can also extend that for linearly not separable data. I am not talking about the soft margin.

Soft margin also basically we use the linear separator only, linear hyperplane only, linear hyperplane but not that. Even for data for which is not linearly separable we can extend a linear SVM to solve to classify that this data. So, let us see how we will do that.
(refer time: 12:55)

So, this can be achieved in two major steps, what it is? Transform the original non-linear input data into higher dimensional space as a linear representation of data. So, our original data was non-linear here we see this data, this data is non-linear. So, this data is a non-linear in how what is the dimension? It was in a three dimensional space. So, transform this non-linear three dimensional space into a higher dimensional which how we will transform so that it becomes linear.

This is non-linear so with the we will do the transformation in such a way of course this transformation will be in a higher dimensional this linear we will transform into non-linear sorry it is non-linear will transform it to linear. So, note that this is feasible because SVM performance is decided by the number of support vectors not by the dimension of the data. Already we have seen in my last lecture that the performance of the SVM that is the complexity of the SVM.

I have already mentioned complexity of the SVM is dependent on the number of support vectors. Remember the summation of λ i y i x i is there. So, it is complexity is totally dependent on the number of support vectors, it is not dependent on the dimension of the data. So, if the complexity would have been dependent on a dimension of the data so if we increase the dimension definitely so our complexity would have increased enormously.

So, that is then that would not have been a feasible solution. Now since SVM performance is

not decided by the dimensionality of the data it is decided by the number of support vectors. So, if we increase the dimension also it will not have a very higher impact. Of course, they are definitely high dimension it goes to higher different dimensions definitely our calculation will be higher but then it will not make an impact on the complexity of the SVM.

Search for the linear distance boundaries to separate the transform higher dimensional data. So, first we will have to transform it the non-linear data into a linear data and then once we transform the nonlinear data to linear data then we will have to search for the linear distance boundaries which will separate the transformed higher trans dimensional data. Non-linear data we will convert it to linear data in a higher dimension.

So, now once we have done that, then what we will have to? We will have to search the decision boundary that is the hyper plane that will separate this transformed higher dimensional data. This we can do in the same way how we have done for linear SVM.
(refer time: 15:36)

In nutshell, to have a non-linear SVM the trick is to transform non-linear data into higher dimension than linear data. This transformation is popularly called non-linear mapping or attribute transformation or phi transformation. This transformation it is called non-linear mapping or attribute transformation or phi transformation more popularly it is called phi transformation.

And then once we transform it rest is same as the linear SVM what we have learned. So, now let us consider a second order polynomial. Say this is the let us consider second order polynomial in a three dimensional input space. So, this is the second order, is not it? So, we have a quadratic term so it is a second order we have a quadratic term. We have this x 1 x 2, x 1 x 3, x 1 ^{2} so it is very much a non-linear polynomial.

So, it is in the what is the input space? Input space is three dimensional so we have x1, x 2, x 3. Now this thing we will convert it to linear in a higher dimension. How will convert it? See here.
(refer time: 16:42)

The 3D input vector this is a vector can be mapped into a 6D space that is Z using the following mappings. This 3D we have mapped with six dimensional space, how? z of one is phi of x that is x 1, so z of 1 is x 1, z of 2 is x 2, z of 3 is x 3, z of 4 is x 1 ^{2}, z of 5 is x 1 x 2, z of 6 is x 1, x 3 all the terms that we have all the monomials that we have here. What are the different monomials we have that mono means basically we have used another one what to say attribute and that is this attributes in a higher dimension.

See here total we have six monomials so we have used six different attributes and so this is

how we have done the transformation. So, now our data X the vector x 1, x 2, x 3 it has become X has become Z where attribute of X was x 1, x 2, x 3 now attribute of Z has become z 1 to z 6. (refer time: 17:51)

So, now basically now the transformed form of the linear data so now this is my transform formula with linear data. I have just replaced the different x 1 and x 2 different form instance of X with Z. Now this is a very much a linear, is not it? Very much a linear equation. Now this is we have to find what is it, we have to find is MMH for this hyperplane, MMH means basically what we need to find out?

We need to find out a different W value we need to find out the Z value and that is all and B value. So, thus if that Z space has input data for its expression x 1, x 2, x 3 and hence Z values then we can classify them using linear distance boundaries. So, this equation we have same method we will be using Lagrangian multiplier method. For Lagrangian multiplier method now what will be our optimization function? What will be our but to say constraint?

So, accordingly we will be what to say we will solve that using the constraint and the Lagrange and multiplier that is del L / z of I del L / B and accordingly we will get the values of z, B, w. Once we get all the values of z, B, w then from this we can find out the we found out all the values of z, B, w then what happens then basically we found out the SVM. Now given any X so that X is in a 3D space that X will convert it into 6D space.

Now once we converted it into six dimensional space so what is x because for this we have we found out the SVM means we found out all the parameters below. So, we then we found out the SVM so once we found out the SVM then once given X will be able to find out which belongs to which class, same. What linear SVM we have done? Same technique we have done. Now just that now our expression has become this that is all nothing got changed.
(refer time: 19:55)

So, just is an example. See here in the figure is so an example of a two dimensional data set consisting of a class level plus and minus that is a not linearly separable as you can see very nicely. So, this is these are plus class, this is minus class definitely it is not linearly separable. (refer time: 20:19)

Suppose these data’s are given and somehow we have found out the function, we have found out the w 1, w 2 value, b value and everything. So, suppose this is the equation, we see that all instead of class minus can be separate from instead of class plus class by circle, is not it? We can separate it by a circle which is a hyperspace. So, without anonymity say the following equation of the decision boundary can be thought of.

So, this is the equation of the distant boundary. We have done the we have done all sort of what to say transformation and like Lagrange multiplier and then we will find out the W value, b value everything and then we will find out the decision boundaries. So, now here since this is circle so we can think that maybe this is maybe the equation of a circle. So, now what happened?
(refer time: 21:17)

So, this is my distance boundary, is not it? This is my hyperplane. So, any data which is in this hyperplane will be this will be equals this minus will be equal to zero, any data which is above the hyperplane will be get a positive value any data below the hyper plane will get a negative value. So, this is the expression. Now this is very much a non-linear representation, is not it? So, now this non-linear representation so we need given a test data, we have to find out.

Because we need to find out all this value, is not it. So, given a test that are done we need to find out whether it falls in which data so for that we need to find out a hyperplane. Now for finding out the hyperspace since this is a non-linear equation separating this two. So, this non-linear equation basically we can find out the hyperspace only. So, if you are interested in finding on the hyper plane so we will be using the linear SVM.
So, what we will do we will just transform this data. So, how we have used this transformation? Suppose for this we have used $x_1^2 - x_1$ we have user transformation that is z_1 and for this $x_2^2 - x_2$ this 2 term we are talking it together and we are taking another one the transformation that is z_2.
(refer time: 22:38)

So, here there was it was a two dimensional data, here I have we have transformed into two dimensional only. It is basically the transformation how you do the transformation. Here without any ambiguity, we can do this transformation, is not it? Because phi x just z_1 we can write it in this pattern because $x_1^2 - x_1$ there is no other separate coefficient also. So, very well I can write it and this also very well I can write it phi to x as that is a z_2.

So, now what is my this is my $z_1 + z_2 = 3.5$, so this is my hyperplane. So, once I know my hyperplane then I will happen I can easily given any data I will be able to tell it falls in which class. So, it was in non-linear X is in 2D space which is non-linear now we have convert Z this is also into this space and now this has become linear.
(refer time: 23:35)

The non-linear to linear transformation however there are some issues. When we have transformed this long linear to linear transformation however there are some issues like recent this last program of problem only you have seen. Here see I have used $x_1^2 - x_1$ is equals to that is one transformation I have used. Why did not I use one transformation for x_1^2 and for x_1 another transformation then again for x_2^2 another x transformation, x_2 another transformation.

So, basically why did I use that now cannot I use that. So, these are some simple questions like. The non-linear mapping and hence a linear descent boundary concept looks pretty simple but there are many potential problems. Whatever how to choose the non-linear mapping to a higher dimensional space that is our how to choose? For the last example the way I have chosen is that better or if I would have chosen instead of z 1, z 2 if I would have chosen z 1, z 2, z 3, z 4 that would have been better which should have been better?

So, for that we will have to do many again experimentation and find out. So, how to choose the non-linear mapping? In fact, as the phi transformation works fine for small examples but it fails for realistically size problems. So, if we do the phi transformation and if there are n dimensional this n dimensional only input space is n dimensional and n dimensional is only quite say high. So, then converting this n dimensional to higher dimensional it may be quite higher.

So, higher you can see here. For n dimension input instance there exists this many number of different monomials. What a monomials? If it has this two dimensional x 1 and x 2 different monomials can be x 1, x 2, x 1 x 2, x 1 ^{2} x 2, x 2 ^{2} x 1. So, different types what I am selling is suppose I have x 1 and x 2. So, different monomials can be x 1, x 2, x 1 x 2, x 1 ^{2} x 2, x 2 ^{2} x 1 then x 1 ^{2} x 2 ^{2}.

There can be different monomials so for each monomial if I use different transformation accordingly my dimensional increases. So, the total number of this is the total number of monomials we will get for a dimensional of D sorry for a maximum degree of the monomial sorry for the dimension of N and if D is the maximum degree so this many monomials we get. So, for this many monomials if we try mapping this each monomial if you try mapping to a different attribute.

So, accordingly our dimensionality increases. So, if a dimensionality increases so what happens? Though SVM as I mentioned it does not support from curves of dimensionality but still one of the dimensionality becomes very high. So, the computation increases computation the number of computation increases when we do the dot product. So, number of additional number of multiplication and number of additions increases when I do w i W dot x.

So, the computer computational it becomes computational intensive. So, cost of mapping this is one of the issue.
(refer time: 26:54)

So, that is the dimensionality problem. It may suffer from the curse of dimension often associated with high dimensional data which I have just already mentioned. More specifically in the calculation of W dot X or X i dot X when we try to find out the class of X then we do X i dot X

is not it, the dot product. So, if it is of higher dimensional what happens? When you do X i dot X how many multiplication and addition we do?

If it is an n dimensional data for n dimensional data, we do n multiplication and add n addition. If accordingly, if our dimensionality increases the number of multiplication number of additions increases. As the number of input instance as well as the support vectors are enormously large it will be computationally intensive. So, that way it will also suffer from the dimensionality problem.

So, again solving the inequality constraint optimization problem in the high dimensional feature space is very computational intensive task. Now when we have to find out the values of this different parameters using the Lagrangian multiplier method what we have seen. So, when using this inequality constraint optimization problem when we solve it for high dimensions space it becomes very computationally intensive task.

Because they will be getting many expressions many equations, is not it? If the n dimensional so we will get when we do del L / del x i. So, if this n dimensional we get n expression. And similarly, if there are more than if the dimensionality is more the number of equation is more. If the number of equation is more than accordingly solving this equation will also be very computationally intensive.
(refer time: 28:31)

Fortunately, mathematicians have cleverly proposed an elegant solution to their work problem. Now discursive dimensionality it would have been a really a case to SVM because non-linearly separable data if we can use linear assume that is very good. Why unnecessarily complicate the issue? Where linear SVM is very simple. So, if you can use it in linear SVM then it is definitely very good but linear SVM we have seen two technique.

One is soft margin which do not which do not work well. If the number of data which are basically which are very near to the hyperplanes number of such data is large so it does not work well, we have seen already and moreover. So, other technique was data transformation. If data transformation if for a while doing data transformation if we face the case of dimensionality then the advantage of data transformation is gone.

Then no point using the linear SVM then we should basically go for nonlinear SVM only. But mathematician has cleverly proposed an elegant solution. So, this cause of the dimensionality problem we will not be we will this even if we do data transformation so we will not face this dimensionality issue. The solution is like dual formation of the optimization problem and kernel trick. So, what is the solution?

The solution as it is and there is a concept called dual formation of the optimization problem whatever optimization problem we had. What is the optimization problem? Minimize W / 2 to a constraint, is not it? So, that was our optimization problem. So, we will find a what to say dual formulation of that which is much easier to calculate. Because here we have seen in this optimization problem when we do del L / del x i and we will have to find out the del L / w.

So, all the number of equations increases tremendously. So, if we can find out some sort of formulations which will reduce the computational thing task in the while what to say while finding this inequality constraint optimization while you are trying to find out the maximum merge in hyperplane while trying to optimize this if you can reduce the computation so then it will be much beneficial for to use the linear SVM for non-linear SVM data.

So, that is called dual formation we will see that in our next lecture. Then and there is another concept called kernel trick that also helps in this what to say becoming this transformation. It has as we have seen it would have it faces the case of dimensional problem but how does kernel trick and dual formulation of automation problem will help in what to say improving the performance of this, we will see that in the next lecture.
(refer time: 31:33)

So, in this lecture we have introduced the concept of linearly separable data and linearly not separable data. Then we have learned about linear SVM, how we can use linear SVM to classify non-linear data. We have seen two methods but soft margin we did not discuss of course. Further we learned about the idea of non-linear to linear mapping to handle non-linear data that is the phi transformation.

So, in the next lecture we will be discussing this kernel trick as well as this one dual formation of optimization problem. This will be discussing in our next lecture.
(refer time: 32:11)

So, thank you guys.

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

Lecture - 58
SVM (Part - V)

(refer time: 00:31)

Hi guys, so and continuation of our lecture on SVM. Today what we will see as I have mentioned in my last lecture so this transformation data transformation from to make it the non-linear separable data which are not separable linearly. So, there is non-linear data, we try to use linear SVM product, for that what we have done we have we try to use so that we can use the linear SVM for this non-linear separable data.

We have transformed the data from non-linear data to linear data for that what we have used. When transformation was in a input space of say X and that transformation we have changed it to a different input space of a higher dimensional. So, when we have changed it to higher dimensional the computational increases. Computational when we try to come compute the optimization constraint problem.

They are also computational increases when we try to find out the different parameter's values there also the computational when we do the dot product W . x w r x i . x. So, this dot product basically we do if it is an anime student space so it is n multiplication plus an addition, so similar when the dimensional increases the computational increases. So, now how to solve that how to address that.

So, these are two concepts one is the dual formulation of optimization problem and another is that we will introduce a new very intelligent method which is called kernel trick.
(refer time: 01:59)

So, as we have discussed, what are the different issues the first of course there are one issue I forgot to mention that one issue is also mapping, how will I map what mapping procedure I will use as I have shown you if in the whether I will use suppose the dimension is in R 2. So, then dimension of R 2 whether I will keep it in R 2 only or I will use I will transform it to R 3 or I will

transform it to R 4 basically how many attributes do I need to use.

So, based on the number of monomials. So, how many how each monomial's, I will be mapping into an attribute of the higher dimension. So, that is this mapping which mapping exactly there is no hard and fast rule that I will have to use this mapping. So, since there is no such hard and fast rule, we how I will do the mapping that is one of the issues then definitely cost of the mapping dimensionality problem and computational costs. So, these are some of the issues.
(refer time: 03:01)

So, mathematicians have clearly proposed an elegant solution to the above problem. So, that is the dual formation of optimization problem and kernel trick. So, let us see what is that?
(refer time: 03:13)

First, we will see dual formally formulation of optimization problem.
(refer time: 03:18)

So, we have already learned a Lagrangian formulation to find a maximum margin hyperplane is not it. So, we have formulated our problem as a maximization problem, earlier it was a minimization problem 2 / W and which we have changed it to W / 2 and we found it we have formulated as a maximization problem. So, we have learned a Lagrangian formulation to find a maximum margin hyper plane of a linear SVM classifier.

Such, formulation what we have seen such formulation is called the primal form of constraint optimization problem. The formulation what we have done till now, this is also called the primal form of the constraint optimization problem. So, primal form of the Lagrangian optimization problem this was my primal form I have to minimize this $W^2 / 2$ is not it subject to this, is not it.

My difference between the two, what was that my difference between 2 vectors, I have to maximize this is not it so what was this that is 2 / W, so I had to maximize this is the difference. So, this I have formulas as a minimization problem how did I formulas as a minimization problem I have used this, is not it. So, I have to minimize this subject to y i W . x i + b greater equals to 1.
(refer time: 04:51)

So, this was my Lagrangian form, this formulation minimize this subject to this, thus I have written in a Lagrangian form as this earlier I have not used the term p, I have just written L is equals to this this was my Lagrangian form remember we have done this. Now, let I am calling

this this thing actually it is basically that time we did not mention specifically this is called the primal form, this is the primal form.

So, a substrate we can use a subscript p to distinguish that this is a primal from. Why we are using this because there is some other form also that is the dual form. Now, what we will see what is the dual form, now here what is the λ, λ is a dummy variable it is also called lagrangian multiplier. So, L p is called the primal form of lagrangian optimization problem.
(refer time: 05:43)

So, now from this is the primal form, now from this we will have to find out the dual form. Why we will find out the dual form? We will see that. The dual form to find out the dwell from what we have done first as we have seen to minimize this lagrangian what we have done we have done del L / del W del L / del b, from that we have found different expression is not it. So, here del of L p / del W we found this W is equals to this is not it.

Because, what was this W minus this we got 0 so, W is equal to this that is what we have done earlier you can refer to the previous lectures. And this also when we have done del L / del b we got this summation of λ i y i = 0. Now, what I will do is that in my primal form lagrangian, I will be substituting this W value and I will be substitute, basically I will be substituting this value and this value and my I will substitute this in my primal lagrangian form.
(refer time: 06:49)

So, this was my primal lagrangian form. So, first I have done first order differentiation with respect to W and I got this then, I have done first order differentiation with respect to b and I got this. This two I will be using in this expression and there is actually a big one-page calculation. So, I did not show it in this slide so this once we substitute this and once, we will be using this.

In this expression we will get something of this form, where there is no b no W all in terms of λ y and x y is the class and x is my observation. So, when I just use these two equations this one, this equation and this equation in this equation after lots of calculation is as I told you it is a big one page calculation then, we will be finding this so in short I can write it this way and this I am calling it a dual form dual lagrangian form.

Because, primal (()) (08:00) that was my lagrangian form and in that lagrangian form what I have done I have done some sort of substitution, I did not do anything else I have done just some sort of substitution. What I have substituted I have differentiated with respect to W, I have the sensitive with respect to b and whatever I got that I have substituted it in primal form and I got a different expression which I am calling it the dual form, dual form of lagrangian.

So, this is my dual form of lagrangian. So, in my dual form of recognition I do not have W term I do not have b term and this is my this my quadratic term is minus here here, my quadratic term is this is my quadratic term is not it this will give me a quadratic term this is minus here.
(refer time: 08:41)

So, the key difference between the primal L p and the dual L p forms of lagrangian optimization problems are given here. So, there are different differences there are many differences between the L P and L D. So, what are the difference let us see, so L p involves a large number of parameters obviously, our L p has last number it has W it has b it is λ I on the other hand L D involves only L i that is the lagrangian multiplier.

L p is the minimization problem as the quadratic term is positive. L p when we use L p there is a minimization problem because there the quadratic term is positive but however in L D our quadratic term is negative, I have already mentioned there hence it turn outs to be a maximization problem. So, since a quadratic term is negative then, it is no longer a minimization problem basically it turns out to be a maximization problem.

So, L p involves the calculation of W . x, in L p what we do when we find out L p so our calculation is w . x dot product of W and x. But whereas in L D involves a calculation of x i . x j x c we are trying to do dot product of x i and x j, two different observation data. Whereas in L p we have done dot product of W and x if W is n dimensional definitely if x is in n dimensional W is also an x dimensional.

W . x means we will have to how we do calculate a dot product is not it, that you know I do not have to tell you. So, but whereas in L D I will do the calculation of x i and x j. If I do the calculation of x i and x j then, things become easier why the what is when we try to find out a dot product of two vectors x i and x j what we try to actually find out everything, it is there in the slide let us see.
(refer time: 10:45)

So, there are key difference, so the SVM classifier, now once we have used the lagrangian multiplier in a primal form earlier we have seen Lagrangian multiplier we have used and then we have found out the different values of the parameters W b and all. And then given any test data we have to find out it falls in which class, if which class so this is this was what we use remember.

So, del of x means which class it belongs to so this is how we do it W . x + b so x is the test vector. So, if we find it a negative sign then it falls in the class below the hyper plane if it falls in a positive sign, it falls the class above the hyperplane, what may be the above class what may be the below class that accordingly what you have specified. Where is the dual version of the classifier?

So, it was in primal form this was our thing whereas in the dual form we will be using this because, this is my W is not it, this is my W. In dual form I will not be using this W . x + b but instead I will be using this formula x i being the is support vector and assume that there are m support vectors. See the difference, here in the primal form to find out the class of a x we have what we have done we have found out the W value.

We have found out the b value W along which the if it is the n dimensional so W 1 to W n we found out the value and definitely x is n dimensional then only W will also be n dimensional. So, accordingly we found out the different x values also and we do the dot product of this and then we find out the magnitude and the sign of x so as to classify (()) (12:41) to predict it falls in which class.

Now, in the dual form, dual form we no longer do the dot product of W . x but whereas in dual form this is our expression. Here, so this is we do we have a dot product of x i . x where x is my test vector which is x i x i ordered different support vectors, we have seen this here x i as the different for all those support factors λ i is not equal to 0. For vectors which are not support vectors λ i is equal to 0 is not it.

So, this equation becomes zero so here will be only multiplying all those support vectors all those vectors which are support vectors and with those I will do the dot product of my test vector. And if I do the dot product of the stress vector I what difference it will make will see. So, till now we have as I told you for solve the time different issues of the transformation using the linear SVM for the non-linear data, what are the different issues we have seen.

So, now to solve those issues as I mentioned there were two ways how we will solve it then, there is not two ways there is a trick how we will solve this problem the first step of is dual formulation. So, we have seen we have seen this dual formulation now next is the kernel trick. Thus, here this does the L p and L D are equivalent is not it, because already what instead of W I am using this the W value is this we have already found out. So, L p and L D are just equivalent.
(refer time: 14:20)

So, it is just a different form, this is the primal form, this is the dual form now we will see what is the kernel trick.
(refer time: 14:27)

So, we have already learned an idea that the training data which are not linearly separable can be transformed into a higher dimensional feature space such that in higher dimensional transform space a hyper plane can be decided to separate the transform data and hence the original data. So, this was not linearly separable we have transformed into the higher dimensional space.

Once we have transformed into higher dimension space we could use a hyper plane, this is the hyperplane we could use a hyperplane to separate this, so non-linearly separable data because now it has become linearly separable.
(refer time: 15:05)

So, clearly the data on the left of figure this is not linearly separable. So, this is not linearly separable yet if we mapped it to a 3D earlier it was supposed in 2D we have mapped it to the 3D space using φ transformation then with the map data it is possible to have a linear dimension boundary, linear decision boundary that is a hyperplane in 3D space.
(refer time: 15:29)

So, see this example suppose there is a set of data in two dimensional in R 2, we have a set of data which is in two dimensional, let the hyperplane in R takes this form. Then, it is data is in the two dimensions suppose, if I consider a hyperplane because it is these are not linearly separable it is nonlinear separable suppose it takes the hyperplane takes this form quadratic form.

Which is the this is nothing but the equation of an ellipse or maybe you can consider this you can consider this sort of data this is an ellipse. So, now suppose we will be doing Φ transformation. Now suppose φ is a mapping from x that is in R 2 to Z z 1 z 2 z 3 to R 3 in 3D space, I will do the mapping. So, say for this term I have used Z 1, Z 1 = x 1 z 2 is equal say this term, Z 3 is equal to say this term.

There are three nominal's, so for each 3 nominal's each nominal I have used one attribute of higher dimension. So, say z 1 = x 1 ^{2} z 2 = root over 2 x 1 x 2 z 3 = x 2 x 2 ^{2}, I have used three different attributes for the different nominal's there were three nominals, so I got this.
(refer time: 17:02)

Now that means when I use this transformation then this has become my equation. So, this is very much a linear equation, is not it this is very much a linear equation so that is that means it is a hyperplane earlier it was a hyperspace. So, now this is very much a hyperplane from which we will have to find out basically the different values. Once you can define the values, we can will be able to classify the test data. This is clearly a linear form in 3D space.

In other words, W . x + b = 0 in R 2 is has mapped into W . z + b dash = 0 in R 3. So, a 2 dimensional it is W . x + b = 0 which we have mapped it to 3 dimensional W . z + b dash if we use the different com intercept that is b dash = 0 that is in R 3, this means the data which are not linearly separable in 2D now R separable in 3D because this is a linear plane. This plane separates the data. Earlier which was separating the data this was separating the data.

It was a hyperspace. A hyperspace was separating the data now what we have we have transformed into a linear plane. This linear plane is now separating the data. So, now we will see the generalization of this formulation which is the key to the kernel trick now we will see what is the kernel tree.
(refer time: 18:38)

See here, so our classifier is earlier our classifier was W . x + b is not it, what is the classifier to given a class X a given an observation at it is false to which class how do I find W . x + b we have seen here, this W . x + b, now it has our thing has become this, this is our classifier now. So, this is my classifier now, I have used a dual form, dual form of the lagrangian. So, this is my classifier now.

So, now this is my classifier in the x dimensional whatever it is my input dimension input dimensional may be 2D 3D over the waves. Now, since I am converting it to an higher dimension I am doing the transformation so as to convert this linear to not sorry non-linear to linear so this transformation I am using say z. So, now I will class for me now if I want to classify given a test that x it falls into which class basically x first.

I will have to convert it to z and that means I will have to find out this z belongs to this class. So, I am now I am telling my classifier z instead of classify x now I can write classify z is not it because, x I will convert it to z. So, classifier z if I convert it this, I can write it in this way x i I got I would have done this φ transformation x also I have done this φ transformation. Earlier, my dot product was between x i . x now my dotproducts is φ . x i to φ . x.

So, what was my dual problem? Since, it was my negative term is the quadratic term is negative

so it is maximized we have already mentioned. So, I will have to maximize this is not it. Once, I maximize this I will find out by maximizing this I will be able to find out my λ value. So, maximize this subject to this constraint we have seen this constraint is not it, when we have found out this expression when we found out this expression then we have seen this constraint.

So, essentially, we will have to maximize this subject to this constraint and when we maximize this subject to this constraint accordingly, we will be able to find out the λ value. Once, we found the λ value then we will be able to do this classification by doing base dot product.
(refer time: 21:14)

Now, question here is how to choose φ the mapping function that is X is mapped to Z, so that linear SVM can be directly applied. So, a breakthrough solution to this problem comes in the form of method known as kernel trick. How to do this mapping? That was our main issue, so that is a one the very good solution is the kernel trick. So, with now what is the kernel trick, we know that a dot product here that I was still talking of this only I told I will discuss later.

So, when we talk of dot product, dot product is what we often try to measure the similarity between two input vector is not it when we do the dot product actually what we have to do we want to find out the similarity between the two vectors. So, this is also we also call it cosine similarity, the angle between the two vectors. So, if the angle between the two vector is 0 then, we can say both the factors are almost similar.

However, if the angle between the vectors is 90 degree, then we can say both the vectors are orthogonal, is not it? So, for here it is written we know that the dot product you know right is what is the dot product what we dotproducts cross product you know all those things, is not it? So, the we know dot product is often regarded as a measure of similarity between two input vectors we call it cosine similarity.

For example, if X and Y are two vectors then we can write it in this way. So, how do we find out the angle from this expression we can find out the angle, is not it? What is cos theta, cos theta is magnitude of X into magnitude of Y / X . y, that will give me cos theta. So, theta will be cos inverse of that what we get, so for similarity between X and Y is measured as a cosine similarity.

So, if theta = 0 then we and then they are more similar the vectors are more similar if the theta is 90 then it is orthogonal as if theta is also 0 then cos theta = 1 then if theta = 90 what will be cos theta? Cos theta will be 0. So, if the cos theta values reduces from 1 to gradually 0 our similarity reduces gradually.
(refer time: 23:42)

So, now analogously if X i and X j are two tuples then X i X j is regarded as measure of similarity between X i and X j, see here when using the classifier what I have used X i . X j I have used the two dot product. Similarly, here also I have used two dotproducts. So, if X i and X j are two tuples then X i . X j is regarded as a measure of similarity between X i and X j.

So, again φ of X i and φ of X j are transformed feature of X j as a respectively, does φ of X i the φ X j is a is also should be regarded as a similarity measure between φ of X i and φ X j. Because, if X i is a vector, definitely φ of X i is also a vector is not it. We just we have transformed it from one dimensional to another dimensional but, it is X i is also a vector to φ of X i is also vector.

So, if X i . X j if you regard is as a measure of similarity then φ of X i . φ of X j also will be regarded as a measure of similarity in the transform space and a different dimension. This is an important revelation and is the basic idea behind a kernel trick. Now, natural equation arises if both measures the similarity then what is the correlation between them. So, now X i . X j this dot product also measures the similarity.

When we try to find out the class of a test data basically, we try to find out this test data how much it is similar to the different support vectors when we are doing the dot product means essentially, we try to find out the state states that is how similar with the support vectors based on the similarity we basically classify it to test data or the other data. So, what we have done for X i . X j X i . X basically that is the test data my test that is X.

We find out the similarity measure between the support vector and the test vector. Here, also it is φ of X i do φ of X j when we do here also, we do the same similarity measure but in a different dimensional in a transform dimension, is not it? So, now what is the relation between that is there now is there any correlation between X i X j and φ X i φ X j. Now let us try to find the answer to this question through an example.
(refer time: 26:11)

So, without loss of generality let us consider a situation stated below. We will be using the same equation which very recently we have used. So, suppose it was from R 2 the input dimensional to input space we have converted into R 3. So, it was x 1 = Z 1 x 2 = Z 2 2 root 2 x 1 x 2 12 is actually the same example which we have used. So, if my x i is this x i 1 x i 2 if my observation given absorption x i is x i 1 x i 2 and x j is this are designed any two vectors in R 2.

Similarly, then my φ of x i will be this and φ of x j will be this is not it nothing but the attributes here the I have three attributes Z 1 Z 2 Z 3, so these are the three attributes in my transform space. So, under two transform function of x i and x j in R 3.
(refer time: 27:12)

So, now let us try to find out the dot product of φ of X i and φ of X j. If I find out the dot product of φ of X i and φ of X j now this is my one vector this is one vector, I cannot, I write it in this way I will just when I have to find out a dot product. So, I have written it in this way, so this is what I got. This is nothing but this is a vector φ of X i is a vector φ of X j is a vector I have written it in this way and I got this.

This again I can write it in this way is not it, this I can write in an a + b ^{2} again, this I can write it in this way and what is this is nothing but X i . X j ^{2}. See here that means φ of X i . φ of X j is X i . X j^2. Ignore the term whole square, whatever it is basically what I found as φ on X i and φ of X j is somehow related to x i x j that means there is a correlation between φ of X i . φ of X j.

And X i . X j there is some correlation, the square cube whatever it is forget about that just that φ of X i . X j is score related to X i of the example because we have seen when we have done φ of x i . φ of x j ultimately, we got X i . X j ^{2}.
(refer time: 28:39)

So, we took reference to the above example this that we have seen we can conclude that φ of X i do φ of X j are correlated to X i . X j. In fact, this we can in general also we can prove that, for any feature vectors and a transform feature vectors. We can take, if we take any picture matters any transformer this is just a simple example, we have seen for any just some feature vectors and is transformed if we see we will be able to prove this there is a correlation between the transform vector as well as the original vector.

And this proof it is not possible to discuss the proof here. More specifically there is a correlation between dotproducts of original data and dotproducts of the transform data. So, based on this above discussion we can write dot product of X i does X j it implies that the dot product of φ of X i of φ of X j it implies this K, this K denotes a function which is more popularly known as kernel function.

Now, what is this kernel function what is this K we will discuss in the next lecture just to give a hint just to understand what is this kernel function. Let me give you an analogy of a very non-technical analogy like suppose, I have here two fruits and I have asked you to compare these two fruits. So, the person whom I will come whom I have asked to compare; this two fruit

suppose he is not a very intelligent person.

So, then what he will compare if both the fruits are red suppose then this both the fruits are red that is once simulate means how does both the fruits are how similar they are if I want to find out. So, he will tell both the future red in colour both the fruits are round size both the fruits have the almost same size. So, this sort of comparison will give but if I ask to compare to a very intelligent person, so on what way the intelligent person will compare.

What she will do is that she will try to find out what are the different values of this food, how what vitamin E has it has B vitamin magnesium, zinc whatever it has and accordingly the other fruits all over it has and how it helps the person who consumes it how it improves the humidity of the person who consumes it. So, accordingly you will find out the similarity measure of these 2 fruits that is how intelligent person will do the other dumb person will just based on the colour shape size will give in compression.

Now you see this the intelligent person will do this sort of compare similarity between this to fruit. Now, likewise suppose if consider two other fruits in the different soil this is in Indian soil suppose in UK soil, in UK soil suppose we have two different fruits. So, in the similarity measures so if we try to find out a similarity measure suppose this when you are comparing that you say two different fruits here.

I have compared to fruits X and Y there in the UK soil another two fruits A and B so similarity measure may be same level of vitamins content zinc content magnesium content whatever it is and how it improves the immunity level. So, if that is there is a correlation between those two fruits and these two fruits in Indian soil then we can say these are correlated, these two maybe these are different fruits the loop wise it is may be very different.

So, but at similarity wise this is the simulative measure of this two, what how it affects a human being whoever is consume it and similarly these two fruits in the Indian soil how it is affects how when a person consumes it this way, we can say this to food if the same similar we can say that they are correlated. So, that is what basically is the kernel function, now what is key there are many standard functions.

Which we will which can be used as a kernel function which and how we can use this for us in this SVM that will be seeing it in my next lecture.
(refer time: 32:51)

So, in this lecture we learned about a dual formulation of the optimization problem, from the primal form how we have converted it to the dual form and we have also learned an important concept that kernel trick we have just learned it we how it can be used and how it helps us in solving this transformation problem that we will be seeing in my next lecture.
(refer time: 33:17)

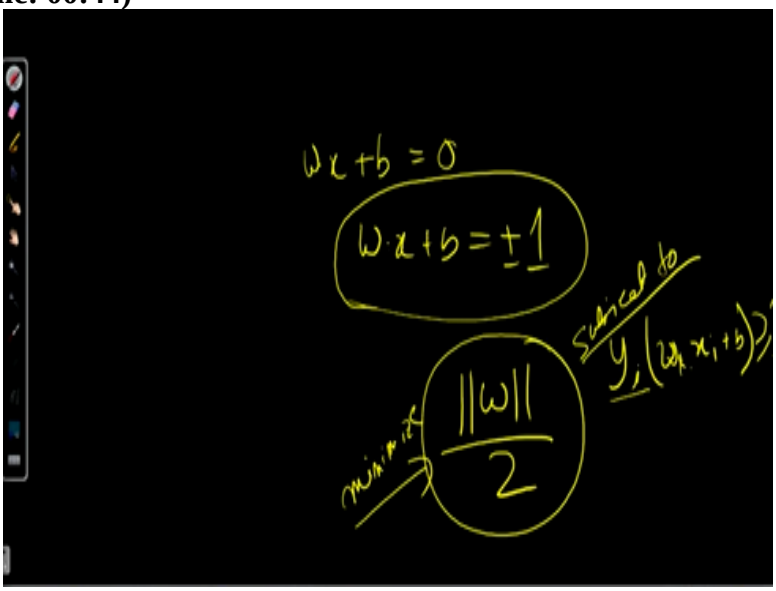
With that thank you guys.

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

Lecture - 59
SVM (Part - VI)

Hello guys, so at last we have come to the end of this lecture. This is the last but one lecture of this course and in today's class we will wrap up our discussion on SVM. And then our last lecture is a tutorial class.

(Refer Slide Time: 00:44)



So, now before going ahead with this lecture today let us just take a quick recap what we have learned in this last four five lectures on SVM. So, basically, we are doing SVM for a classification purpose we were talking of SVM for a binary classification problem (()) **(00:59)** datas which are linearly separable. Then we went on to find out the equation of the hyperplane. So, what was the equation of the hyperplane.

So, basically it was W . x + b = 0 this is my equation of the hyperplane. Now I we have already discussed this is W . x this is a dot product. So, we have already discussed that this expression if we divide by any constant value, we will get the then the hyperplane will that way we will get will be getting infinite hyperplane but all with the same slope and the same intercept. So, now to just to avoid this confusion we went ahead.

And we found out that W x + b then W x + b = + - 1 which is nothing but two pipeline basically two line parallel to the hyper plane we can say it is a there is a parallel boundary we can say. So, now and then this was our equation of the hyperplane these are just two parallel hyperplane. Now our task was to find out the maximum margin hyperplane. So, in optimization task in the classification in SVM our main objective was to find out the maximum margin hyperplane.

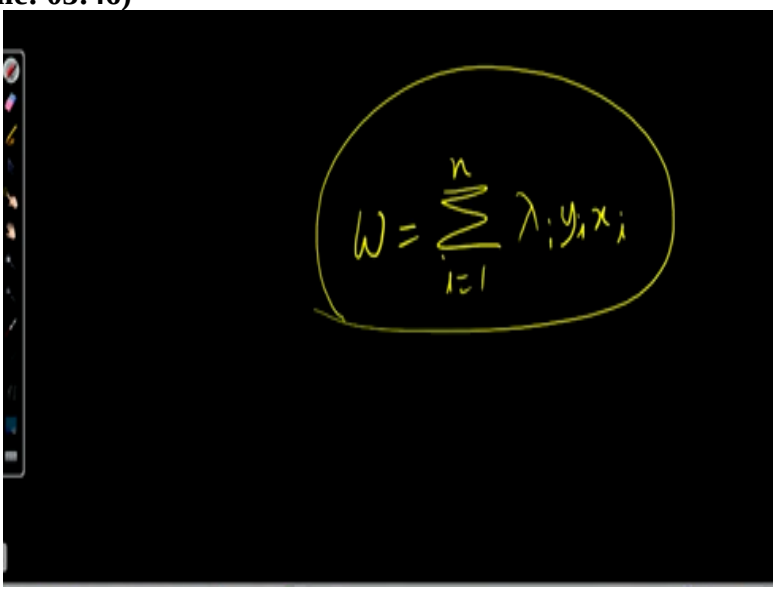
The hyperplane that will give us the maximum margin is not it that was our objective. So, with that objective then we went on to see that this is very much an optimization problem where we have used this the way we try to optimize this function let me write it here we try to optimize

this is not it. And subject to there was a constraint subject to the constraint what was the constraint w i y i w i and w i . x i + b this is greater equals to 1 subject to this constraint.

So, this is we have formulated this computation of maximum margin hyperplane. Our task is to compute the maximum margin hyperplane. So, computation of Maximum hyper margin hyperplane we have formulated it as an optimization problem so what is the optimization, we need to minimize this thing minimize. We need to minimize this subject to what subject to y i then to w i . x i + b which is greater equals to one y i is the different classes.

Different means there will be two class. So, subject to this we have to minimize this and then we went ahead and to use lagrangian multiplier for optimizing this problem is not it for optimizing this formulation.

(Refer Slide Time: 03:46)



So, like now than what we have done. So, we have while trying to while using lagrangian multiplier we have introduced a dummy variable the dummy variable was λ remember. So, now then after we had formulated the lagrangian then we have differentiated with respect to the input variable that is x i x i of all we using all the attributes. So, we have differentiate the lagrangian with respect to that as well as we have differentiated the lagrangian with respect to the λ.

Then we got different functions for that we try to find out the solution using any of the equation solving method. And then we found out the different values of the w n b so, while trying to do that we found out our W what is our W is equals to we found out our W = i = 1 to n λ i w i x i is not it. This is this was our W so n is the total support vectors. Now what happens once this is then again till this it was ok for us if we have data which was linearly separable.

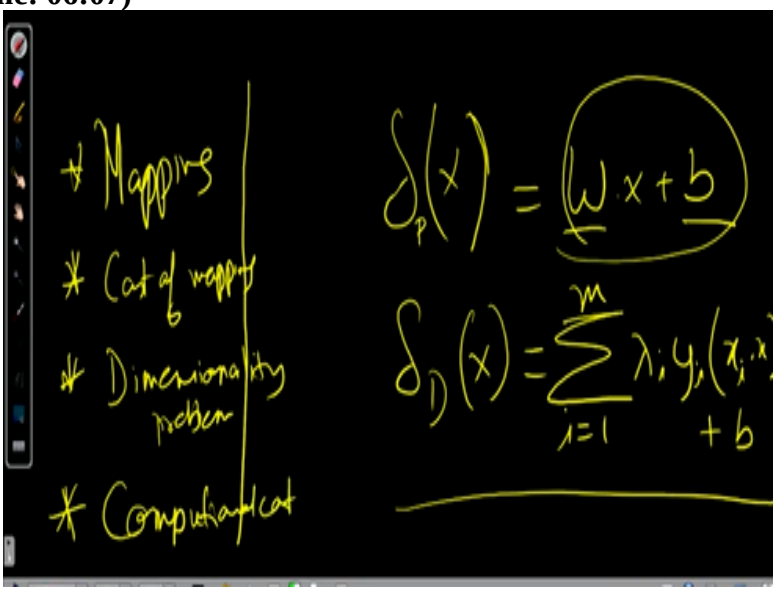
Now if the data's are not linearly separable then what we have seen there are one way is to use the if some of the data not be very few data's are not linearly separable then we could use the soft margin. And then if the soft margin is not applicable then we went ahead to use the transformation method like trans we have transformed the data from one space from one input space basically from the original space.

And we have transformed the data to some other space of a higher dimensional by doing what we could solve the non-linearity of the data. If the data is non-linear so what happens what we

have done this data we have transformed into some other space and once we have transformed it into some other space then we have used the same what to say same optimization problem and then we have solve it using lagrangian.

But there what we have said instead of this initially the lagrangian form which you have used we have called it as a primal form that is the initial form. Then we have changed a bit and we are calling it a dual form.

(Refer Slide Time: 06:07)



So, dual what is the dual form of the lagrangian let me before going to the dual form now using the initial primal form of the lagrangian when we had to classify a data suppose we are classifying a test data x so, I am using it as to find out the classification updates test data we had to this is my step. So, from optimizing the lagrangian I got the value of w I got a value of b then using this w . x + b I could find out where my x data belongs to which class.

So, this was till then. Now if the data is not linear then we have transformed and we have used the primal I mean so sorry the dual form of this, when we have used a dual form what was my classifier now. So, let me call it as this let me call it that primeval this may call is a dual form what was my dual formula classifier. Now during my duel from the classify and water equivalent just that it is i = 1 to m then λ i y i x i . x + b this was my dual form of the classifier.

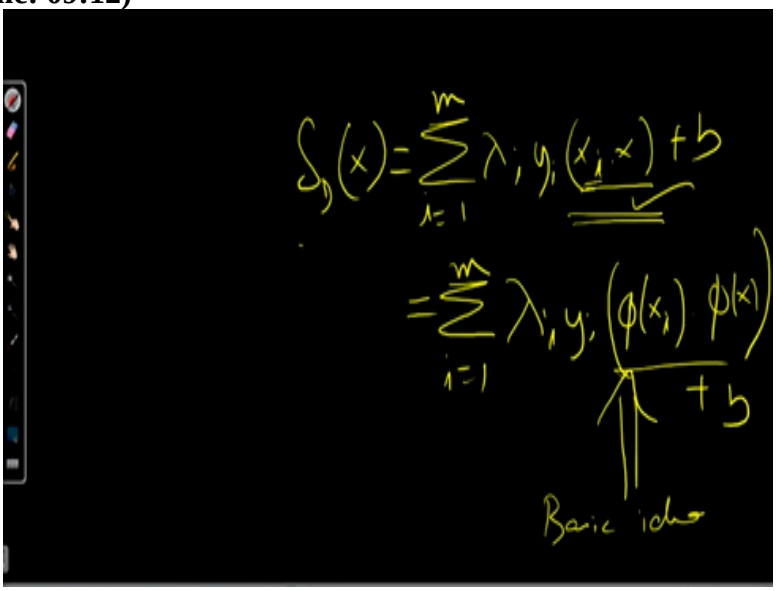
And then the optimization function also change it was a big function I am not writing it again you can see it in my slide. So, this was the dual form of the classifier. So, updating was till here but then we had some other problem. When we have transformed the data from one input dimensional space to some other dimensional space which is usually higher than this it is never lesser at sometimes, we are in the same dimensional but most of the time we go today higher dimension.

So, when we do this transformation there are some issues with this. What are the issues the different issues? One of the issues is mapping. How do we do the mapping? That is how because for each monomial we will select a different mapping function or combination of two three monomial I will select a mapping function. So, how we will do the mapping this is one problem. Second problem is the cost of mapping.

Then the third my third problem is the dimensionality problem. Dimensionality problem usually SVM does not suffer from the curse of dimensionality but and when it goes to a very high and high dimensional space computation it becomes very computational internship. Then it may suffer from dimensionality problem as well. The dimensional problem as well as computation cost dimensionally problem means computation cost will increase both are related actually as well as computational cost.

So, these were the issues of the transformation of data from a lower dimensional space to a higher dimensional space to make the data linearly separable. Now then so mathematician have found a very intelligent way of dealing with this so, where we do not at all have to do the mapping without doing the mapping how we can go about it. So, what was the way?

(Refer Slide Time: 09:12)

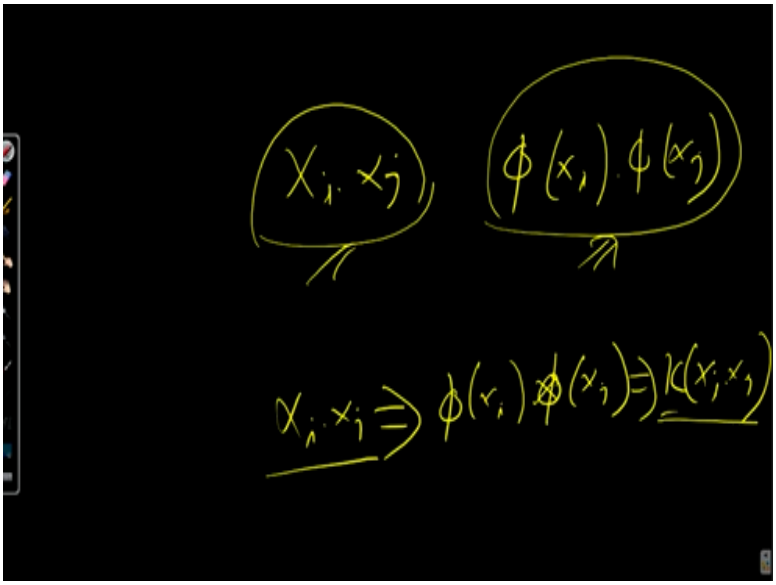


So, what we have done here? Then next what we tried doing is that like so what was our what to say classify function let me write my classifier function again. Classify function in the dual form so what was my classifier function? My classifier function was λ i y i x i . x + d i = 1 to m this is my classifier function and a dual form. So, now this is in this classifier function what we see this is the dot product between the this is one test data for each days that I will see a dot product of this x is a test data and this is my training data.

So, with the test data is a dot product of each and every training data. So, this is what when I try to find a class of x this is what we do dot product of this. Now when we; talk of dot product. What is dot product? Dot product actually we try to find out the similarity. When we find out a dot product between two vectors so if we try to find out the similarity is not it. The similarity of the two vectors x i . x so, now when we have transformed it into a difference space.

So, this equation again and in the transform thing dimensional transforms dimensional space I can write this φ of x i this is the transform data and this is φ of x + b just I have transformed this is not it. So, here this is the; if now the condition is if the x i . x gives a similarity measure. Similarly, φ i x φ x i . φ x is also the similarity measure between φ x i and φ x this is the basic idea between behind a kernel trick, this is the basic idea. This is the basic idea behind a kernel trick.

(Refer Slide Time: 11:26)



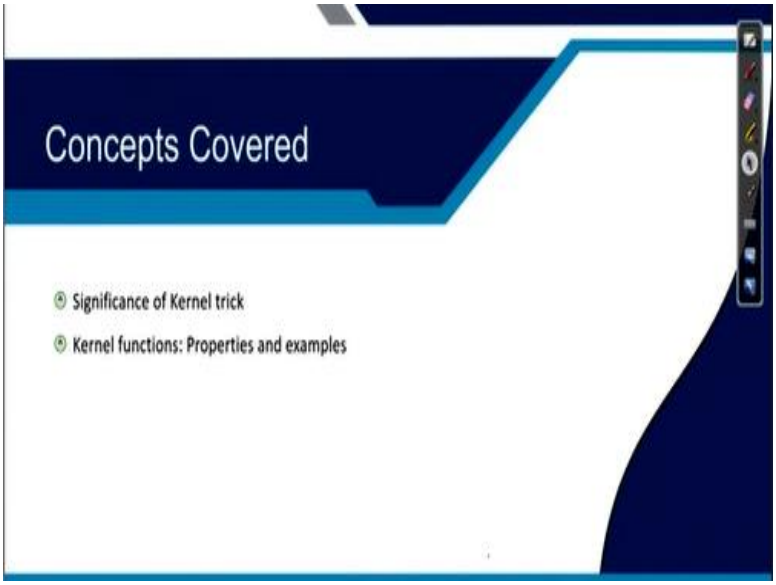
So, what we went on to do is that so, now first thing is that when we have seen this X i X i . X j and φ of X i . φ of X j so this gives the similarity between these two data and these two vectors this gives the similarity between these two vectors the what is φ X i is a transform vector of X i φ X i is the transform vector of X j. Now this also gives similarity this also gives similarity. So, we were wondering is there any correlation between these two.

And then we went on to find that this two are correlated φ of X i . φ of X j is very much correlated to X i . X j so, because then we can do, we could write that X i . X j implies to φ of X i . φ of X j implies to K of X i . X j, what does this mean. This kernel functions physically implies similarity this kernel function K it physically implies similarity in the transform space consider using the input space.

So, I am repeating again what does this kernel function it implies similarity in the transform states φ of X i and φ of X is a similarity in the transform space. So, K of X i it gives the similarity in the transform space considering the input space taking the input space. We have taken the taken the data of the input the dimensional in the input space where it is whatever dimensional it is whatever two dimensional three dimension it is there have taken that data.

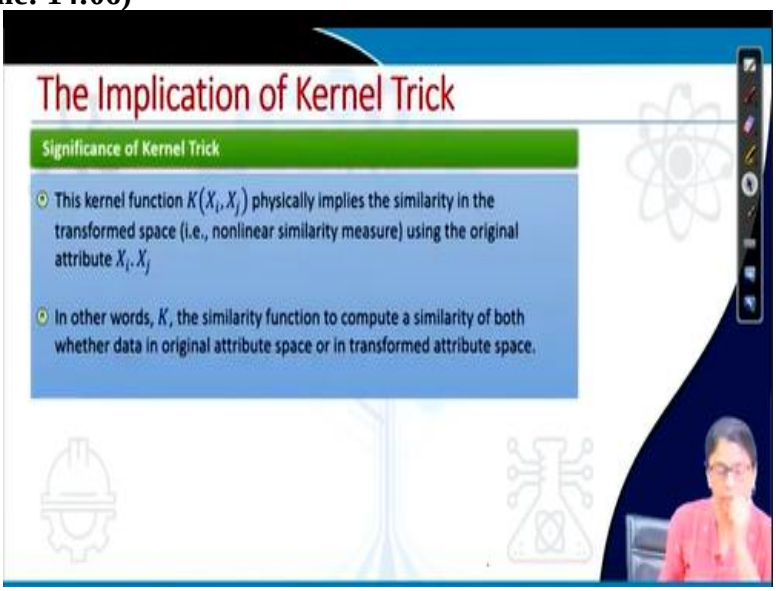
And now we do not literally have to go to the transform states without going I can use a kernel function where I can see the similarity of the transform space data taking the input of the input dimensional input space. So, this is the now what is this K function? K function is such a function we are taking the input space data I can tell that similarity of the transform space data. Now what is this K function. So, essentially, we need to find this K function. So, that is what that is where we have stopped in my last lecture.

(Refer Slide Time: 13:54)



So, now starting from that so, what we will see is that the significance of the kernel trick how this we could use this kernel tree and we will also see different properties and some examples of kernel function.

(Refer Slide Time: 14:06)



So, basically what is the significance of kernel trick which I have already mentioned let us read it out the kernel function K of X i . X j physically implies the similarity in the transform space. What does K X i X j thus? It implies the similarity in the transform space that is the non-linear similarity measure using the original attribute. What is using? It is using the original data original X i X j like in my last lecture I have given you an example two fruits in the Indian soil.

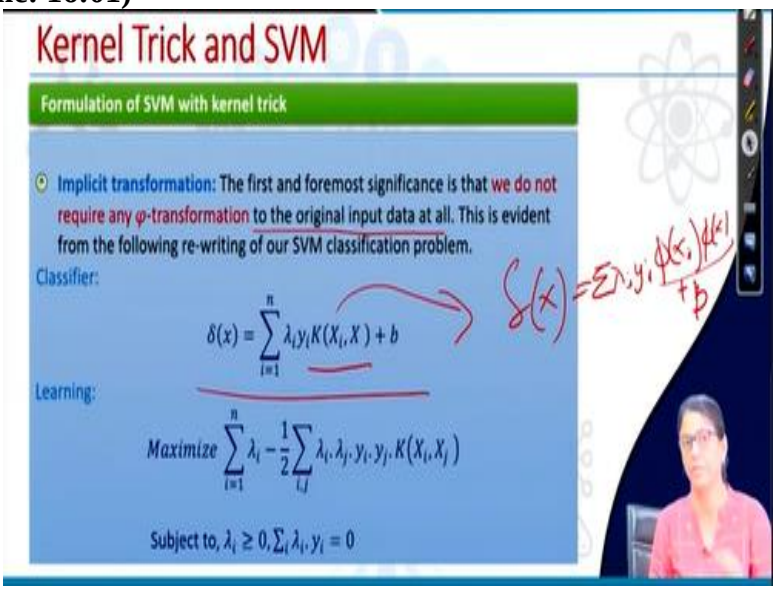
If I know the similarity between these two, similarly, if I have some function which can tell the similarity between two fruits and say UK soil, I do not need to go there I do not need to know what I do I can just tell the similarity between two fruits. If I have such function which can tell the similarity in the transform space there my transform space is a UK soil. So, if I can tell the similarity in the transforms is taking the using the original attribute using the original attribute.

So, if I can know that; then I do not need to transform it my requirement is that why I am transforming a data so that I can the non-linearity I can make it linear. So, and actually what I

am doing when I am trying to do the classification, I am basically trying to find out a similarity between two factors is not it what are this input inputs are nothing but the different vectors. So, when I am trying to classify whether it test data falls in this or this.

I am trying to find out how much less data is similar to which the different two types of data. So, I am just finding the similarity so that is it. In other words, K the similarity function to compute a similarity of both whether data in the original attribute space or in that transform attribute space.

(Refer Slide Time: 16:01)



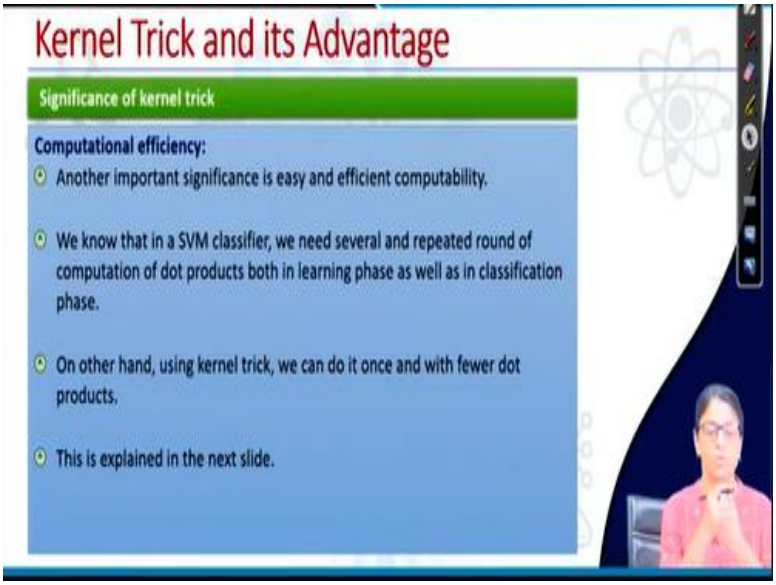
So, the first and foremost significance is that we do not require any φ transformation. We have such a function which can give the similarity and my transform space taking the data from the input space original input space. So, we do not require any φ transformation to the original input data at all. So, now what is my classifier? This is my classifier. Earlier classify λ i y i φ of X i φ of X is not it.

This I was writing as λ of X; the X is the test data λ i y i φ of x i φ of X + b so φ of X i and φ of X means transforms this. Now I do not need this transform what is that instead of this I can write K of X i, X the similarity between these two factor K I have X i and X and subjected so this is my classifier now. And what is the learning what else I need to maximize this we have already seen the dual form of the lagrangian.

Now in the last lecture we have seen. So, I need to maximize this subject to this so now my problem has become this if my data is linear fine if my data is non-linear so I will use this as my what to say optimization problem maximize this and subject to this. If when I will do this, I will be able to find out the different values of the lambdas different values of the basically I need to find out the values of λ only.

Once I find out the values here only λ is in the primal form, I had I required W I require to find out W I require to find out B in this form I only need to find out the values of λ. So, that is sufficient for me to find out the expression of the hyper plane. And once I know the value of Λ then I can use the classifier to find out the class of a test data.

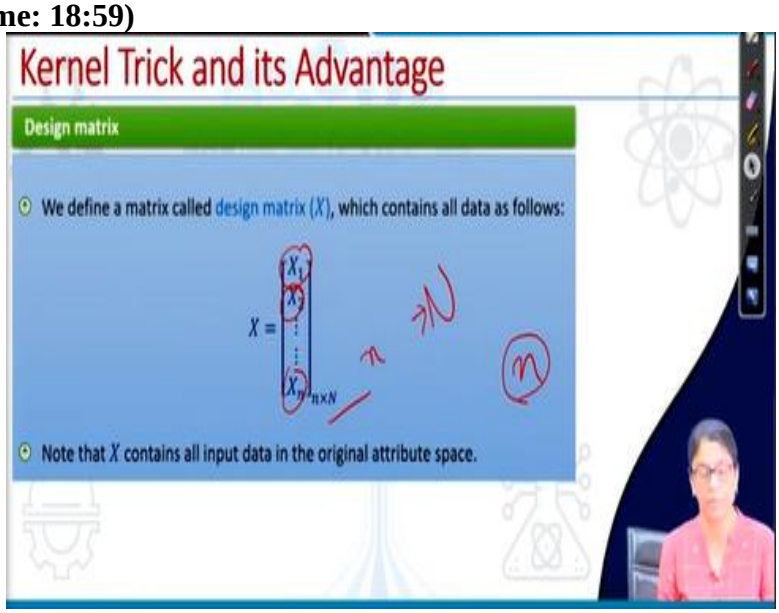
(Refer Slide Time: 18:09)



So, now what is the significance? One important significance it is easy and efficient computability. So, this first of all this is we do not have this already we have learned a last lecture we do not need W we do not need B just we need λ. We know that in SVM classifier we need several and repeated round of computation of dot product both in the learning phase as well as classical classification phase.

So, dot product X i . X X i . X j so, we need to do repeated round of this dot product both as in the learning phase as well as the and classification phase. But on the other hand, using the kernel trick we can do it once and with fewer dot product we will see it how.

(Refer Slide Time: 18:59)

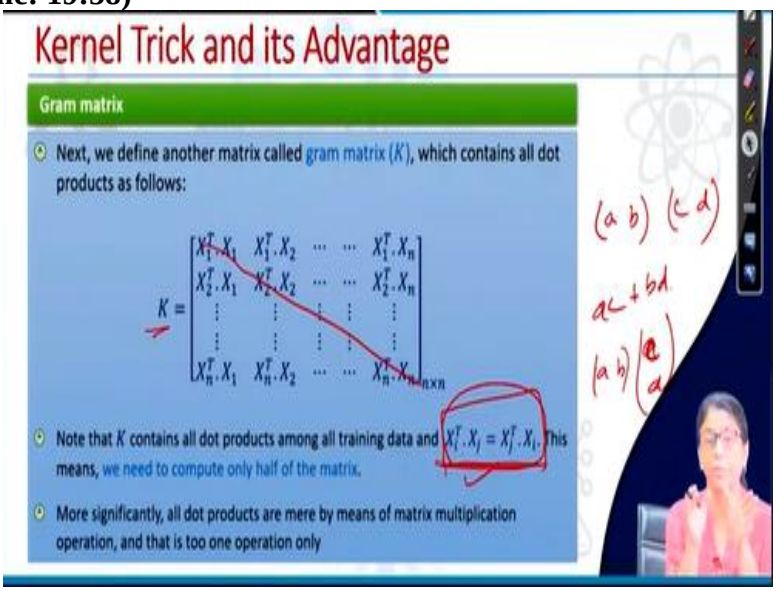


So, we define a matrix called we are calling it a design matrix which contains the data as follows. So, this is my design matrix where X 1 to X n these are the input data. So, total my total how many input data I have that means I can say it is the training data that is n training data or how many observation I have along with the class level that I call I may call it instead of trading data I can call it observation also along with the class level.

So, I have total n data and it is in how much dimensional say it is an N dimensional that means each X 1 is a vector of N dimensional. Each input data this is one input data this is one input data so total there are small n data and each data each dimension of data is capital N so; I am

defining a design matrix where my design matrix is small n cross begin. And so, where X contains all input data in the original input space.

(Refer Slide Time: 19:58)



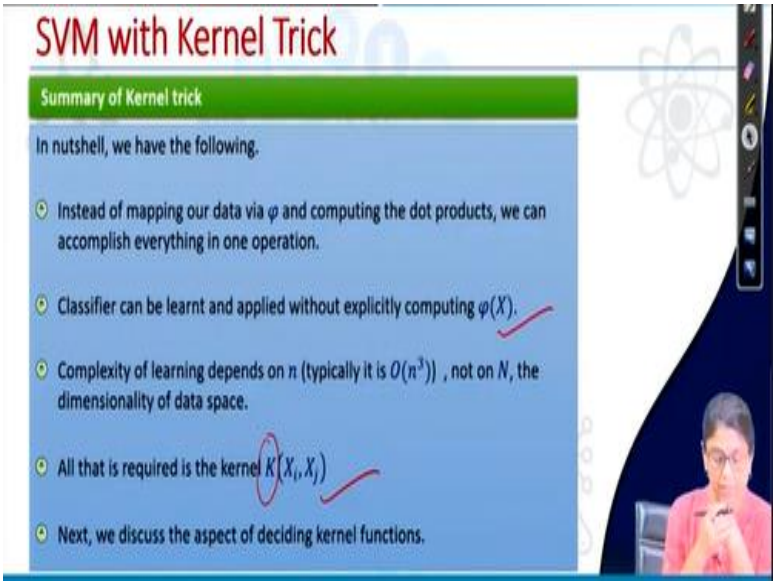
Now what I will do I will define another matrix which I call a gram matrix this basically this is K what K we are talking about which contains all dotproducts as follows. We are calculating all the dotproducts what is dot product X so, when they suppose these two vectors a b two vectors say c d how do we do the dotproducts what do we do the dot product you know this so, it is ac + bd is not it. So, now here if I write in this way, we get the same thing was not it.

So, that is all that is what we are trying we are finding out a dot product between all this input observation so this is how we are doing X 1 T X 1 X 1 T X 2 similar this one is all with X 1 then the second row with X 2 X 2 T in X 1 X 2 T X 2. So, the third row with X 3 the third input data that way we find all the dot product. Now if we see here the note that K contains all that purpose among all training data that means among all observation.

And note that this is definitely X i transpose into X i X j = X j transpose into X i this is obviously is not it, this is an obvious thing. So, when this is true that means, we do not need to compute the whole thing only if we compute half of it if you compute this, this value we will get because this is this, this is equal to this we are computing this and we are storing it. We are calling this as a gram matrix. This means we need to compute only half of the matrix.

More significantly all dotproducts are made by means of matrix multiplication operation and that is to one operation only. So, now this is how we have computed all the dotproducts and essentially, we try to find a similarity between these two and we have kept it.

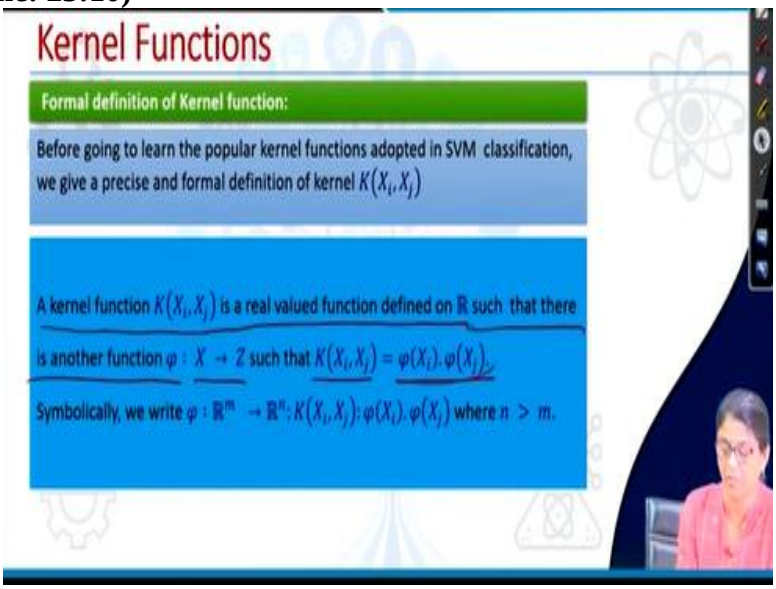
(Refer Slide Time: 22:13)



So, in nutshell we have the following, instead of mapping our data via φ and computing the dotproducts we can accomplish everything in one operation. Classifier can be learnt and applied without explicitly computing φ of X and here the complexity of learning depends on the number of input instance number of observation number of training data. It does not depends on the dimensionality because it is a simple matrix and **(()) (22:40).**

It does not depends on the dimensionality it just depends on how many input data we have. So, it depends on N typically order of n to the power 3 and all that is required is the kernel all that is required is that we need to do this the risk K of X i X j is it simple exercise is a simple dot product or whatever what we need to do we need to find this kernel function. So, next we discuss the aspect of deciding kernel function.

(Refer Slide Time: 23:10)



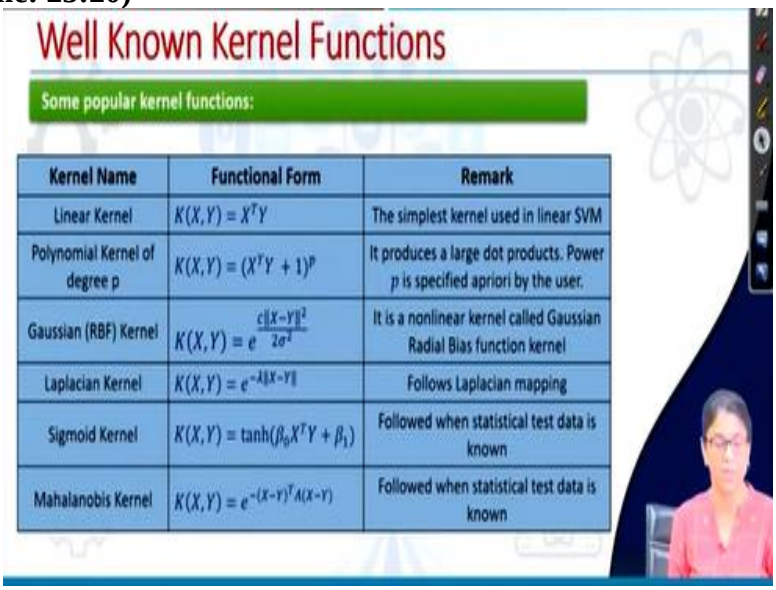
So, before going to learn the thing different type of kernel function let us formally define kernel function. This is the formal definition you see the formal definition of the kernel function. A kernel function K of X i, X j is a real valued function defined on R definitely it is a real valued function is not it defined on R such that such that there is another function φ. So, which such as φ of X gives Z such that K of X i X j = φ of X i . φ of X j.

So, this is the definition of kernel function if I formally define it. Please go to it again a kernel function K X i X j is a real valued function defined on R such that there is another function φ such that X gives Z φ of X gives that such that K of X i X j = φ of X i X j. K of X i X j physically implies that it gives a similarity in the transform stress that is it gives the similarity between φ of X i and φ of X j, what is φ of X i and φ of X j?

Φ of X i and φ of X j is the vectors in the transformation the input vector centre transform space. So, K of X i as this gives the similarity in this transform space how does it give the similarity for similarity it takes the input attributes in the input space it takes the vectors in the input space. So, symbolically we can write R of m, R m it is transforms transform to R n where n is greater than m.

Dimensionality n is dimensionality of m where here dimensionality is m it gets transformed to n dimensionality where n is greater than n and K of X i X j = φ of X i . φ of X j this is the formal definition of kernel function. So, now I think you can understand you understood what is kernel function and we have defined a formally same thing just a formal definition.

(Refer Slide Time: 25:20)



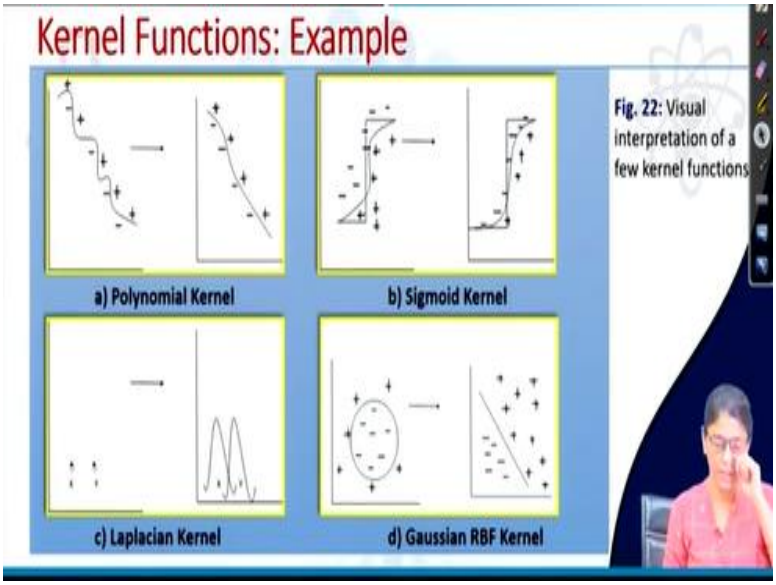
Kernel Name	Functional Form	Remark
Linear Kernel	$K(X,Y) = X^TY$	The simplest kernel used in linear SVM
Polynomial Kernel of degree p	$K(X,Y) = (X^TY + 1)^p$	It produces a large dot products. Power p is specified apriori by the user.
Gaussian (RBF) Kernel	$K(X,Y) = e^{-\frac{c\|X-Y\|^2}{2\sigma^2}}$	It is a nonlinear kernel called Gaussian Radial Bias function kernel
Laplacian Kernel	$K(X,Y) = e^{-\lambda\|X-Y\|}$	Follows Laplacian mapping
Sigmoid Kernel	$K(X,Y) = \tanh(\beta_0 X^TY + \beta_1)$	Followed when statistical test data is known
Mahalanobis Kernel	$K(X,Y) = e^{-(X-Y)^TA(X-Y)}$	Followed when statistical test data is known

So, these are some popular kernel function. So, what we have just shown this was this kernel function this is a popular kernel function. So, we can use any of the kernel function now in a given set of data we do not know which kernel function will work properly. So, it will just we will try each and every kernel function if we already notice that of this type and accordingly, we will select a kernel function which is suitable for that data.

But usually, it is very difficult to know data of which type we have our training data wafer observation but from the observation it is very difficult to make out it the data is of what type. Data is a linearly separable written a nonlinear separable or what type it follows which distribution nothing so, better find out try with all sort of kernel and then the kernel fish gives the best result that we will be using for our and for SVM of that particular data process.

So, these are the different I am not going to the details of this so, these are the different kernel functions this is a functional form and it is wave what is used it is given here. You can go the each of the for learning each of the kernel function if you can just Google it and you can learn each and every kernel function.

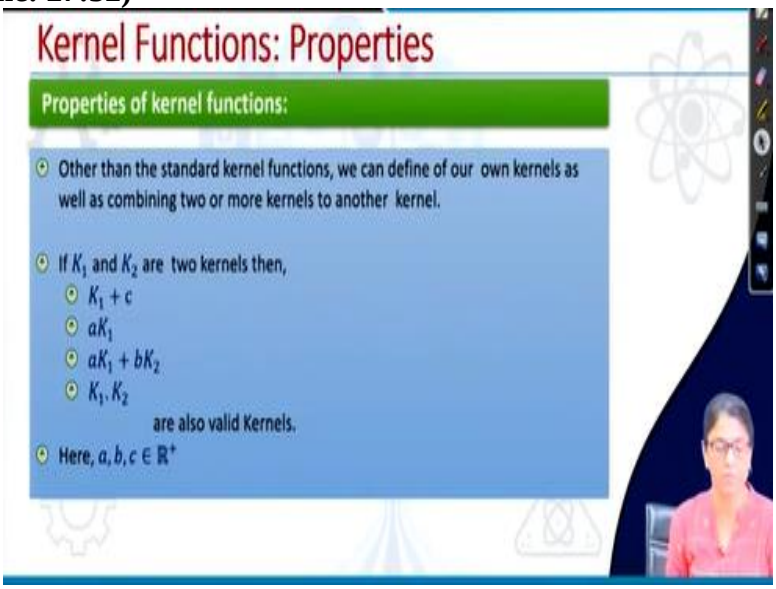
(Refer Slide Time: 26:35)



So, and like here is the inner figure I have just shown what is what effect we get if we use the polynomial kernel. So, this is the polynomial kernel this p value, p value is again we can you can take any value you can take p 1 to any value which will have to do it and see which value will give the better result. So, in polynomial kernel this was initially our data was in this form now in the transform space if we see we got the data in this now which can be easily linearly separable.

Similarly, if I use the sigmoid kernel if data is in this form my data got data has transformed to this space which is again linearly separable. Similarly, this this these are the different you can see this data if it is totally non-linear data if I use the Gaussian rbf 10 **(()) (27:21)** and radial based function kernel so it my data gets transformed this.

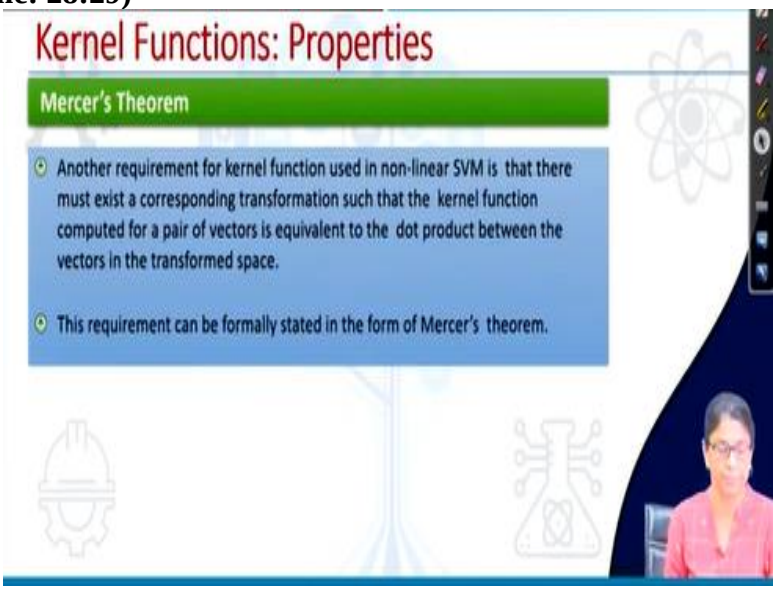
(Refer Slide Time: 27:32)



So, some properties of kernel function. So, other than the standard kernel function we can define of our own kernel as well as combining two or more kernels to another kernel. These are the standard kernel function we can define our own kernel function also. But our kernel function has to satisfy certain criteria just arbitrarily we cannot define it has to satisfy certain criteria and that way we can define our cone kernel function.

Moreover, given a kernel function by joining two or more kernel we can have another kernel we do not know which kernel will work best for a type for class of for a what to say category for a group of data. So, if K 1 and K 2 are kernel then this also can be a kernel this also this also different combination where a b c is nothing as just positive real numbers. These are all these are valid kernels we can have some of this kernel in this way by using different constant a b c which belongs to the real number and positive space.

(Refer Slide Time: 28:29)

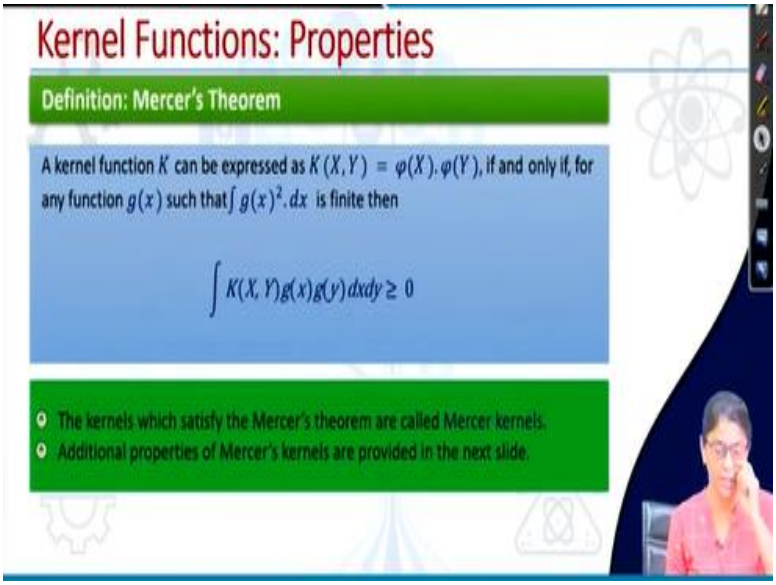


So, in kernel function used in non-linear SVM one of the requirements is that there must exist which I already talked basically. There must exist a corresponding transformation such that the kernel function computed for a pair of vectors is equivalent to the dot product between the vectors in the transform space. This is the main requirement of the kernel function for classification purpose what we need for classification purpose.

What is the main requirement of kernel function is that there must exist so we have what we have seen and we are repeating it again and again there must exist a corresponding transformation such that the kernel function completed for a pair of vectors in the non-transform space will give us the similarity in the transform space. Such that a kernel function computed for a pair of vectors is equivalent mark this term is equivalent to the dot product between the vectors in the transform space.

This requirement can be formally stated in the form of Mercer's theorem we have a theorem called Mercer's theorem.

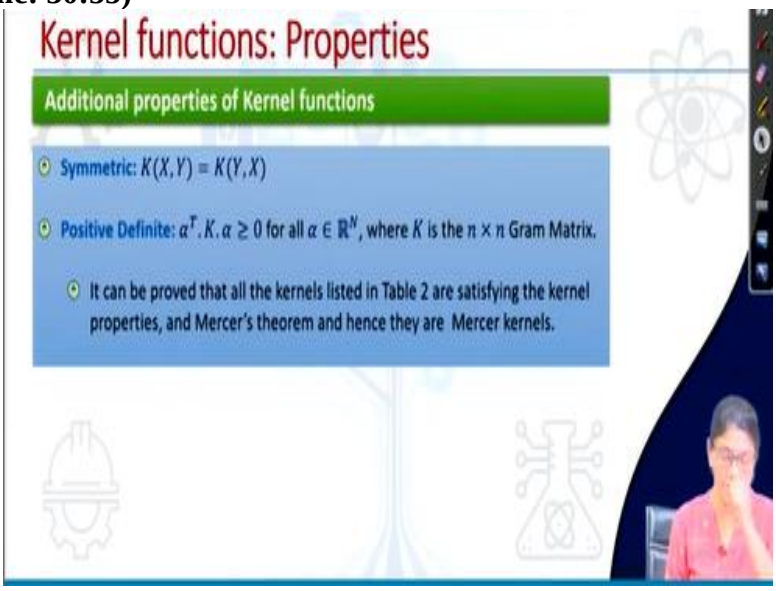
(Refer Slide Time: 29:53)



So, what does martial theorem says if kernel function K can be expressed as this if and only for any function g x any function g x you can normally is for any function g x such that integration of g x ^{2} dx is finite you can select any function g x but that it should satisfy this should be finite and this should be greater equals to zero. Then we as here we have seen that we can define our own kernels we can very well define our own kernel.

But our kernel function should satisfy this it should satisfy the Mercer's theorem what is this Mercers' theorem this is the Mercer's theorem it should satisfy this. The kernel's theorem which satisfy the Mercer's theorems are called Mercer kernel.

(Refer Slide Time: 30:55)



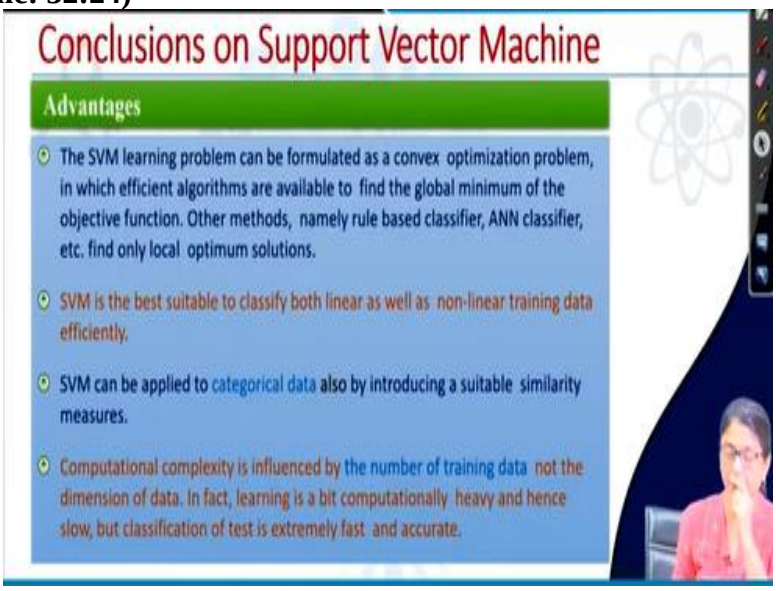
There are some additional properties which will be seen here the kernel one thing is our kernel function should be symmetric that is the reason why we have computed half of the matrix within compute the other half. If you compute the half, we can get the other half. So, in the example what we have seen. So, that is one kernel function simple dot product so it has to be symmetry and it should be positive definite.

What does positive definite means? A to the power T K . α should be greater equals to 0 where α is this so, α ^{T} . K . α is greater equals to 0 where α belongs to this where K is a n cross n gram matrix. It can be proved that all the kernels that we have listed in table 2 this is the table two

this is the table two oh sorry I did not mark it here table. So, anyway this is the table two are satisfying the kernel properties.

And Mercer's theorem enhance their Mercer's kernel all the kernel function that we have mentioned all those are called Mercer's kernel because they are satisfying the kernel properties water is properties these two properties and as well as the Mercer's theorem.

(Refer Slide Time: 32:24)



So, now in coming to the end of it so what is the advantage of support vector machine. The support vector machine learning problem can be formulated as a convex optimization problem. We have seen it can be formulated as a convex optimization problem which we have solved it using then like then Lagrangian multiplier in which efficient algorithms are available. One of the efficient algorithm is like the Lagrangian multiplier to find the global minimum of the objective function.

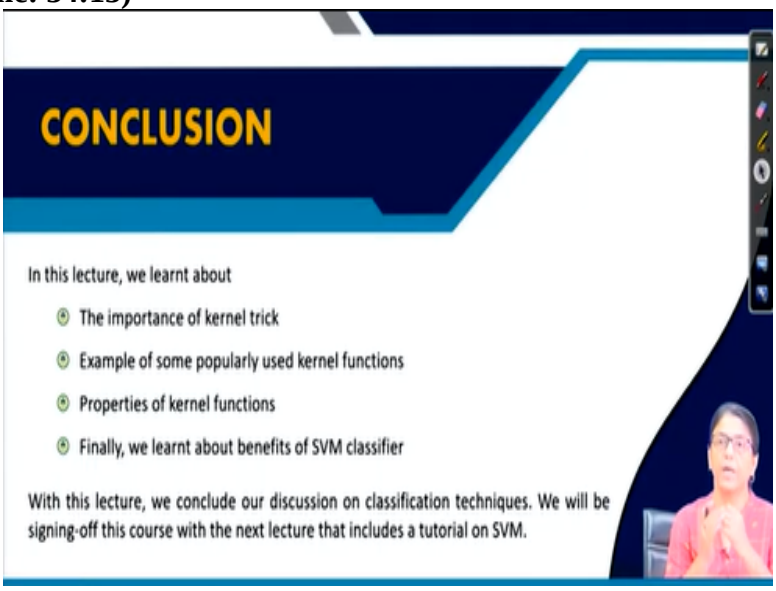
Other methods namely rule based classifier ANN classify etcetera finally local optimum solution. This things ANN classifier rule based classifier it finds optimal solution when we try to find out the classified machine SVM that function. So, it using this convex optimization problems and what we get we get the global optimum. Then SVM is best suitable for classify both linear as well as non-linear training data efficiently.

Linear training data we have already seen non-linear trending determines using this kernel function it does it efficiently. SVM can be applied to categorical data also by introducing a suitable to similarity measures because category data we do not have those as such we do not have a value. So, we can introduce a suitability measures and accordingly it can be used for classification. Computational complexity is influenced by the number of training data we have already seen.

Computation complexities very much depend on the input space it does not depend on the dimension of the space not the dimension of the data. In fact, learning is a bit computationally heavy because computational heavy means we do not know which kernel function will be best suitable. So, and during the learning function we may have to try different learning function and if we have to use the polynomial kernel.

Then we have to use maybe we have to use on different values of p and find out which is best suitable and hence slow. So, learning is based a bit slow but once learning is done classification of text is extremely fast and accurate. So, these are the advantages of SVM.

(Refer Slide Time: 34:13)



With this we come to the end of this lecture today. We have learned about the importance of the kernel trick example of some popularly used kernel function we have also seen the properties of kernel function and finally we have also mentioned the benefits of the SVM classifier. So, with this as I told this is the last lecture of this course means from the discussion on point of view and will be signing up this course with a tutorial class and my next lecture.

(Refer Slide Time: 34:44)

Thank you, guys thank, you.

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

Lecture - 60
Tutorial on SVM

(refer time: 00:35)

Hi guys. So, the last tutorial of this class last tutorial as well as the last lecture. So, now as usual in all my tutorials the way I have covered. The first will be solving few objective type equations and then we will go on to solve few problems. Like here but one thing is that for solving problems for SVM classifier it is very difficult to get some toy problems for that. Because here solving some problems means we will be; needing some small problem.

That is way we can find out the optimization function and what to say from the optimization function we need to compute the MMH and if we need to do transformation also so it is very difficult to do using pen and paper and assets form, formulates such a toy example it is really difficult. So, the problems what I have kept here is just one simple toy example I have kept for classifications.

But the other problems just how to use the Lagrangian multiplier from how to use, how to solve the optimization problem using the language that is the convex optimization following. Using Lagrangian multiply I have just given some problems on that because that experience will also help.
(refer time: 01:44)

And now first coming to the objective type questions. So, these are also very easy ones. So, first is support vector machine is based on what? It is based unsupervised learning algorithm; it is based on the unsupervised learning algorithm, regression analysis, logistic analysis. So, support vector machine is not based on the regression or logistic and logistic regression and it is a classification supervised SVM we have used for classification.

So, when we have used first classification, we have used our training data has the class level. So, definitely when our training data is a class level so answer to this question is the supervised

learning algorithm.
(refer time: 02:26)

Then the vectors which are lying on the decision boundary that basically gives the complexity of the SVM also. What are those vectors called? Those vectors are called training vectors, those vectors are called test vectors. What are training factors? Training factors are which those vector which you use to train our SVM that is called training vectors. What is a test vector? Test vector test data which you want to test that is called the test data test back to this and the support vector training vector due to noise out of question.

So, this is the lying on the distant boundary this we call it a support vector and support vector also gives the complexity of the SVM as well.
(refer time: 03:08)

So, support vector machine can be used for this question support vector machine this we have seen that we have used support vector machine for classification type. But however, support vector machine can also be used for regression and also be used for ranking. So, we have not seen those two, this is a bit advanced topic. If you are interested you can just learn it by yourself. Once you have the idea the thing is that basically in SVM also the amount of lecture I have covered in SVM.

That is actually if you consider the SVM literature such a huge literature that is for just covering SVM literature. If I try to cover go into the depth of SVM go into every integrity of SVM it will be taking at least say 15, 20 lectures which is really not feasible. Here the objective behind is that once you know what is once you can understand the concept once you know what it is then you yourself can do some self-study and you yourself can learn those things it becomes very easier.

Once you know the direction know the concept then you can go ahead. So, that is a; be that logic actually have what to say discuss the SVM. I did not go into depth of the SVM so now similarly this SVM machine which we have used for only classification. But however, it is it can be also used for regression analysis and ranking tasks. So, the answer is this.
(refer time: 04:29)

The goal of SVM is to find the optimal separating hyperplane such that so what is the goal of SVM? The goal of SVM is to find out the maximum margin hyperplane, is not it? That is our goal. So, what it is given? It maximizes the margin of the hyperplane, it minimizes the margin of the hyperplane that is definitely not, it minimizes training error maximizes test error. When we are from trying to from our classification and always in any tasks our goal is to minimize training

error minimize test error both the time.

We do not need more test error and less training error or the vice versa, we need optimal of both. So, this is also definitely out of position so it maximizes the margin of the hyperplane, this is our goal.
(refer time: 05:17)

It is not here consider the following data set. We are performing a binary classification problem using linear SVM classify. The points are circled red that are representing support vectors. So, these are the support vector this is one support vector, this is one support vector, this is one support vector. If I remove any one red points from this data if I remove any one red point if I remove this or if I remove this or if I remove this.

If I remove any one of these data does the decision boundary will change? Yes. The equation here the answer to this is yes, because the support vectors decide the decision boundary. It is the support vectors which decide the distance boundary. So, if the support vector if I play around with a support vector definitely my decision boundary will also change. So, does the decision boundary will change if you remove any one point? Answer is yes.
(refer time: 06:12)

Next question is the same question. If you remove any non-red circle points if I remove any of this point, I do not reform the support vector. It may remove any of these points does the decision boundary will change? No, the decision boundary will not change because the decision boundary is very much dependent on the support vectors.
(refer time: 06:28)

What is true about the kernel in SVM? Kernel function map low dimensional data to higher dimensional space this is obviously does that, is not it? Kernel function map is that it implies the similarity in the higher in the transform space taking the data from the input space, is not it? So, it is a similarity function so both are correct. So, only one is true, both one and two are true this is the answer.
(refer time: 07:02)

Which is not used as a kernel function in support vector machine? Sigmoid kernel is used we have seen that, radial bias RBF kernel also is we have seen that, the table in the last lecture I have given some lists of some kernel functions where this is the polynomial kernel also, we have seen where we have the speed to the power p, p is a user defined. We can use any non-linear polynomial definitely it has to satisfy certain constant and only we can say is a kernel

function. So, this is the answer.
(refer time: 07:35)

What is the purpose of the kernel trick? What is the purpose? Now you know what is the purpose of the kernel treatment. Let us see what are the options given. To transform the problem from regression to classification is it the purpose definitely not. To transform the data from non-linearly separable to linearly separable, yes of course. To transform this problem from supervised to unsupervised, definitely not none of the above definitely not B is the answer.
(refer time: 08:03)

Which one of the following is a correct equation for a linear separator that is a hyperplane all symbols bear the usual meaning. Correct equation of a linear separator, which one? Definitely this one is not a linear separator less than equals to zero, in the hyperplane should be equals to zero. So, this is the one this. If I take decide and it will be equals to zero so in a hyper plane it will be when it is when it lies on the hyperplane and the point will lie on the hyper plane.

The value will be equal to zero. But here on the above the hyper plane will get a positive value below the hyper plane will get a negative value. So, this is the equation.
(refer time: 08:46)

Now some problems, as I told you these problems are not classification problem see.
(refer time: 08:55)

Suppose a company produced two goods x and y. As per company's policy total 42 goods need to be produced each month. The total production cost of the company follows a cost function so this is the cost function of the production cost. So, it produces two types of put x and y and this is the cost function. The owner wants to minimize the production cost while maintaining the constraint that the total number product in each month is $x + y = 42$.

Find a solution for x y using the Lagrangian multiply method. Even if you have mentioned Lagrangian we could we know we can very well use this Lagrange multiplier method because this is the optimization problem where we need to minimize this subject to this constraint. So, which one will be will it is a convex automation problem or with equality constraint? This is the equality constraint we have $x + y = 42$.

So, equality constraint means we will have will find out the Lagrangian, once we find out the Lagrangian and Lagrangian will do the differentiation of Lagrangian within each input attributes.

So, in our support vector we have seen we have used the Lagrangian we have differentiated a Lagrangian multiplier with the different attributes. The different attributes we have differentiated with different x i.

If there are n dimensional so δ (L) / δ (1 x 1), δ (L)/ δ (x 2), δ (L) / δ (x 3) likewise δ (L) / δ(x n) then also we have differentiated with respect to B δ (L) / δ (B), is not it? That is how we have found out the different equations. And then once we found out a different expression then we found out the values of the W and b. And of course, we have also differentiated and δ (L) with respect to the different dummy parameters this λ also, δ (L / λ).

How many dummy parameters? If there are two dummy parameters δ (L) / δ (λ 1) δ (L) / δ (λ 2). The number of dummy parameters a day will have to differentiate with respect to those also. So, the Lagrangian will differentiate the lagrangian and with respect to all of the input parameters as well as the all of the dummy parameters and then we have a couple of equations.

We will solve these equations to find the value. We are keeping in the mind constraint that we need to satisfy. So, that was equality constraint problem.
(refer time: 11:21)

So, here also it is a equality constraint problem. So, like we have to minimize this subject to this. So, how we can form the Lagrangian? This is how we can form the Lagrangian. So, this is f x + λ of the constraint so my constraint is this x + y. So, this I can write it in this form, is not it? 42 - x - y = 0 I want to equalize it to 0 so this is my what to say constraint. Constraint = 0 equality constraint so this is here.

So, I will do the differentiation of L with respect to x, differentiation of L with respect to y, differentiation of L with respect to λ.
(refer time: 12:05)

So, with respect to x I got this, with respect to y I got this with respect to λ I got this. Then I solved this tree expression and I found this is the value for x. This is this very simple equation so I could easily solve it but for a bit complex equation I will have to use some numerical techniques method. So, I got this x value this y value this and λ value this. So, the company must produce 25 of the first group and you see this satisfies the constraint also x + y = 42.

It is satisfying the constraint as well. So, the company must produce 25 of the first good and 17 of the second good to minimize the cost maintaining the constraint.

(refer time: 12:46)

Again, another one problem of optimization, solve the following optimization problem. Maximize this subject to this. Remember our convex optimization problem, we use for minimizing a particular function subject to inequality constraints where are inequality constrained is less than equals to. If you can remember our inequality constraint was less than equals to and we are of minimizing the optimization function.

So, if it is greater than equal to then what we will do? We will be in our Lagrangian will be subtracting it. So, this maximization problem we can make it minimization easily we can do that. So, our Lagrangian is this.
(refer time: 13:29)

I have changed it to minimization because this was maximization and this since it was greater than so I have used minus. Instead of plus you put it less than I would have used plus of λ 1 into these plus of λ 2 into this. So, this is my Lagrangian and in the Lagrangian I have to have some KKT constraint remember for what to say convex optimization problem you for inequality constraint so I have to define the Lagrangian and then it should satisfy certain KKT constraints.

So, what are the KKT constraint? KKT constraint is that δ (L) / δ x 1 δ (L) / δ (x 2) whatever I get that those two is equals to 0 and I do not differentiate with λ 1 and λ 2. I do not differentiate L with the dummy parameters for the inequality constant rather I have some other con constraint KKT constraint if you can remember KKT constraint. So, what are the KKT constraint?
(refer time: 14:26)

So, first I have been δ (L) / δ (x 1), δ (L) / δ x 2 these are the two KKT constraint. Now distinct this I have just kept it to just to remain remind you that this is not required for inequality constraint. For equality constraint only we differentiate with the dummy parameter. So, for inequality we do not differentiate this, this is not required. Then another what are the other KKT constraint?

Other KKT constraints are the constraints inequality constraints what is already there. So, these are the inequality constraint λ 1, λ all my λ 1, λ 2 should be greater equals to zero λ i should be greater equals to zero, this is one KKT constraint and then whatever the constraint specified. Then another one is λ i x i = 0 this is one KKT constraint. Λ i x i = 0 and then x i = 0 and then this is λ i greater equals to zero.

These are the different KKT constraint, is not it? So, we have this is one KKT constraint x 1 + x 2 less than equals to four and my λ 1 greater equals to zero I can write it this way, this is equals to 0, this into this. And similarly, it is x 1 + 3 x 2 - 9 this is less than equals to 0 and this λ to greater equals to 0. I can write as this this is one KKT constraint this is one KKT, this is one KKT constraint and these are the two KKT constraint.

We have the different KKT constraint. Now using this KKT constraint we will be solving the value finding out the value for x 1 and x 2 taking different combination of λ 1 and λ 2 value. We have two dummy parameters that is λ 1 and λ 2 and it specified that λ i should not be greater it should be greater or equal to zero, it cannot be less than zero.
(refer time: 16:25)

So, I will take different combination, I will take λ 1 λ to 0, λ 1 greater than 0, λ 2 = 0, λ 1 = 0, λ 2 greater than zero, λ 1 greater than zero, λ 2 greater than zero for all these combinations there will be four such combination. For so four such combination I will be using these four combinations and will try to find out the values.
(refer time: 16:49)

So, basically these are my obtained constraints whatever I have in the last slide I have written it here. Now with different case this already we have solved it, you can just go through different steps I do not need to explain this. We will take for each case we will take separately. First case is λ 1 λ 2 = 0. If λ 1 and λ 2 = 0 what does this expression becomes? This expression become - 2 of x 1 - 4 = 0, this expression become - 2 of x^2 - 4 = 0.

And λ 1 = 0 λ 2 = 0 then what happen? If λ 1 = 0 λ 2 = 0 then we will be using this x 1 + x 2 and this. Suppose λ 1 is greater than 0, if λ 1 is greater than 0 then I can write x 1 + x 2 - 4 = 0 then I can use this also as the equation. Again, if λ 2 is greater than 0 then this will be equals to 0, x 1 + 3 x 2 - 9 = 0. So, that way I will be having all the equation then I can find out all the solutions and once finding out for there are total four cases as I told you.

For the four cases you will be finding different solution. Now from among this solution you will have to find out which is the feasible solution. if there is one feasible solution then that is the only solution. If there is more than one for feasible solution then among the feasible solution you will have to find out which is the optimized solution and accordingly that is the solution.
(refer time: 18:18)

So, you can just go through the steps in fact please do not go through these steps you do it yourself. And once you get the answer then you try to match it with this.

(refer time: 18:29)

(refer time: 18:31)

So, this is how decided I am not going to all the steps you can do it on its own. So, we got the finally we got the solution x 1 and x 2 = 2 and it satisfies the constraint also. What is the constraint x + x 2 greater equals to 4 that is it is satisfying the constraint as well.
(refer time: 18:47)

Now here one is a small example of a classification problem. As I told you it is a toy answer this is not at all a realistic because for classification for developing an SVM we need more and more observation. We need more and more training data with a class level then only we can come up with a proper maximum margin hyperplane. So, it is just a toy example just to see just to utilize the concept. So, suppose there are three points, so what are my points?

This is one point, this is another point, this is my third point that means it is that each point is a two dimensional it is in a two dimensional space. Here P 1 and P 2 are the negative class this two belongs to negative class and P 3 of positive class two class negative and one positive. Compute the decision boundary as learned by the SVM classifier.
(refer time: 19:36)

So, first what will I do? So, this I will consider this is my decision boundary W x + b = 0. So, this is in two dimensional so I will be having two parameters w 1 and w 2. So, w 1 x 1 + w 2 x 2 + b = 0. So, we need to compute the values w 1, w 2 and b. Find this SVM learning problem is given by minimize? We will need to minimize this subject to this for all i. So, first we will define the Lagrangian that so my Lagrangian is this, minimize this.

So, it is greater then I will since my convex optimization problem it is less than equals to zero. Since; I have >= 0 so I will be using minus. So, this is how I have defined my Lagrangian. So, I have defined my Lagrangian now I will have to find out a different KKT constraint.
(refer time: 20:33)

So, these are the data points I have given here in the figure x 1, x 2, x 3 these are a different data point and other this is a class level which is already given class level and these are the λ value λ 1, λ 2, λ 3 we will have to find out this these three values now. So, now the KKT constraint this is one KKT constraint δ L / δ w. So, then this already we know λ i and y i should be equals to zero.

So, in our case n = 3 so we will first we have to find out what is w, w it says we know this is the w this is the formula for w. so putting this y i - and y i I will take value at a minus or plus then similarly we found out the values for w. And this is another one expression λ i w i = 0 means λ greater equals to zero. So, λ greater equals to zero minus means this is our λ this is equals to zero.

So, we will get this so these are the two equations. Simple putting it and putting it in the equation and trying to solve it that is all.
(refer time: 21:41)

Now basically if I dual problem is maximizing this, is not it? Subject to this greater equals to zero. Now we will put all this in this dual problem then we find L is equals to this just simply putting this here. Λ y i λ j y j x i x j and whatever data we have just put in the data, simple dot product. Excuse me if you do it you can simply do the dot product and see you will get this and simplifications and using this equation 1, we have this λ 1 + λ 2 = λ 3.

Using this we will get this is my Lagrangian. So, now I founded the constraint λ L by what to say differentiation of L / δ λ 1 / δ of λ 2 we found these different values. So, solving we got the λ value so λ 1 we got 0, λ 2 1 / 8 and λ 2 = λ 3 both is 1 / 8. So, I got a value λ 1, λ 2 and λ 3. If any values if the λ value is 0 that means that is not a support vector. So, my first point is not a support vector, the second and third point is my support vector.
(refer time: 23:04)

So, here this is not a support vector because λ 1 value is zero for value is zero means it does not lie in the parallel line, is not it? So, since it is zero so it is not a support vector these two are the support vector. But actually, this is a bit trivial see here, this is let us come to that later. So, now here we got this support vector λ 1, λ 2, λ 3 value we got.
(refer time: 23:30)

Once we know this then from here, we can find out the W values as well. We know λ 1, λ 2, λ 3 values we can put it here and we can find out the w value so we got w 1, w 2. So, we need to find out w 1, w 2 two values we will put it here and we found out the values. Once we found out the W values, we can find out the b values also. Once we find out the λ values w value, b value so we can find out the equation of the hyper plane.
(refer time: 23:54)

Now my equation of the hyper plane turns out to be x 2 = 0, very trivial example. Just this example is to show just to practice whatever you have just learned very simple example. Now here this x 2 = 0 means my hyperplane is this is my hyperplane and this are this point and this point I got support vector. See if this is my hyper plane this point, I got this is at a distance of two I got a this is my support vector this is a distance of 2 I got a support vector.

Since this is the x axis which is my hyper plane and that means distance of two my these are the two parallel lines. So, these are my distant boundaries, is not it? Now since again has surprisingly this I got as a support vector and this point if this point also lies in this side did not get as a support vector, why? So, this is a bit contrary what we have learned, is not it? Because if this is a support vector; definitely this also should be a support vector.

But actually, considering the case that this is just we are walking with three data's in considering three data's this. If considering this as a support vector a sufficient what to say it is actually it makes it sufficient for this data. If I just consider this as a support data this is no longer it needs to be considered as a support vector. In fact, when we see that some such a case whenever the line is parallel to the x axis the hyperplane is parallel hyperplane is on the x-axis or hyperplane is on the y axis.

So, then what happens? The decision boundary is just parallel to this two lines. So, in that case among there will be there may be many data which falls on the line parallel to that but all those data is not necessary to be considered as a support vectors. Just considering few of those as the support vectors what will help us in solving the problem? Because the main reason behind is that so it is very what to say first of all the reason behind is that since we are working in a very small example.

Very small and working in a small data we will definitely not get a proper SVM. So, to move the data we train to move the better and SVM we get. So, once we get a better SVM stand like which will be in a support vector which is not in the support vector then it is easier to find out. Now it is very toy example so it makes no sense just to find see how it works basically that is why we have kept this example.
(refer time: 26:36)

So, now another one, consider following two dimensional two dimensional data point for a binary classification problem. A non-linear transformation is used to transform the input vector. We have we are using this transform to infer to a transform space, this is the transform space. So, what is my transform function? Φ 1 of x 1 x 2 is this function and φ 2 of x 1 x 2 is this function. This is how we are transforming.

This one this whole thing we are transforming to φ 1 of x 1 x 2 to a different dimension and this whole thing we are transforming using φ 2 x 2 x 2 to a different function. Transform the data to the transform says is there any specific observation.
(refer time: 27:22)

So, this is my original data x 1, x 2. Suppose this original data is given and the class level are also given, the original data is given and the class level is written. Now from this original data I tried finding out my Z values z 1 and z 2. What is my z 1? This is z 1 and this is z 2, I tried finding my z 1 and z 2 value. Now I will try to these are also two attributes these are also two this is also two dimensional this is also two dimension.
(refer time: 27:51)

I tried plotting this data. See when I plotted this data so these are initially these are my data's. When I tried plotting this data if you see if you know you will get this sort of there is from this figure. We can see from this figure edition they are not linearly separable, is not it? Now once I have transformed it when I transform it so these are my values. Now in my transform space you see it is very much linearly separable. These are the data usually this set us triangle.

Now here in the transform space all into comes to one side, all the square comes to the other side is very much it has become very much linearly separable. So, we can see that the data points in our original input says on the left is not linearly separable. But when the data points are projected from input space to the transform space, they become linearly separable.
(refer time: 28:53)

So, with this I end this lecture I hope you have enjoyed this course. So, with that I had like to tell thank you to all of you and then conclude this course, thank you.

THIS BOOK IS NOT FOR SALE

NOR COMMERCIAL USE

(044) 2257 5905/08

nptel.ac.in

swayam.gov.in