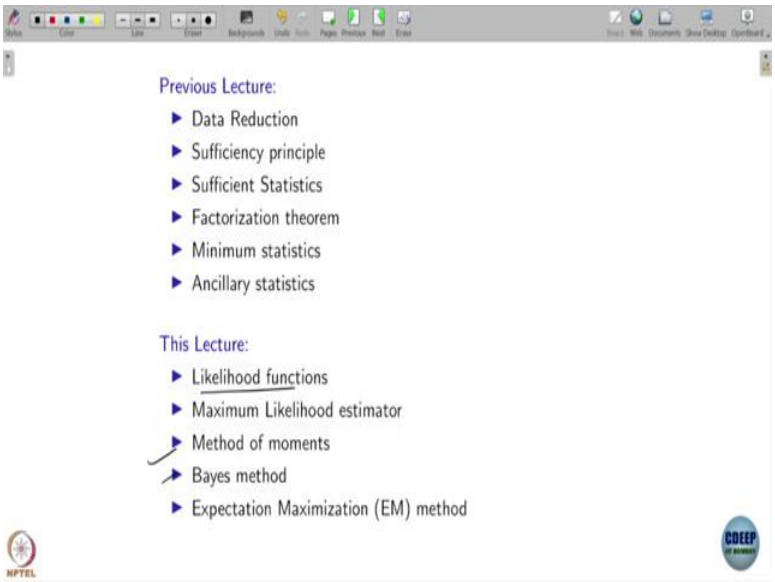
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operations Research,
Indian Institute of Technology, Bombay
Lecture 37
Likelihood Functions, Maximum Likelihood Estimator

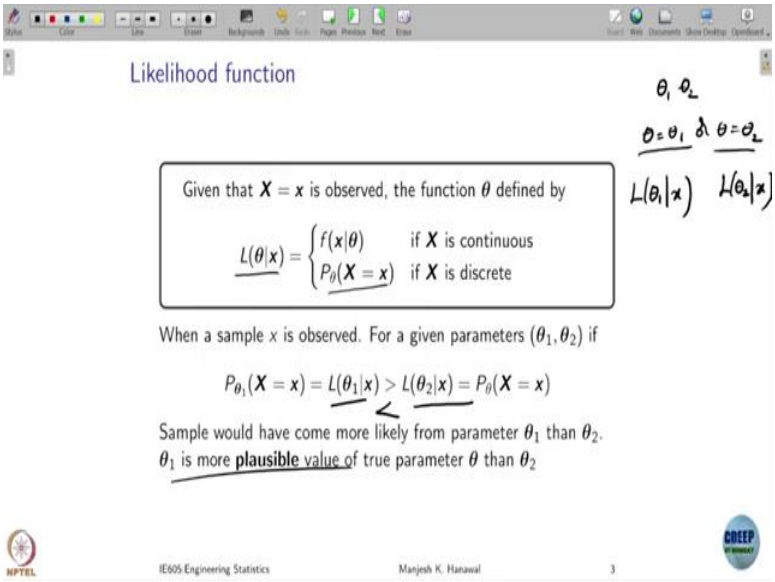
So we, so far whatever we covered in terms of this data reduction principle, sufficiency principles, if you have any questions you can ask now again.

(Refer Slide Time: 0:28)



So what we will do now is we will look into the another aspect that is the likelihood principle today and try to cover like some of the concept like likelihood function, like likelihood estimators, and we will also look into other estimators called moment of methods and Bayesian methods.

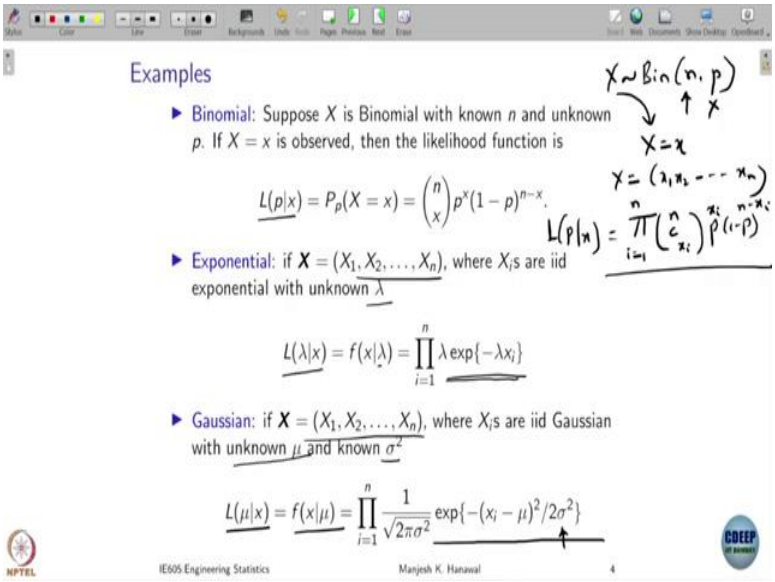
(Refer Slide Time: 0:50)

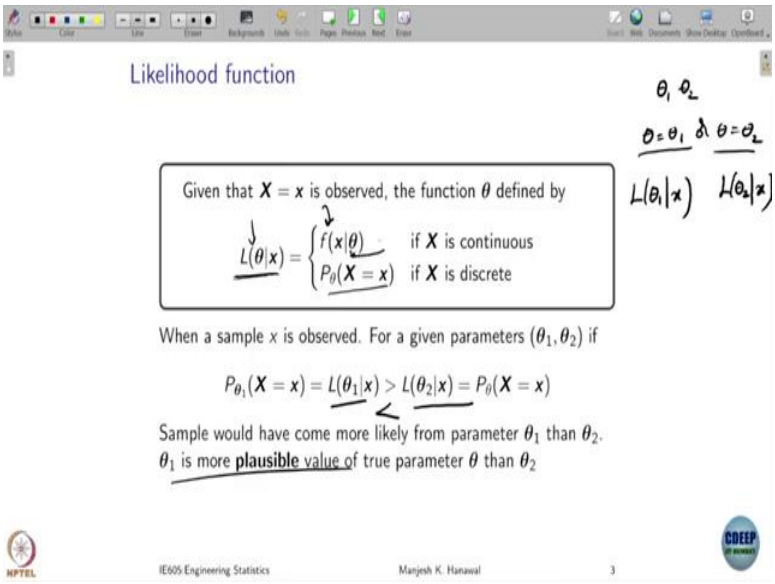


So we wanted to see - what is the likelihood function. Suppose let us say a sample X has been observed. And now I want to see that I now know that X has been observed, how it influences my theta. So to do that I am going to define a function L(theta| x), which we are going to call it as likelihood function. And it is defined simply as f (x|theta), that is probability density function of that point x under my parameter theta when that is continuous.

And I am simply take it as a probability mass function when that is discrete. So we said that suppose let us say we have two parameters theta 1, theta 2 and my true value could be either of this theta 1 and theta 2, I do not know that and I want to identify if theta| equals to theta 1 or theta equals to theta 2, I want to ask this. How I can potentially decide this? One potential way to do is compute its likelihood of observing theta 1 under your function x. And similarly compute likelihood observing theta 2 under this, under your sample x and see which parameter is more likely. And this happens to be more likely than this, then we are going to say that parameter theta 1 better explains my samples x, so I can say my true parameter is possibly theta 1. And if the other case then may I may go with saying that maybe the true parameter is theta 2. So in this case we are going to say theta 1 is more plausible value of the true parameter than theta 2.

(Refer Slide Time: 2:59)





And we saw some examples also which we will repeat again. Suppose let us say you have a binomial distribution with parameter n and p and assume that this n is known and this part is unknown. The parameter p is unknown. And now for time being assumed you are given one sample from this. You have just observed one sample and using this one sample you need to decide or get some information about this parameter p.

So how to connect this observed value with my parameter p? I am going to connect through my likelihood function, which is defined as n choose x, p to the power x, 1 minus p, n minus x and whatever x is the observed value here. This is one way I have defined. I am just giving example of what could be the potential likelihood function one can use.

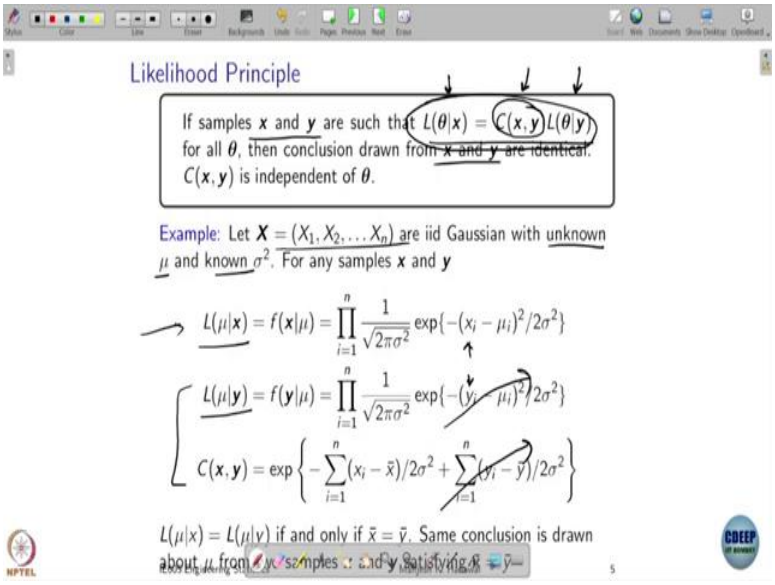
Now instead let us say, I mean, here I have just taken one example, one sample, but it could be a vector also here. It could be, you have observed like this. In this case you can rewrite this likelihood that under this sample p, in this case you are simply going to take product of those n and choose xi, then p to the power xi and 1 minus p to the power n minus xi. Assuming that this sample all of them are independent.

So now similarly, suppose if you are given a random sample where xi's are exponential with parameter lambda. Now one way to write the likelihood function of this is using my pdf function, which I have simply written here. So this is what the probability of observing x under the parameter lambda and this is the PDF at point xi, so I have just taken and multiply all of them.

Now similarly, if you have Gaussian where I have unknown value mu and known value sigma square, then also I can simply go and write a likelihood function in terms my PDF function using this expression. So here sigma square is known and this mu is unknown and now this is a function of mu given by xi's. So notice that you should not get disturbed, confused here, the interpretation of L theta given x and f x given theta.

So here this one, when we wrote this, we said probability density of x under the parameter theta. In this case theta was given and we try to find PDF of x, whereas, in this case my x is given and now I am trying to see this as a function in theta, x is fixed here. But now for a given theta x, I am just taking this to be f of x given theta. And if my theta changes here, so my theta changes here also accordingly.

(Refer Slide Time: 7:00)



Now the likelihood principle - what is that likelihood principle we are talking about. We are saying that if I have two samples x and y, such that one can be expressed the likelihood of parameter theta and these two samples can be expressed in this form for all theta, then we are basically going to say that the conclusion drawn from x and y about the point theta is the same, provided this quantity c of x y does not depend on my parameter theta.

So in a moment it will become clear to you why that, this transformation are going to be identical, when I am looking to obtain information about theta. And once this guy is not dependent on theta, actually all the information about the theta is only coming through this and if they are transformed to this function, which does not depend on theta, in a sense it is telling us they are kind of identical. The information provided by x and y are identical.

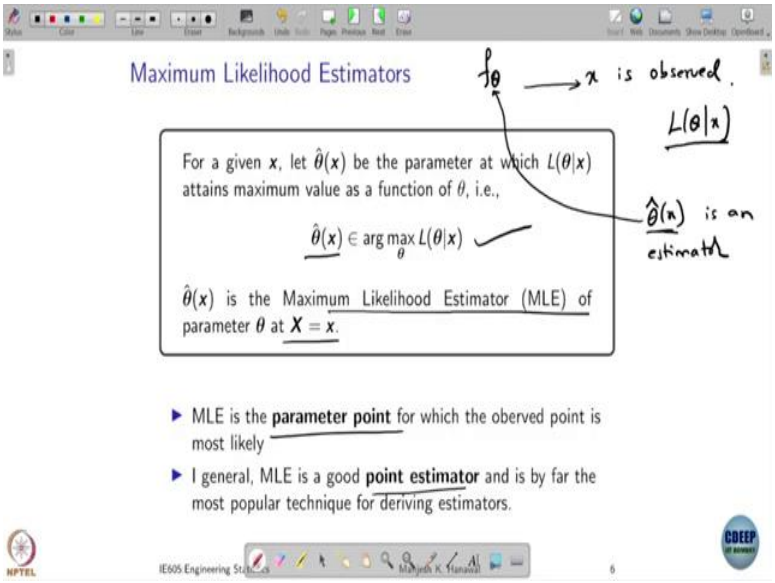
So let us write an example for this. Suppose I have this random sample where again for simplicity we will take only mu is unknown, but sigma square is known. So this is the likelihood function under observation x for the parameter mu and this is my likelihood function for parameter mu under my observation y and I have just written one function C(x,y) here ,which does not depend on mu but depends on only the known parameter sigma square.

So as far as parameterization concerned, the C(x,y) does not depend on my parameter theta. So theta is mu here. Now if these things you want to compare, these two are going to be identical, if all this samples are going to be identical, but now, if you take this l, you can write this l mu x y here into the product of l x L mu given y and C(x,y), you can write it. So now let us simply consider l mu, sorry, lLmu given y and C(x,y).

If I take this product what is going to happen? This entire thing get knocked off with this guy and I am only going to get exactly L mu x. Nothing changes, so this factor is still there and this guy is there. That is fine. Now as we said, now I am able to write L mu x as a product of these two things and L mu given x and l mu y are going to be same for me in this case, they are going to be same through this if this x bar is going to be equals to y bar.

So what we are saying is the information provided by these two samples x and y they are going to be same if I can write them through this transformation, provided x bar is equals to y bar in a way, I mean, we also get to know the sense, earlier we know that x bar and y bar, they are trying to essentially capture information about mu, but if they are same that means they have captured the same amount of information about the parameter.

(Refer Slide Time: 11:40)



Now the thing is we have this likelihood functions, which is now functions of my parameter theta. Now my goal is to identify the best theta there, which explains my observed samples. So x is what I have observed, x is observed. Now I have this l theta x and now what I can do is try to find a theta, which maximizes this likelihood function and that is what I am going to call it as theta hat x.

And I am calling it as a function of x because if you are going to change your x this theta hat x is going to change and this theta hat x that maximizes your likelihood function over all your parameters theta is called maximum likelihood estimator of your parameter theta, the unknown parameter theta at the point x. So notice that now this theta hat x I am calling it as, is an estimator.

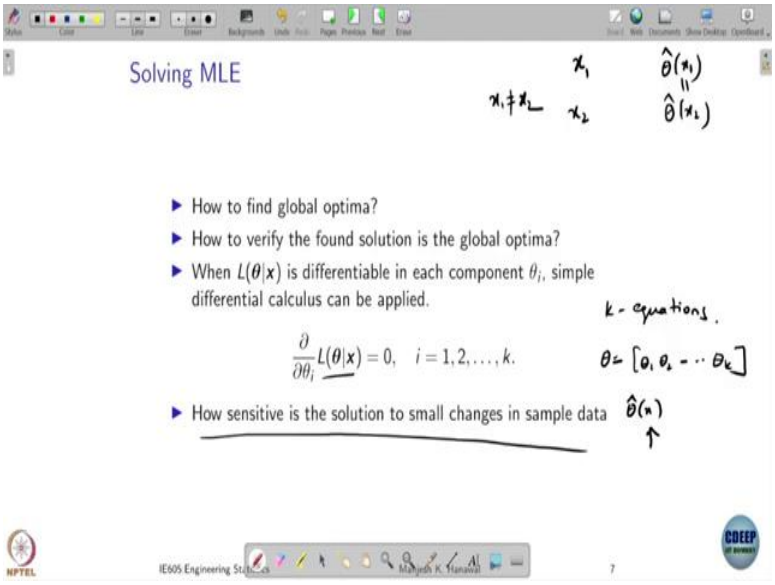
So earlier also we had an estimators. We talked about sample mean as an estimator, sample variance as estimator for mean, sorry, variance and all, but now given a parameter, given a sample x and you are interested in understanding how this, you are interested in estimating this parameter theta, now I am obtaining it by maximizing this likelihood function, which I am calling it as theta hat.

And now this MLE we are going to call it as a, we are going to call it as a point estimator, but before that this MLE is the parameter point for which observed parameter is most likely. So whatever theta hat you are going to get what you are going to say this is the parameter that most likely that had generated your samples, when you consider that parameterized probability density function.

So what you did? You have this underlying theta hat, we did not know about this theta and from this sample x is observed. Now what I got is, we got a theta hat x. Now we are just saying that this theta hat 1 is the closest approximation of this theta. So the PDF, whatever the parameter is associated with is exactly now we are believing this is that parameter based on the observed x.

So in general this MLE is a good point estimator and this is one of the most popular estimators that you are going to see. So to find an maximum likelihood estimator, first you need to appropriately define your likelihood function and then optimize it using this criteria.

(Refer Slide Time: 15:39)



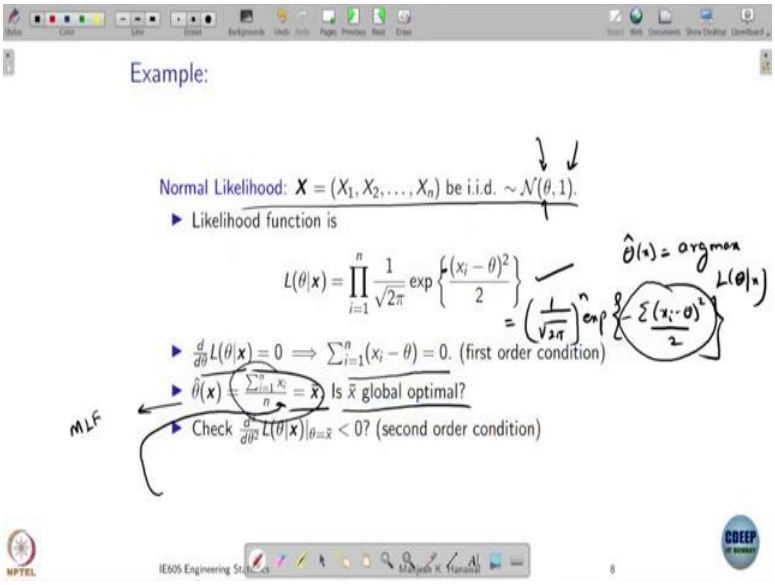
Now there is an optimization here. You have a, defined a function and there is an optimization here for your parameter space. Now the first question is, whenever I try to optimize a function whether whatever I am getting is a global optima or not. Now I need to have a mechanism to check whether my solution obtained is global optima or not. Now how you are going to obtain this solution?

The standard technique you know is you differentiate your likelihood function assuming that assume for time being that this function is differentiable in your components theta and then you are going to equate them to 0 for each of this I equals to 1 to k and then when you solve them, let us say by doing this you get k equations and solving those k equations you get the cape components of your parameter theta like this is theta 1 theta 2, like this theta k.

Now notice that as I said this theta we are trying to find and this theta hat we get this is going to be a function of your samples. Let us say you observed initially one random sample, let us call that X 1 and you got theta 1 x 1. And you had gotten another sample and for that you got theta x 2. This these two are necessarily same. Let us assume that x 1 is not same as x 2. If x 1 is not same as x 2, do you expect theta 1 of x 1 and theta hat of x 2 to be the same?

Not necessarily. If there is a small change in sample data you are estimator can also change, so it is going to be sensitive to your sample, but then the question is how sensitive is a solution to the small changes in the sample data.

(Refer Slide Time: 18:03)



So to understand that let us start calculating this maximum likelihood estimator using some samples itself, like samples from some particular distributions itself. Let us take normal functions, normal distribution, where theta is its mean value and its variance is going to be 1. So I have deliberately put the variance equals to 1 because now this is fixed, the only unknown quantity here is the mean value.

Now how you are going to write the likelihood function? Likelihood function is simply going to be the product of the PDF computed at each of these sample points. So everybody agree with this likelihood function here. So notice that sigma square is set to 1 in this. Now let us try to find out, let us try to…

Student: (())(19:16)

Professor: Yeah, that is right. There is a minus here. Now I want to maximize this to find an estimator at point argmax of L theta given x. So to do that first I am going to differentiate it. There is only one parameter here theta, one dimensional parameter so I am just going to differentiate this with respect to the parameter theta. So if I differentiate it, I am going to get this condition.

Everybody agree with this? So one way to write this is, another way to write this is 1 upon square root of 2 pi to the power n exponential minus xi theta whole square by 2. Now the only way, part where theta is coming is in the exponent. This is a constant. So it is as good as, just focusing on this part and differentiating this. Everybody agree? So if you are going to differentiate this part, this is going to be 2 times xi minus theta by 2 into minus 1.

But that 2 is a constant, so I will get simply this condition. Everybody agree? Now if I simplify this theta x theta hat x, what is the solution I am going to get, I am going to get summation of xi by n. Now let us say summation xi by n, I got it as theta hat of x, but now is this the global optimal, is this the global optima or is this globally optimal. How to check that? How do you check whether the solution from the first order start condition is globally optimum?

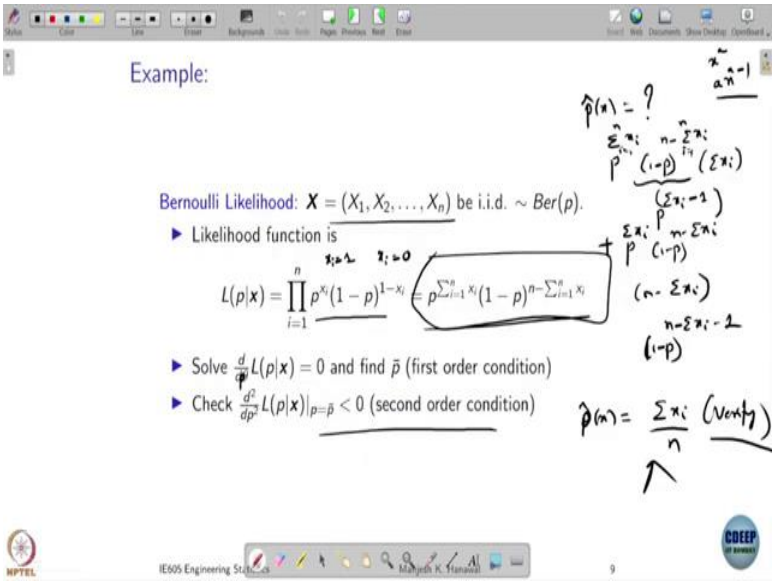
Student: (())(21:25)

Professor: Yeah, you check for a second order condition and it so happens that in this case when you check this value at this quantity x bar their second order derivative is negative, so indeed this quantity be is going to be global optima in this case. Now we have this seen this quantity many times, what is this quantity, it is a sample mean. So what we are saying is we are actually saying likelihood estimator is saying that sample estimators…

Sorry, sample mean that we have seen many times is actually a good estimator for estimating this parameter theta, which happens to be the mean of this distribution in this case. So we know that theta is the mean of the Gaussian random variable. We know already that sample mean, we call this quantity as a sample mean and took it as an estimator for mean. Now likelihood estimator is saying that okay, if you are interested in estimating is parameter theta simply take the sample mean.

So sample mean is the best estimator for estimating the parameter theta here which is the mean of my distribution. So anybody has any question in this the way you are computing our sample mean or how we are computing the… So this is my maximum likelihood estimator what I am calling is as MLE, which is the sample mean.

(Refer Slide Time: 23:08)



Now let us look into another example. Let us say you have n samples which are i.i.d drawn from Bernoulli with parameter p. Now first thing I will do is I will write the likelihood function for this. The likelihood function for this is going to be simply this one, everybody agree. We have done this exercise before also. For example, if xi equals to 1, this quantity is simply going to be p.

And if xi equals to 0, this quantity is going to be 1 minus p. Now this could be written in this format. The product become summation in the exponent now. Now let us try to optimize this with respect to my parameter, which is p here. I think I should have written p here. Can somebody quickly differentiate this function and see what is the p hat function I am going to get. p hat of x is what?

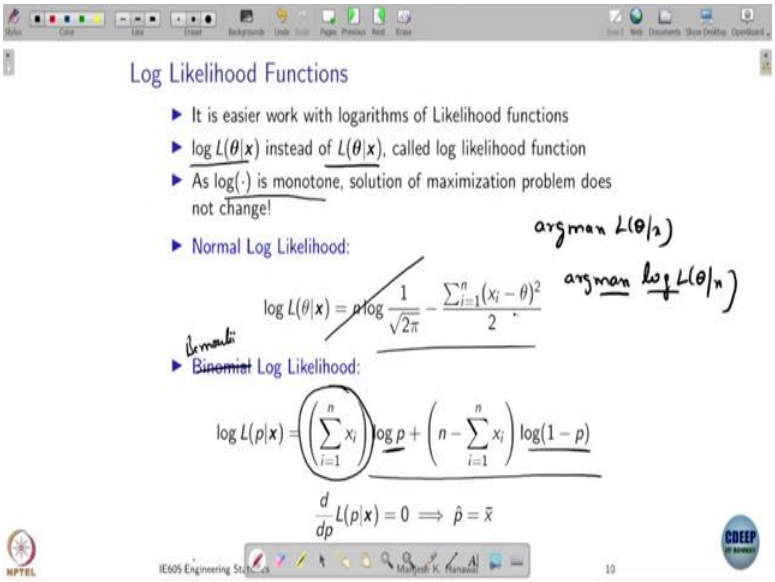
Is it easy to differentiate this with respect to p? Let us differentiate it. You are going to get something p summation i equals to 1 to n xi. I am going to hold it n minus summation I equals to 1 to n xi and so this is going to be summation xi and p summation xi minus 1, something like this we get. I hold this term as a constant and trying to differentiate the first term. So what is the differentiation of x to the power a?

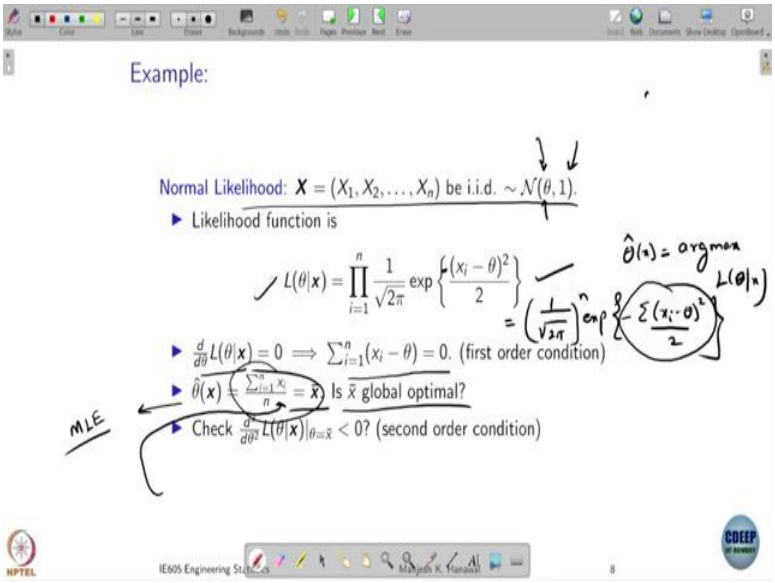
Is it a x a minus 1? So then, yeah, this is what I have done. And now similarly, differentiate the other term p summation xi and then n minus 1 minus p n summation xi, n summation, sorry, 1 minus p n to summation xi minus 1. So I am deliberately trying to do this because there is a simpler, much simpler way of doing this than all this complication derivations we have to do.

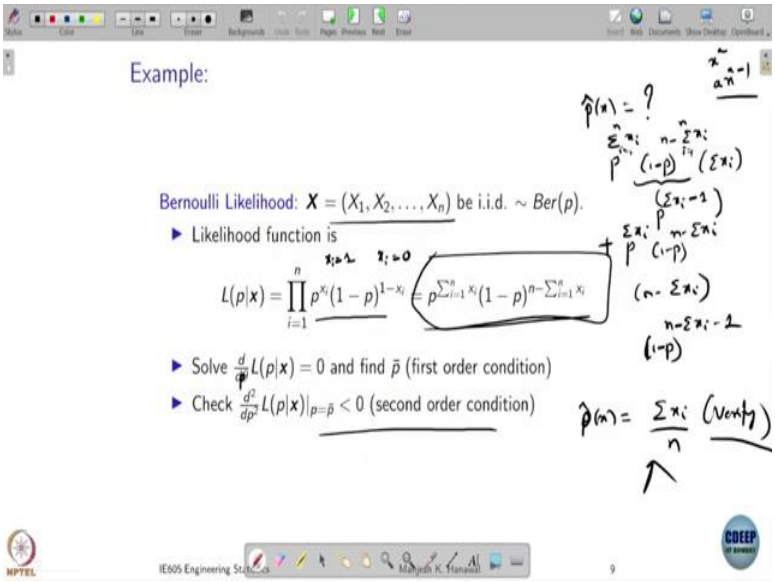
But you see that like, I mean, there is a little bit differentiations I have to do, but if I do couple of steps I will be able to do this and you will see that the theta, the p hat that maximizes this is again given by summation of xi by n. This simple calculations you can verify. Now you can check that that same solution xi by r is also actually makes your second order derivative negative that means this is going to be the optimal point.

You see that, like once I have this log likelihood function I just need to follow the steps of differentiation and get this value. Now there is something called log likelihood come function which will help us simplify our life when we are faced with such functions.

(Refer Slide Time: 27:04)







So what it says? Often it is easier to work with log of your likelihood function than with the likelihood function itself. And since log o is a monotone function, the solution to the maximization problem does not change. So suppose let us say I am going to find an argmax of some log function and again I am going to find argmax of log of l theta by x. Does this value, optimal values change?

Because log is a monotone and I am maximizing this, it does not change. Now let us see how does the log help. So you have this function here initially when you computed for the Gaussian, if I look, apply the log function here, you will see that this will become simply this. And now it becomes easier to optimize this, because the first part is constant does not belong to, does not depend on theta, only the part depends on this.

This is already simply linear in theta and if you can optimize this, sorry, this is quadratic in theta and you can easily optimize this. And similarly, for the Bernoulli case we had this complicated function, so this should be Bernoulli. If you take the log function, this again simplifies in this format. And now this is like a constant for a given x only thing is log p is there and log 1 minus p and this is lot easier to differentiate and compute it.

So that is why often instead of dealing with the likelihood functions, we deal with a log likelihood function. This will make our calculations easier.