Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 16
MLE

Very good morning, good afternoon, good evening, my dear students; welcome to this DADM, which is Decision Analysis and Data Analysis and Decision Making. One another NPTEL MOOC series and we are on the 16th lecture. This is a 30 hours lecture, which means it will be spread over 12 weeks with 60 classes each being of half an hour each week as you know we have 5 lectures.

So, we have we have just started 16th lecture means we are starting in the 4th week. Now if you remember we were discussing about I will just repeat to go back a little bit more in the past, we discussed the three distributions, the t distribution. Initially we did definitely discuss about the z, the t distribution the chi square and the f distribution. And I did mention that I will be coming back to that time and again that the z and t would be utilized to find out something related to the mean. And the chi and f would be something to relate it to the standard deviation of the variance.

Now, we also saw in the last part of the 15th lecture, that dependent the concept of unbiasedness and consistency were important we are not going to go into the proofs. And based on that we also gave examples that consider for an exponential distribution with a value of zero, you are picking up a sample. Then the sample mean is the best estimate of the parameter of the exponential distribution.

Consider for the Poisson distribution, again you are picking out observations from the sample of which consist a sample from the population you are picking them up. And you know the sample readings then the best estimate for the Poisson parameter is the sample mean. Similarly, you have the concept that for the normal distribution the sample mean is the best estimate for the population mean. And we also saw that the and I did mention that time and again then the standard deviation or the standard error to be right standard error or the square of the standard error for the sample would be the best estimate for the population variance of the standard deviation.

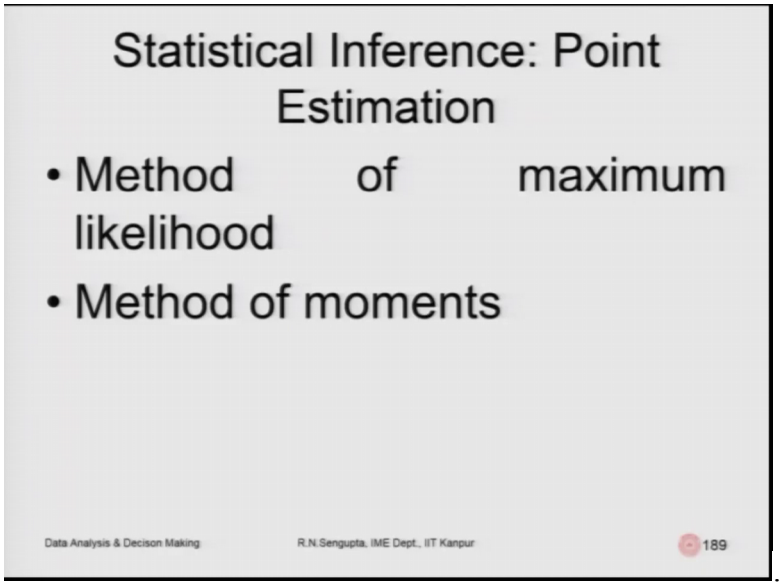
Now, remember that I mentioned about the concept of s dash and s without a dash and in one case you had that whole factor divided by 1 minus 1. Where the population mean mu was not known hence, it was replaced by the sample mean. And in another case, it was divided by n because the population mean was known, hence we did you did not lose any degrees of freedom in order to basically find out the sample mean because the population mean was known.

Now, I will proceed further with the concept of in this statistical inference problem. gGo through the basics for point estimation, interval estimation, hypothesis testing and then further on consider for the multivariate analysis and so on and so forth.

Now in statistical inferences you have three areas which is basically point estimation, when you find out a point of value like the sample mean is the best estimate of the population mean or say for example, s dash or s without the dash is the best estimate for the standard deviation for the population giving some conditions.

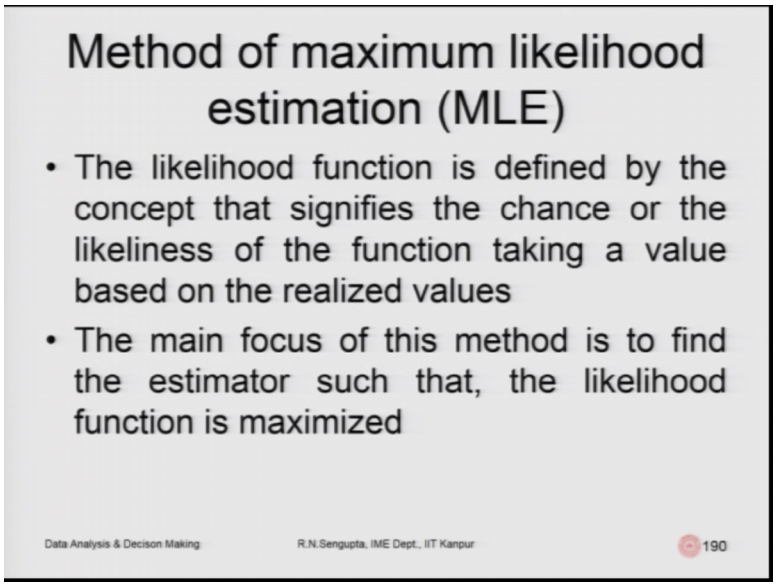
So, whatever it is we want to find out a particular value and check that the concept of unbiasedness and consistent inconsistency are met by that point estimate. So, there are different ways how you do that.

(Refer Slide Time: 03:50)



Two of the methods are quite popularly used on the method of maximum likelihood and the method of moments. I will very briefly discuss them without going to the proof I will basically come to the results.

(Refer Slide Time: 04:04)



Now, in the method in the method of maximum likelihood estimation M L E method, what we do is that? We want to find out the values of the parameters which are unknown, based on the fact that the sample which we pick up with the set of observation x 1 to x n, are such that they give us the maximum chance of finding out the highest probability such that the corresponding parameter values basically substantiates that. I will come to that concept and a little more detail through qualitative as we talk without in going to the quantitative concepts.

So, in the MLE method the likelihood function; so, we would basically define a likelihood function is defined by the concept, that signifies the chance or the likeness likeliness or likelihoodness of the function taking a value based on the realized value. So, consider the realized values are basically x 1 to x n.

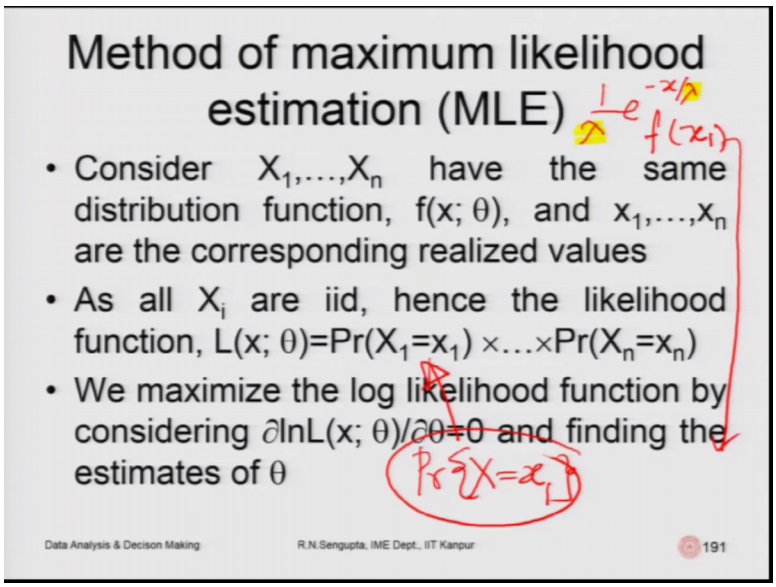
So, considering that they are if x 1 to x n; obviously, the corresponding to the first reading which is x 1 you will have a corresponding probability for the from corresponding from the PMF of the PDF or whatever you are going to consider for which the parameter value theta is unknown.

Similarly, when x in the second set when you pick up the observation and it is x 2 small x 2 realized value. Then the corresponding P M F value would basically be given by the value where you put on this x as is equal to x 2, but considering the parameter theta is unknown and then we find out that probability given parameter value is unknown. Similarly we do it through x 3 x four x 5 and so on and so forth.

So, what we want to do is that we want to basically and there are IID's remember that. So, if there are IID's, what we want to find out? If we want to find out the probabilities being multiplied such that the gives of the maximum value. And based on the fact they give the maximum value would basically like to differentiate that functional form, which is an unknown in theta with respect to theta put it to 0 and find out the best estimate of theta for which that case would be true.

The main focus of this method is to find the estimators such that the likelihood function is maximized. So, log likelihood it is generally mentioned because we convert the likelihood function into the log likelihood function with a logarithmic. And hence it is easier for us to differentiate because, logarithmic is here increasing function. Then hence trying to find out at what value of theta that log like log likelihood value takes the maximum one is easy for us to find out and we can basically comment in gently what is the value of theta which is basically theta hat. And that characteristics, is from the sample which basically gives us the best estimate of the population parameter theta.

(Refer Slide Time: 06:55)



So, consider we take in a set of absorption x 1 to x n. And we have the distribution function consider the PDF or the PMF whatever it is. There is given by f of x comma theta so; that means, if it is say for example. So, say for example, you have I am using the red color so, consider you have the exponential 1. So, in this case theta and lambda this is unknown.

So, say for example, when x takes the x x value x 1 x of x 1, then you put that x as x suffix 1 and find out what is the corresponding PDF or the PMF based on the fact that lambda is an unknown. So, that would basically be your case, where we get f is equal to x 1. Technically this means probability. So, similarly we find out for x 2 I am just writing it out. Then we find out for the next reading third reading, x 3 then x 4 x 5 x 6 continue and multiply all these values in order to get the so called maximum value.

So, this is what we are trying to do. So, consider x1to x n have the same distribution function they are I I D; that means, Identical and Independently Distributed. The same distribution function f of x given theta which is the parameter theta is unknown remember that. And x1to x n are the corresponding realized values for capital x1for capital x 2 for a capital x 3 till capital x x n respectively. As all the x is are I I D hence, the likelihood function corresponding to the fact that X takes x1in the first case X takes x 2 in the first second case X takes x 3 in the third case so on and so forth. In the last case when x takes x n in the last case, is given by the corresponding probability being multiplied.

So, when I mentioned, this actually this value is basically what I have written. So that means, I have written suffix X1for the case of capital X that corresponding to the fact that is the random variable. Which will be a key and a keen or the characteristics will be given for the first trading, similarly, capital X 2 capital X 3 so on and so forth.

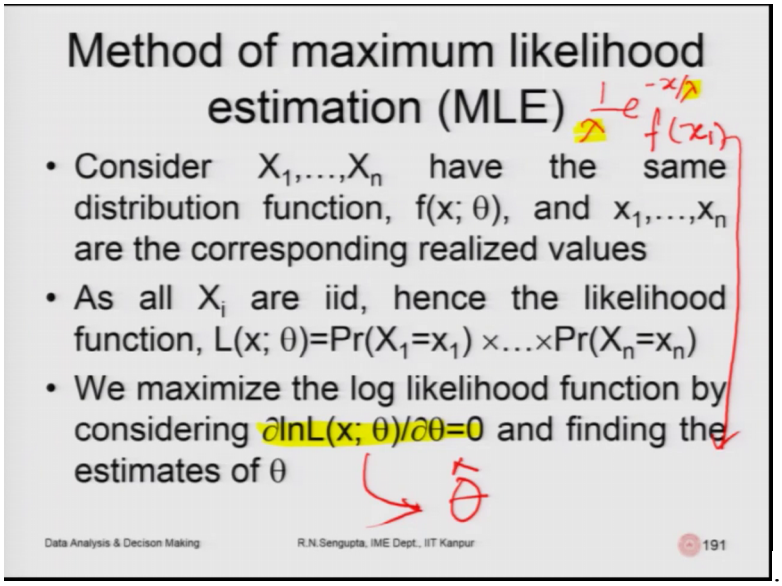
So, you multiply these values, these are multiplication signs where I am hovering my pen. Now considering that we want to find out some theta. So, if you want to basically maximize the probability so, which means, that that value of theta which gives us the maximum chance on speaking of this observation should be the parameter value such that, we are able to maximize the probability.

So, once we basically multiply the probability corresponding probabilities for the IID's we find out the log likelihood function because, as I said this is a monitoring increasing function. Once you find out the log likelihood function x 1 to x n are all known? Now the value which is unknown is basically a value of theta. So, if theta can be in case of a normal distribution they can be two parameters mu and sigma, for the case of exponential it can be a nth lambda depending of a is nonzero.

In case a is 0, then it will be lambda. We maximize the log likelihood function by differentiating. So, I should remove this because or else it will be. So, we maximize the log likelihood function by considering and differentiating this. So, these are the so the del of l n of log of l of x theta is the log likelihood. So, likelihood function we convert into log likelihood and differentiate with respect to theta put it to 0 and finding the estimates of theta.

So, once we put it to 0 you will have different values of theta. So, those would be the estimates of theta.

(Refer Slide Time: 11:31)



And we technically write that as theta hat. So, once we differentiate put it is 0, that value of theta which we found out would be basically the corresponding characteristics would be coming from the sample only because, here inherently the x 1 and x n are known to us.

So, once we once we find out x 1 to x n, we put it back again I am repeating find out the likelihood function find first then find out the log likelihood function differentiate with value of theta partial differential put it to 0 find out this theta hat. In the method of moments we work in a slightly different way. So, moments if you know the mean value is the first moment the variance is basically the common is actually the second moment concept which is being utilized to find out the variance.

So, variance basically uses the second moment and the third moment. Similarly we have for the variance we use the second moment and the first moment my apologies. For the third moment we use the concept that given the third moment you can you utilize the first the second and the third to find out this skewness. Then once we find out the fourth moment we can find utilize the first the second the third and the fourth to find out the kurtosis and so on and so forth.

So, what we do is that depending on the parameter values how many of them are to be found. We derived the equations that relate the population moments to the parameters of the interest. So, say for example, parameters for the case of exponential distribution or for the case for normal distribution so the parameters would have some relation with the moments. So, the first moment would basically the first parameter for the normal distribution.

Second moment would basically given by the mu and sigma square such that we can find out the relationship between the second moment and the first moment and the variance. So, if we have the relationship between the moments and the parameters we basically write down those equations. So, if consider that you have basically k number of parameters you will basically have the first moment, second moment, third moment, fourth moment till the kth moment. And they would be given by the equations g 1 theta 1 to theta k because, all the parameters would be taken into consideration some may be 0 due to the calculations, but we will take basically all of them in the calculations.

Similarly, we will find out the second moment which is mu 2 which will be given by g 2 a function of theta 1to theta k till the kth moment would be given by mu k is equal to g k and then functions of theta 1to theta k. So, once we have basically have the formulas as per the theoretical norm, what we do is that? First we basically find out the first moment with respect to the sample characteristics which you have found out. And basically

utilize that sample characteristics in order to basically estimate and find out the moments one at a time.

So, say for example, we have the sample mean. Sample mean is utilized to find out the first moment. So, we get some information of the mu. Similarly we find have the sample we from it we find out the sample variance, utilize the sample variance and the sample mean to find out what is the population second moment or also find out what is the variance of the population? We go step by step find out if they are close form solution well and good.

If they are no close form solution you do some iteration method, to find out that how the sample information set of information from the moment point of view like the first second third fourth from the sample can be utilized to find out the first second third fourth moment for the population, hence, we can find out the parameters for the population as such.

So, it is written there in the second point after that a sample x 1 to x n is drawn and the population moments are estimated for that. So, here we consider mu is and this is suffix e, e means the estimate. The estimate of the sample the moments would basically be found out from the observations x 1 to x n these are the characteristics corresponding to the set of observations x 1 to x n.

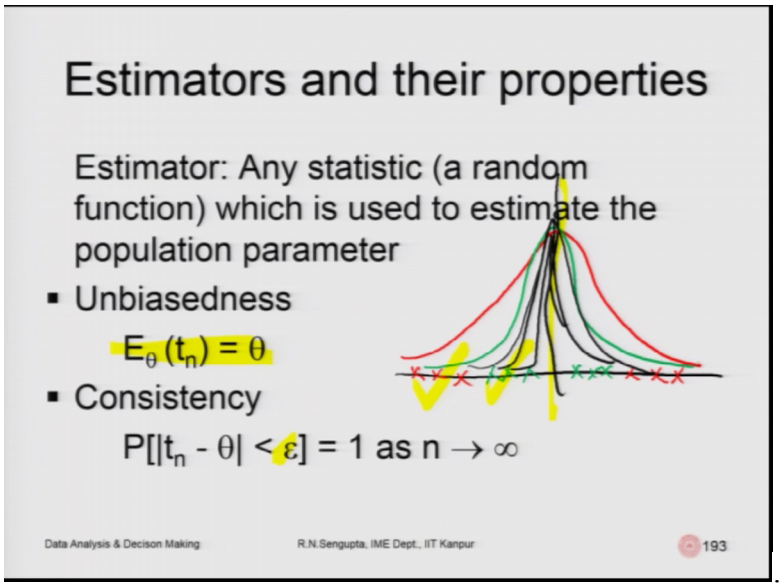
They can be equated to basically the functional form which you have initially derived, but from there what we will find out is basically not the parameter. The parameter estimate which will be found out and then we will try to compare the estimate of the parameter with the parameter and then basically go into the characteristics of the unbiasedness and consistency and comment intelligently accordingly.

So, from the sample we find out let me read it mu i e. When i means the ith moment, e is basically the estimate is the function of x 1 to x n. So, that will be given by so for the first case it will be given by g 1 and interval it will be a function of the estimate of theta 1 to theta k. Second case it again would be basically g 2 a function of theta 2 theta 1 to theta k, but here also remember they are the estimates. We continue doing it such that we are able to find out this theta 1 estimate theta 2 estimate theta 3 estimates so on and so

forth, which are the best estimates for the actual parameters which are theta 1 theta 2, theta 3 till theta k.

So, if you are basically the mean value only, you will use the first moment. If you have basically the variance you will use the first and the second moment and so on and so forth. This is a slide which is a repetition, but I will still repeat it because this would be coming time up again generally when you are trying to solve the problems.

(Refer Slide Time: 16:49)



But obviously, the answers generally the proofs would not be given or they are not required the final answers would be given for those proofs. And these those answers will be utilized to basically find out the point estimate and utilize the point estimate to find out basically the interval estimate and the do the hypothesis testing.

So, any statistic a random function which is used to estimate the population parameter, basically should have basically two characteristics one is the unbiasedness and one is the consistency. In unbiasedness if you remember that what we want to do is that? I will draw the same diagram which have done. So, let me change the color such that is easy for us to follow. So, this is the x axis this is the y axis and consider we have chosen observation for two different populations and consider them both of them are normal.

So, let the set of observations for the first one be red in color and we have we have picked up say for example, three observe six observations and drawing arbitrarily and the

mean value is like this. So, the mean value is the central line the variance is basically this and the distribution looks like this so this is normal. Next consider we pick up a second set of observation against six in number from a normal distribution, but now the color combination is green.

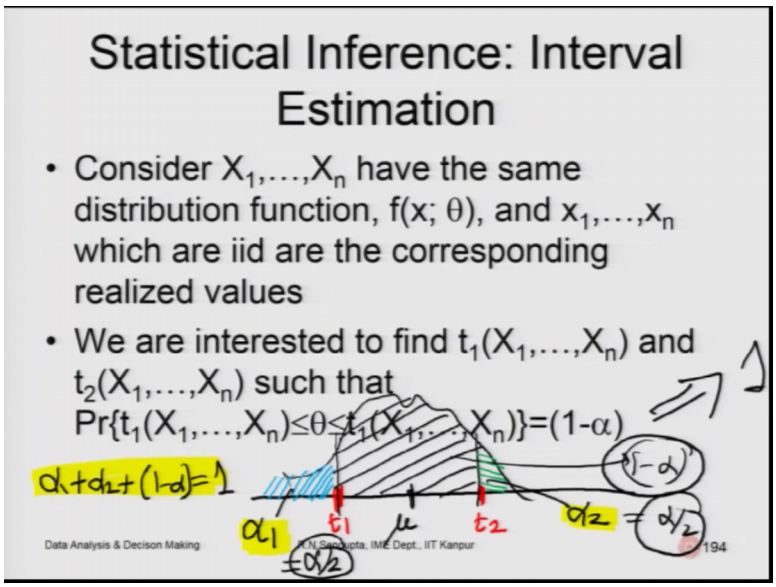
So, here also the mean value is same. So, the expected value let me use the highlighter. The expected value in both the cases for both for the red as well as for the green are same which is this is the central line. But if we go to the second characteristic which is consistency you will find out the difference between the actual estimate value and the parameter value from the population. They are basically a function of n because as n increases; obviously, t n would basic actually approach the value of theta which is the population parameter.

And the distance between whether it is on the positive side or the negative side will start decreasing. Such that the epsilon value which you have put for ourselves in this case, is a function of n and in the long run as n tends to infinity the difference basically goes to 0 with a probability 1.

So, which means that, as sample size increases the distribution let me use the dark color. So, these are I am utilizing as for any of them, whether for the green or the red as the sample size increases in the long run the distribution looks like this. So, it basically peaks and basically gets concentrated in around the mean and hence the variance decreases.

Now, consider the concept of interval estimation. So, so this point estimation would be followed up by interval distribution and they would be logical flow from there I will come to that later as we solve the problem. I will point it out to all my students.

(Refer Slide Time: 20:01)



Consider x 1 to x n have the same distribution function. Which is basically so, I am considering they are I have a population from there I am picking observations and they would be IID's.

So, obviously, because, 1 observation does not affect the other so, we will consider I D and they are coming from the same distribution. Consider x 1 to x n have the same distribution function f of x given theta as the parameter that theta can be vector or scalar and that does not matter. And we consider x 1 to x n are II D's and the corresponding realized values of them are given. So, what we are interested in is given that x 1 and x x n and depending on the level of confidence or the of the probability value of how good or bad are set of observations are we want to basically find out two estimates or give us two functional form.

So, those two estimates would basically be a function of the sample estimate, such that within the bound of the lower and the upper estimate or the lower control value and the upper control value we will find our actual population parameter. Which is the mean or the standard deviation or the mode or the median; whatever it is with a certain probability. Where that probability has already been being dictated by us before we solve the problem or it has been given by the external sources.

So, those basically functional forms based on the sample set of observations would be given as t 1 or t 2. So, those functional forms I will come to that in few minutes. So, t 1 and t 2 would basically with the functional forms based on which we will find or the lower control limit and the upper control limit. And these t 1 and t 2 are just functions based on the fact that you have picked up observations x 1 to x n.

So, what we want to find out is this I will draw it here. And use the red color for either understands first I let me draw the he line and straight line and then try to basically give the color combination concept. So, consider this is the line let me check this is the x axis and consider this is the mean value.

Now, what t 1 and t 2 me are actually this. So, t 1 is a value on the left hand side which is a function of x 1 x 2, t 2 is a value on the right hand side which is again a function of x 1 to x n sorry. Such that the overall area in which so, consider any distribution I am not drawing the normal case any distribution which you have. Such that the bound which I have would contain the mean value or the parameter value from the population with a probability of 1 minus alpha.

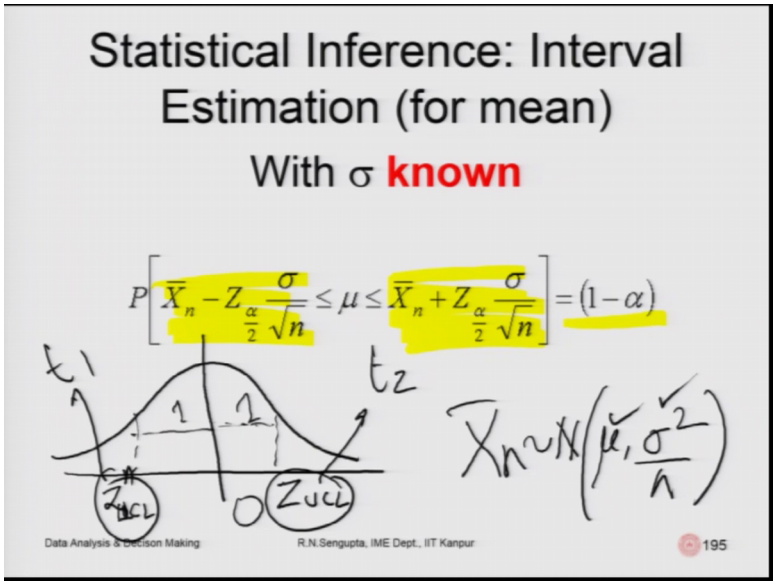
So, this is the one minus alpha or the level of confidence we are seeing. Now if it is a symmetric distribution what we will do is that I will use a different color no this yellow one go let me change to blue. So, this left hand side probability and this right hand side probability. They would be such that the sum of considered I am using the black color also again this is alpha 1 this is alpha 2.

So, alpha 1 plus alpha 2 plus1 minus alpha which is the level of confidence between should add up to1 which technically would be the case as discussed. But the beauty is that we considered a symmetric distribution and we considered alpha1 and alpha 2 be equal and equal to half alpha by half. So that means, the central portion will be1 minus alpha the left and the right portions on to the left of the l c l on to the right of the l c l would be equally divided.

So, technically alpha 1 value alpha 2 value would be equal to alpha 2 alpha 2. So, as the addition of this plus this plus this gives me 1. So, we are interested to find out lower control limit and the upper control limit which is given by functional forms t 1 and t 2. Such that the probability of the mean value or the parameter value of the population be

bounded by that lower control and upper control is given by 1 minus alpha which is the level of confidence. Without the derivations I am just and I am just giving the results.

(Refer Slide Time: 25:25)



So, statistical inference for the interval estimation for the mean with them the standard deviation of the population variance for or the standard deviation for the population on the variance of the population is known. Because that is why it is highlighted in a red color? So, if that is the case what you will do is that; if you know the best estimate so, I am I am basically now going through the background of the story without going into the details.

So, obviously, if the mean value is unknown so, obviously, have to estimate from the sample. So, your first question would be if I want to basically estimate from the sample for the normal distribution, what is the best mean value? So, we know that is the sample mean. Now we will also ask that if the population variance or the population standard deviation is known then how we should solve it.

So, we will we will think and remember that the sample mean as it is the best estimate hence, it is distribution is normal which is true the mean value for that that a sample mean is mu. Which is also true and the standard deviation of the variance of the sample mean is given if you remember by sigma square by n.

Now, if it is give it my sigma square by n. So, now, we have x bar n, which is basically the sample mean. As normal distributed with mean mu and variance sigma square by n hence, if it is normal it can be converted into a Z normal Z value which is the standard normal distribution. And it can be converted into a standard normal distribution we can find out t 1 and t t 2 which is the lower control and upper control based on which you can find out that given a probability that how what are these t 1 and t 2? Such that they would encompass the sample and the population mean within itself with a certain level of confidence which has to be specified beforehand.

So, now if I have normal distribution I am drawing it in order to make you understand this should be applicable for all the cases. This is for the actual sample mean mu value the mean value is mu, which is here the standard deviation is given by sigma square by n. So, this is the standard deviation.

Now, I convert it into a standard normal. So, let me we remove it. So, this becomes 0 and these are the variance1 1 this value is Z1 corresponding to X 1. So, I have to find out X1 means those are not the observations X1 to X n those are not those are basically the let me put it a Z l c l and Z u c l. So, they are basically coming from t 2 which is coming from t1 using the standard normal deviate.

So, once basically I find out that t1 and t 2 corresponding to the Z distribution and when we use the m the X bar distributions are. So, this is the l c l if we remember alpha by 2, because if you remember I said that alpha1 alpha 2 our equally divided on the right side the u c l is this plus value. So, this is minus and this is plus. So, they are used accordingly such that you are equidistant from the mean value. That will be equal to1 minus alpha based on this you can find out the interval estimate and conclude your problem accordingly.

So, I will keep repeating these type of problems for the mean values given standard deviation known not known for the standard deviation given mean values you know not known and try to basically utilize the concept of the Z distribution the t distribution the chi square distribution and the f distribution. And then utilize them further on when we come to the multivariate statistics. With this I will end the sixteenth lecture.

And thank you very much for your attention have a nice day bye.