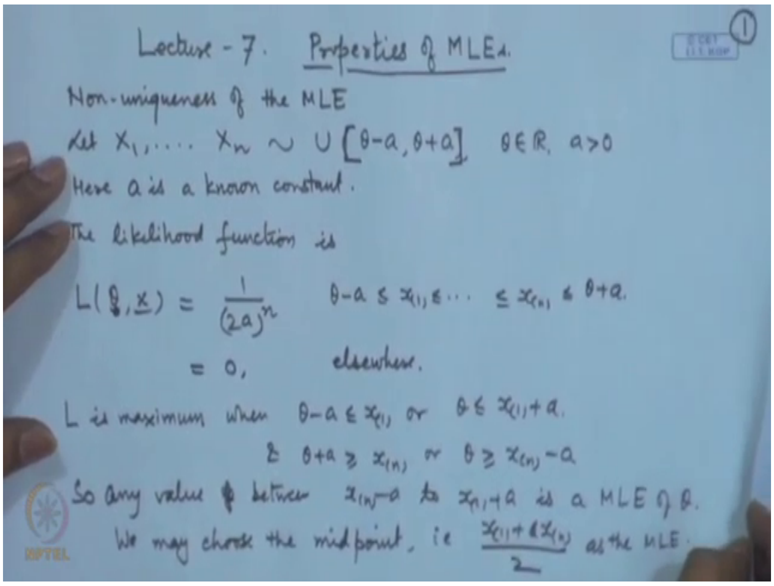
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 13
Properties of MLES - I

In the last lectures, I have derived the form of the maximum likelihood estimators for various probability models. I have demonstrated how the role of likelihood is there in determining the final form of the estimator. Suppose, there is a prior information, then it has a effect. Now, in today's lecture, I will spend some time on discussing important properties of the maximum likelihood estimators. First of all, we note that see various problems we have done and in most of those problems you have got a value of the maximum likelihood estimator; that means, there is a function which is corresponding to the estimator. However, that is not necessarily the case; sometimes we may have a non uniqueness.

(Refer Slide Time: 01:06)



So, let me give an example of that non uniqueness of the MLE. Let me discuss one example of that let we have a sample X 1, X 2 X n from a uniform distribution on the interval theta minus a 2 theta plus a where theta is a real number and a is a positive number, here a is a known constant. So, the problem is here to estimate the parameter theta of this uniform distribution; that means, the spread is from theta minus a to theta plus a. Since, theta is unknown; we do not know the starting and the end point of the spread.

So, let us consider the likelihood function. The likelihood function is now here the density function is 1 by 2 a. So, if I consider the joint distribution of X 1, X 2 X n, it will become 1 by 2 a to the power n and each of the X i s lies from theta minus a 2 theta plus a. Therefore, we can summarize this information in the form that theta minus a less than or equal to x 1 and so on, less than or equal to x n less than or equal to theta plus a. Let me put it here closed interval because I am including the endpoints here it is equal to 0 elsewhere.

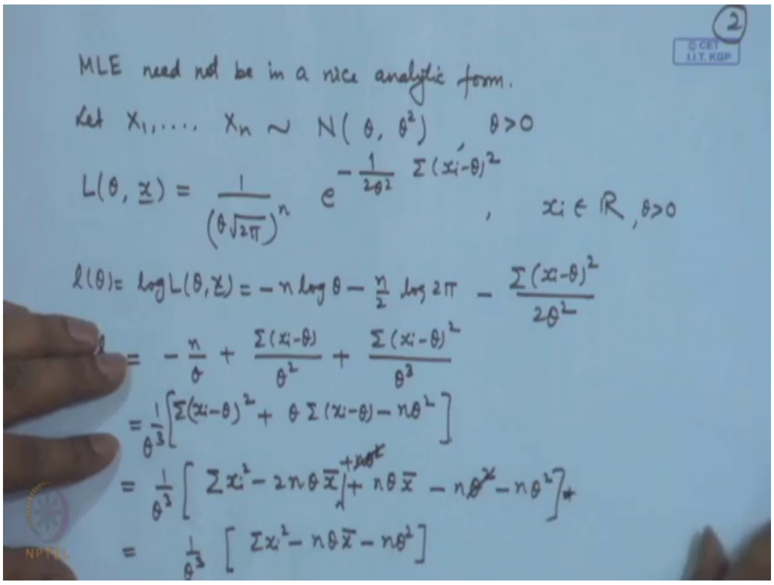
Naturally, you can see that the maximum value of the likelihood function is 1 by 2 a to the power n because, this is a constant value here at other points it is 0. Now, this is satisfied then this inequality holds true. Therefore, let us see the optimal range of theta for which this value is attained. So, L is maximum when theta minus a is less than or equal to x 1 or you can say that theta is less than or equal to x 1 plus a and theta plus a is greater than or equal to x n or theta is greater than or equal to x n minus a.

Naturally, if I choose any value of theta in the interval x n minus a to x 1 plus a that will be the maximum likelihood estimator. So, any value of between x n minus a to x 1 plus a is a maximum likelihood estimator of theta. So, this is an example where the maximum likelihood estimator is not unique.

However, we may choose the for example, the midpoint of this that will be the x 1 plus x n by 2. We may choose the midpoint that is x 1 plus x n by 2 as the MLE.

Now, another feature which we noticed in the various problems that we have done that in most of the cases, we got it a very nice function. For example, we got it as x bar 1 by n sigma x i minus x bar whole square the median, the largest or the smallest etcetera. In most of these cases the maximum likelihood estimator is in a closed form and also a mathematically elegant form, but even that is not necessary.

(Refer Slide Time: 05:30)



Let me take another example where we do not get a nice analytic form, MLE need not be in a nice analytic form. Let me take one example here. Let X 1, X 2 x n be a random sample from a normal distribution with mean, theta and variance theta square.

So, naturally here theta is a positive parameter here, the likelihood function. So, here your variance is actually the square of the mean. So, there is an interrelationship here. So, the problem reduces to one parameter. So, if we consider the likelihood function, it is a joint distribution 1 by theta root 2 pi to the power n minus e to the power minus 1 by 2 theta square sigma x i minus theta whole square and here each of the xi is on the real line whereas, theta is positive.

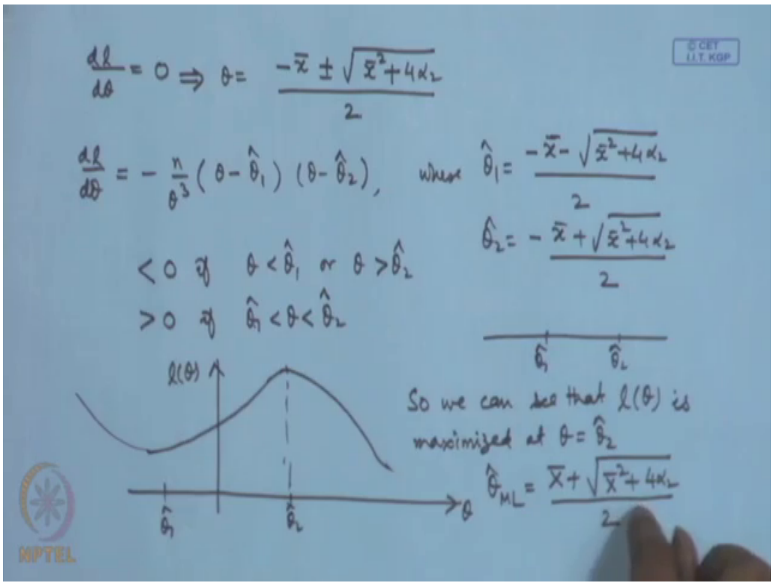
So, if we consider the log likelihood that is equal to minus n log theta minus n by 2 log 2 pi minus sigma x i minus theta whole square divided by 2 theta square. There is naturally a difference from the situation when we had considered mu sigma square because then we had two parameters and we had considered the maximization with respect to both of them. Now, since mu has been replaced by theta, so, this is a consolidated function of theta that is coming here and we have to maximize this with respect to theta, but nevertheless this is a differentiable function and therefore, we can think of the usual calculus procedure.

Let us look at d l by d theta. So, that is equal to minus n by theta plus sigma x i minus theta divided by theta square plus sigma x i minus theta whole square divided by theta cube because the derivative of 1 by theta square will be minus 2 by theta cube. So, that simplifies to this. This term we can write as sigma x i minus theta square plus theta sigma x i minus theta minus n theta square divided by theta cube.

Now, we can expand these terms here, you will get 1 by theta cube sigma xi square minus twice. Now, you get 2 theta x i when you put summation here it becomes sigma x i which we can write as n x bar. So, minus 2 n theta x bar, then you have theta sigma x i again which you can we can write as n x bar. So, this becomes n theta x bar minus theta square and this is summation here. So, minus n theta square minus n theta square and of course, there was another term here which we missed here, theta square here. So, n theta square will come here with a plus sign plus n theta square.

So, naturally this term cancels out and here one of the n theta x bar cancels out, we get here 1 by theta cube sigma x i square minus n theta x bar minus n theta square. Now, we can consider the if we put this equal to 0, then this is nothing but a quadratic equation in theta which will have two roots because the denominator is theta cube which is always positive.

(Refer Slide Time: 09:58)



So, we can look at this d l by d theta is equal to 0 gives theta is equal to. So, we can write this equation the sigma x i square term is coming we can take out n here. So, this we can write as minus n by theta cube theta square plus theta x bar minus 1 by n sigma x i square, I will use the notation alpha 2. This alpha 2 notation we have introduced in the method of moments. This is the second sample moment, this alpha 2 is 1 by n sigma x i square.

So, if we put this equal to 0 we can straightforwardly apply the b square minus 4 a c formula. So, we get theta is equal to minus x bar plus minus square root x bar square plus 4 alpha 2 divided by 2. So, naturally there are two solutions and we have to see the increasing and the decreasing nature of this. So, we can express d l by d theta as equal to minus n by theta cube theta minus, let me call it theta 1 hat into theta minus theta 2 hat where I am taking theta 1 hat to be the solution with the negative that is minus x bar minus square root x bar square plus 4 alpha 2 divided by 2 and theta 2 hat is equal to minus x bar plus square root x bar square plus 4 alpha 2 pi 2.

Now, let us look at the sin scheme of this. This term will be negative if theta is less than theta 1 hat or theta is greater than theta 2 hat because if theta is less than theta 1 hat, this term is negative here we can see that theta 1 hat is less than theta 2 hat. So, if we are considering theta 1 hat and theta 2 hat here. So, if theta is below theta 1 hat and theta to hat, then both of these terms are negative their product is positive. So, this entire term dl by d theta will become negative. Similarly, if theta is greater than theta 2 hat then this term is positive as well as this term is positive. So, the overall term will become negative and this will become positive, if theta 1 hat is less than theta is less than theta 2 hat.

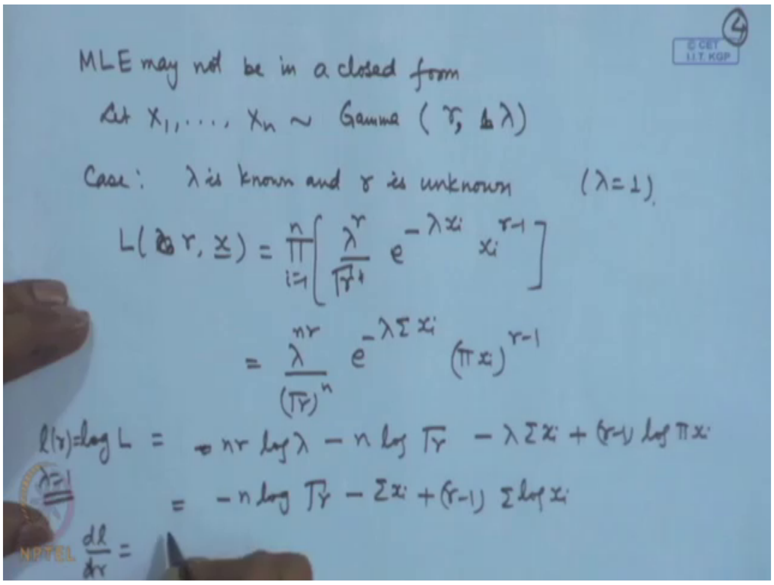
Therefore, we can look at the behavior of the likelihood function as theta varies of course, we can actually plot it here theta 1 hat will be somewhere here because both the terms are negative here, minus x bar of course, minus x bar could be because x bar can be negative or positive; so, but this term is certainly negative and it is bigger. So, naturally I think this will become negative whereas, theta 2 hat is going to be positive.

The function is decreasing before theta less than theta 1 hat. So, something like this and then it will between theta 1 hat to theta 2 hat it increases and from theta 2 hat onwards again it will start decreasing. So, you can see that at theta is equal to theta 2 hat, we get a maximizing value. So, we can see that l theta is maximized at theta is equal to theta 2 hat and another point which we notice here that this is actually a positive value and from our model that we have considered here, theta should be positive. So, it is natural that our maximum likelihood estimator conforms to that range here and it is happening here.

So, the maximum likelihood estimator is X bar plus square root X bar square plus 4 alpha 2 by 2. Naturally, you can see that the form of the maximum likelihood estimator is not in a nice analytic form. In fact, you are getting a square root. So, once again taking expectations etcetera checking whether it is unbiased and all those things will be quite complicated. So, the statement that maximum likelihood estimator need not be in a nice analytic form.

We may have even more difficult situation that is we may not be able to solve the likelihood equation. In this case, although solution is coming it is not in a good form, but there may be a situation where we may not be able to solve it explicitly.

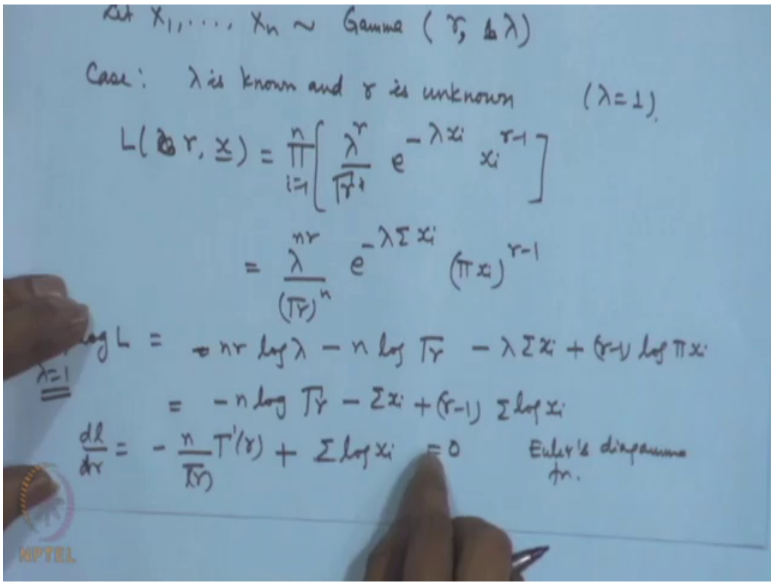
(Refer Slide Time: 15:37)



So, let me give an example of that situation also MLE may not be in a closed form. Let us consider say a random sample from a gamma distribution with parameter say r and lambda. Now, there can be two cases as we have seen earlier; r could be known, lambda may be unknown and r could be unknown and lambda may be known or both may be unknown. Let us consider the case lambda is known and r is unknown. If lambda is known, since it is occurring at a scale parameter we may take it to be 1.

Now, let us consider the likelihood function. So, likelihood function will be a function of r now lambda to the power r by r gamma r e to the power minus lambda x i x i to the power r minus 1 product i is equal to 1 to n. So, this if you take lambda to the power nr by gamma r to the power n e to the power minus lambda sigma xi product xi to the power r minus 1. Let us take the log of this that is equal to n r log lambda plus minus n log gamma r minus lambda sigma x i plus r minus 1 log of product x i. Now, let us put lambda is equal to 1 here, then this term becomes much simpler this particular term vanishes here you get minus n log of gamma r minus sigma x i plus r minus 1 sigma log of xi.

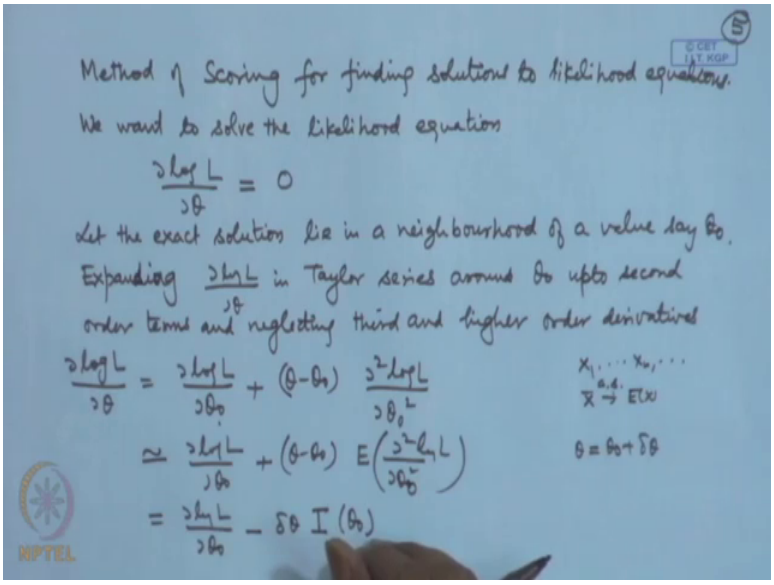
(Refer Slide Time: 18:15)



Now, if we treat it as a function of r, then derivative of this with respect to r will give me minus n by gamma r into gamma prime r. This is known as diagram of function minus this will become 0 plus sigma log of x i. So, this if you put 0 gamma prime r by gamma r, this is known as Euler's diagram function. So, it is not a very nice analytic function and you cannot solve that this is equal to what will be the solution. You will have to use a numerical method such as say Newton Raphson method or any other numerical method to solve this non-linear equation.

Now, in the cases when the explicit solution of the likelihood equation is not possible, a modification to the Newton Raphson method was suggested by Fisher and this is known as the method of scoring, the method of scoring for finding solutions to likelihood equations.

(Refer Slide Time: 19:20)



So, the method is briefly as follows.

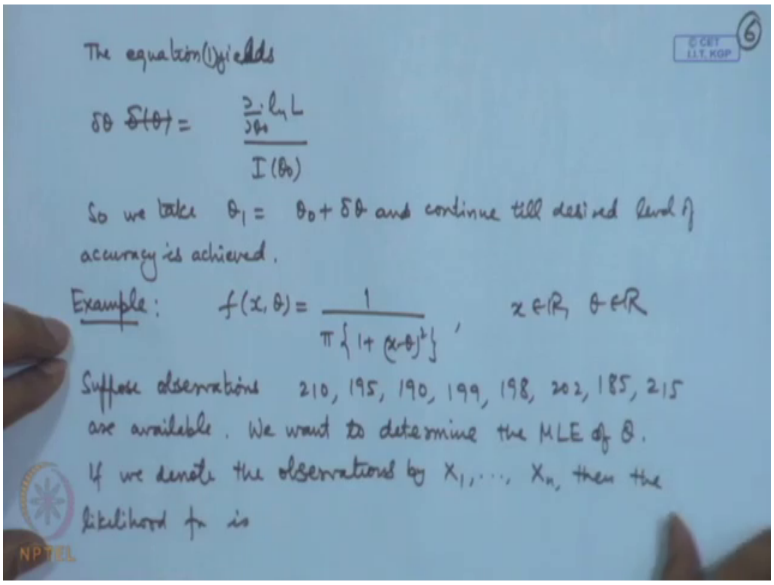
We are actually looking at del by del theta log of L is equal to 0, we are trying to solve this equation. So, in case the solution is existing, there is no problem; however, there may be cases when the solution exists but we are not able to get exact analytic form. So, what we will do, consider we want to solve the likelihood equation del log L by del theta is equal to 0 and suppose the exact solution lies in the neighborhood of theta naught or you can say it lies a so, it could be theta naught or it could we can assume that it is near about theta naught.

Let the exact solution lie in a neighborhood of a value say theta naught. Now, once again here we will use the techniques of analysis to determine. So, for example, if we want to find out the roots of an in general non-linear equation then what do we do? We study the behavior of the function, for example, if we are saying f x is equal to 0, then we look at the behavior of the function we try to locate the roots where they may be lying and then we apply any numerical method. Because generally, the initial approximation is important for example, in Newton Raphson method in one initial approximation is required. If we are using say bisection method, then two initial approximations are required such that both of them are on the either side of the solution.

So, similarly here we guess the initial root say theta naught and let us consider expansion expanding say del log L by del theta in Taylor series around theta naught. Now, once we say Taylor series expansion, we are making the assumption that the derivatives of this exist. So, let us consider only up to second order, up to second order terms and neglecting third and higher order derivatives. So, basically, it means the derivative evaluated at theta naught; similarly, here it means the derivative evaluated at theta naught, the second derivative. Now, this term what Fisher suggested we approximate by; that means, in place of this term we have written expectation. Now, there are certain justifications for this.

For example this likelihood function is the joint distribution. So, when we are taking log it is becoming summation here. So, this becomes summation term here. Now, we know by the laws of large numbers that if I have X 1, X 2 X n a sequence of IID random variables then X bar converges to expectation X bar almost surely that is with probability 1; that means, if n is large enough, this approximation is all right. So, we have replaced this term by it is expectation and this expect expectation of this with a minus sign is known as the Fisher's information. So, this is equal to del log L by del theta naught minus now this theta point we consider in the neighborhood of theta. So, let us write it as delta theta. So, this is delta theta and minus expectation this is called Fisher's information at the point theta naught.

(Refer Slide Time: 25:02)

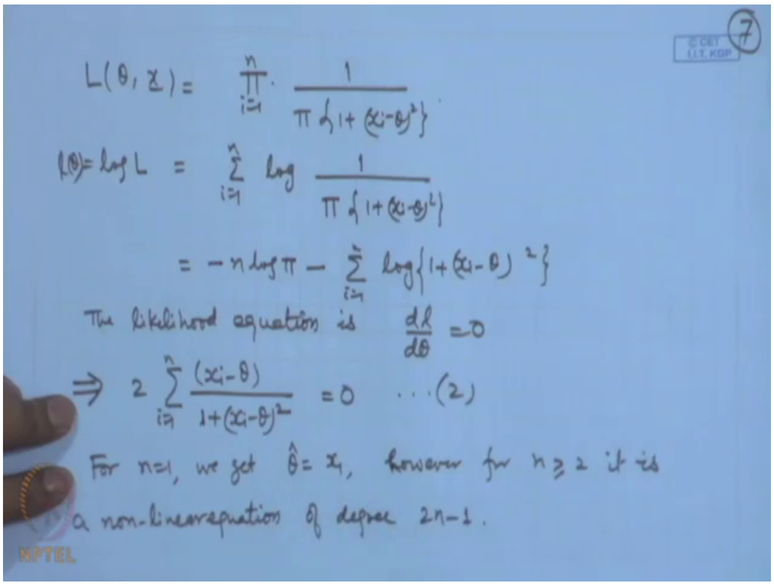


So, now from here what we get, the equation yields let me call it equation number 1; the equation 1 yields delta theta is equal to delta by delta theta naught log L divided by I theta naught because, what we are going to do, we are having delta log L by delta theta is equal to 0; that means, we are putting this term is equal to 0 here. So, if we put this 0 then we can simplify this and we get the first approximation delta theta as del by del theta not log L by I theta naught.

So, we take the next iterate as theta naught plus this delta theta and continue. So, theta 2 will then again become where delta theta will be evaluated at theta 1 and so on. So, continue till desired level of accuracy is achieved. So, this modified method it is known as Fisher Newton Raphson method or Fisher scoring method, I will explain it through one example. Let us consider say we have observations from a Cauchy distribution where x is any real number and theta is any real parameter. Suppose, observations 210, 195, 190, 199, 198, 202, 185 and 215; eight observations are available and we want to determine the maximum likelihood estimator of theta based on this sample.

Now, in general if we write X 1, X 2, X n then what will be the likelihood function in the case of if we denote the observations by say X 1, X 2, X n, then the likelihood function is product i is equal to 1 to n 1 by pi 1 plus x i minus theta square.

(Refer Slide Time: 28:10)

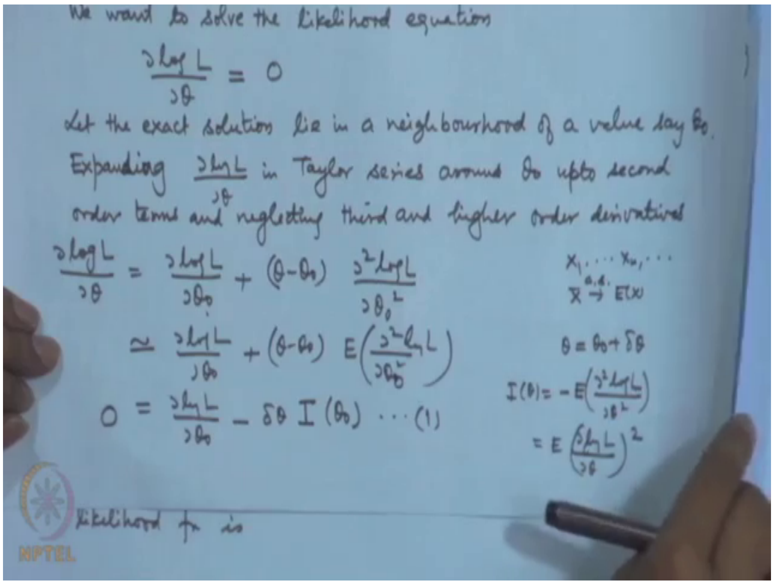


So, if you take log of this, we will get sigma log of 1 by pi 1 plus x i minus theta square which we can write as minus n log pi minus sigma log of 1 plus xi minus theta whole square. So, if we look at the likelihood equation, d l by d theta is equal to 0, this is equivalent to this term will yield 0 if you differentiate and here you will get 1 by 1 plus x i minus theta square and then derivative of that that will be 2 times x i minus theta with a minus sign. So, we will get twice sigma xi minus theta divided by 1 plus xi minus theta square i is equal to 1 to n.

Now, naturally you can see here this equation you cannot solve for n greater than or equal to 2 for n is equal to 1 this will give simply theta is equal to x 1 for n is equal to 1, we get theta hat is equal to x 1. However, for n greater than or equal to 2, it is a non-linear equation. In fact, even if you write two terms here, then you will get x 1 minus theta by 1 plus x 1 minus theta square plus x 2 minus theta divided by 1 plus x 2 minus theta square and obviously, that equation will be having terms up to theta cube in the numerator. So, in general if I am writing n terms here then n in each of the terms you will get a square in the denominator. So, if you multiply n minus 1 of them, you will get 2 n minus 2 and then numerator.

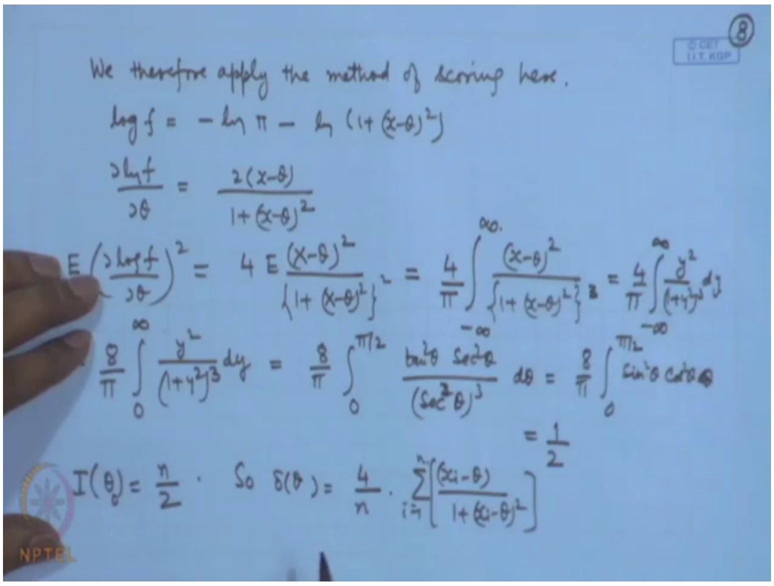
So, this will give me a non-linear equation of degree 2 n minus 1. So, naturally we cannot solve this theoretically. Let us apply the method of scoring in this problem. Now, method of scoring involves as we have seen just now that we should calculate the term called I theta naught. I theta naught is obtained as minus expectation del 2 log L by del theta naught square.

(Refer Slide Time: 31:37)



And that is also equal to let me write it here I theta is equal to minus expectation del 2 log L by del theta square. This is also equal to expectation of del log L by del theta whole square so, one can do it in either way. Let us look at the calculation for this part.

(Refer Slide Time: 32:09)



So, we therefore, apply the method of scoring here. Now, for Cauchy distribution log of f is equal to minus log of pi minus log of 1 plus x minus theta square. So, del log f by del theta is equal to twice x minus theta divided by 1 plus x minus theta square. So, expectation of del log f by del theta whole square that is equal to 4 times expectation of x minus theta square divided by 1 plus x minus theta square whole square. So, we evaluate this that is equal to 4 times integral x minus theta square divided by 1 plus x minus theta whole square and whole square of that multiplied by the density function of x and the density function of that is 1 by pi 1 plus x minus theta square. So, you will get power 3 here and 1 by pi I will write here this is from minus infinity to infinity.

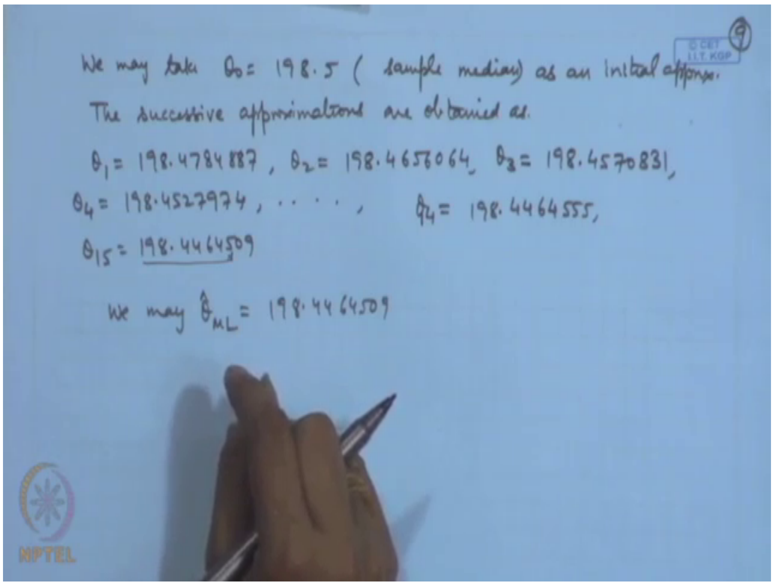
Now, you can easily transform this by putting x minus theta is equal to y. So, you get this as equal to 4 by pi y square divided by 1 plus y square whole cube d y; easily you can see that this is an even function. So, this becomes 8 by pi 0 to infinity y square divided by 1 plus y square whole cube d y.

Now, this type of integral is standard we can substitute something like y is equal to tan theta. So, this will give me 8 by pi 0 to pi by 2 tan square theta sec square theta divided by sec cube theta sec square theta whole cube d theta and that is equal to 8 by pi 0 to pi by 2 sin square theta cos square theta d theta and that is equal to half.

So, I, so, this you can see it is free from theta, so information at the point theta naught that will become n by 2. Now, the function that you need to calculate for the scoring method is delta theta. Delta theta is equal to del by del theta naught log of theta divided by I of theta naught. So, if we look at this term here delta theta that will be equal to 4 by n sigma x i minus theta divided by 1 plus x i minus theta whole square I is equal to 1 to n.

Now, the question is that what should be the initial approximation now in the Cauchy distribution. The sample mean is inconsistent because we have seen the distribution of the sample mean is the same of same as that of the initial observation x i each x i; however, we can see that sample median will be a consistent estimator here. Now, from the given data of this the middle observation will turn out to be two middle observations are there that is 198 and 199 because there are if you arrange it in the ascending or descending order, then these are the two middle observations. So, the midpoint of that can be considered as the initial approximation.

(Refer Slide Time: 26:33)



So, we may take; we may take theta naught is equal to 198.5 the sample median as an initial approximation. Now, for each theta, so if I take theta naught here 198.5 here, we substitute n is equal to 8 and we have the data available to us x is in the form of these values 210, 195 etcetera. So, if you substitute these values from the initial approximation we can get the successive approximations. The successive approximations are obtained as theta 1 is equal to 198.4784887 you have done up to 7 decimal places; next approximation gives 198.4656064, theta 3 is 198.4570831; theta 4 is equal to 198.4527974 and so on.

If you look at 14, that is 198.4464555 and theta 15 is equal to 198.4464509. So, it is accurate up to five decimal places. So, we may take we may take the solution as 198.4464509. Of course, you can see that this is not much different from the sample median because, the sample median was 198.5. So, this method of a scoring can be applied to various cases whenever we are getting a non-linear equation for which the solution is not in a tractable form.