Statistical Inference

Prof. Somesh Kumar

Department of Mathematics

Indian Institute of Technology, Kharagpur

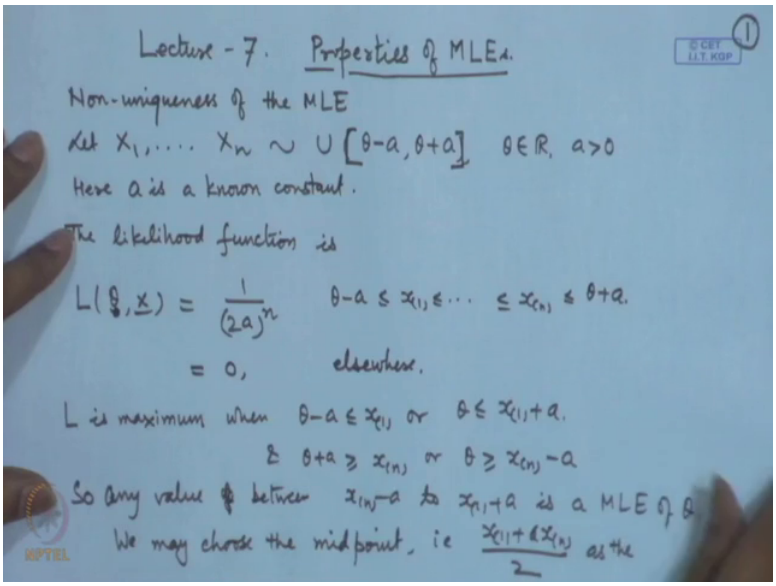
Module No. # 01

Lecture No. # 07

Properties of MLEs

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: 01:07)



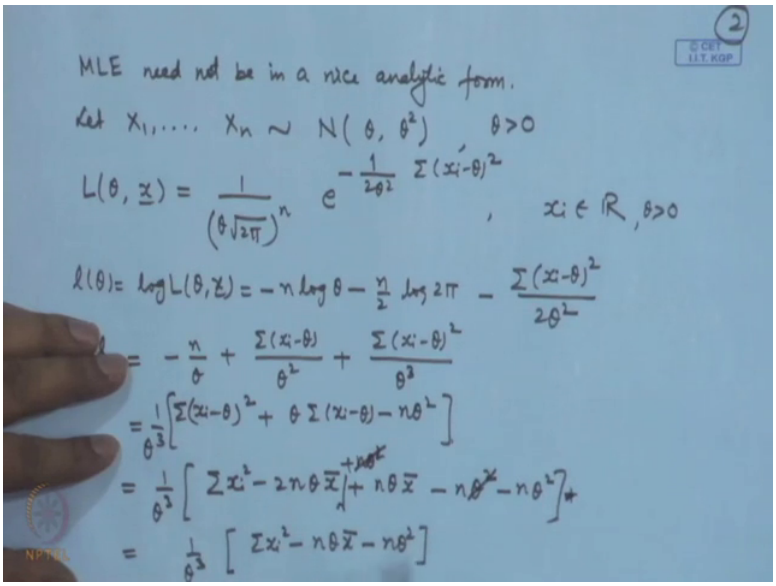
So, let me give an example of that, non-uniqueness of the M L E. Let me discuss one example of that, let we have a sample x 1 x 2 x n, from a uniform distribution on the interval

theta minus a to 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 to theta plus a, therefore we can summarize this information in the form, that theta minus a less 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 close interval, because I am including the end points 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 when 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 M L E. 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 go it as x bar, one 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. Let me take another example where we do not get a nice analytic form.

(Refer Slide Time: 05:32)



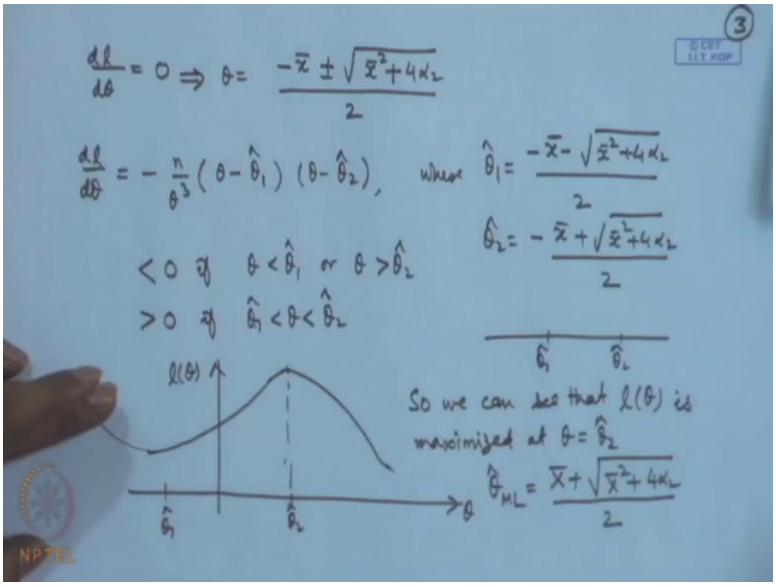
M L E need not be in a nice analytic form. Let me take one example here, let x 1 x 2 x n be a random sample from a normal distribution, with mean theta and variance theta square. So, naturally here theta is a positive parameter here, the likelihood function. So, here your variance is actually the square of the mean, so there is an inter relationship 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 x i 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 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 q, because the derivative of 1 by theta square will be minus 2 by theta cube, so that simplifies it 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 x i 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 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 cancel's 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: 09:59)



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, this 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 straight forwardly 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 head into theta minus theta 2 head. Where I am taking theta 1 head 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 head is equal to minus x bar plus square root x bar square, plus 4 alpha 2 by 2.

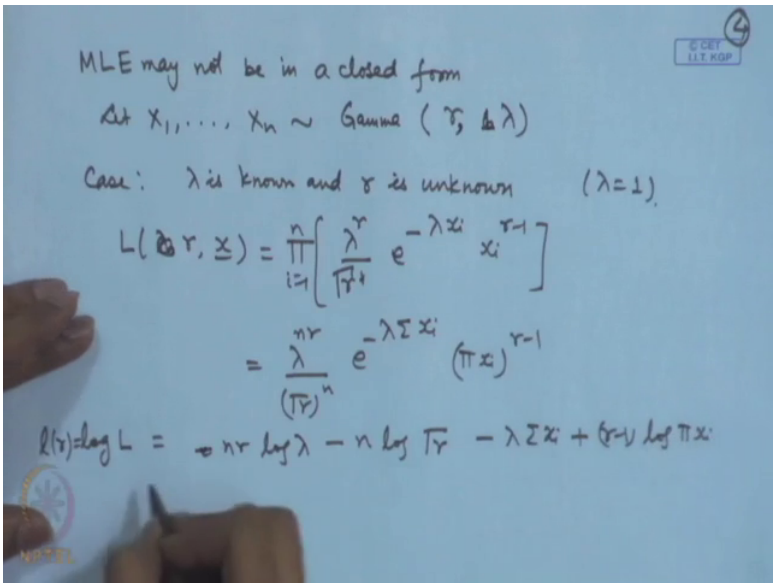
Now let us look at the sign scheme of this. This term will be negative, if theta is less than theta 1 head or theta is greater than theta 2 head. Because if theta is less than theta one head, this term is negative. Here we can see that theta 1 head is less than theta 2 head. So, if we are considering theta 1 head and theta 2 head here. So, if theta is below theta 1 head and theta 2 head, then both of these terms are negative, their product is positive, so this entire term d l by d theta will become negative. Similarly, if theta is greater than theta 2 head, 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 head is less than theta is less than theta 2 head. Therefore, we can look at the behavior of the, likelihood function as theta varies.

Of course, we can actually plot it here; theta 1 head 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, but this term is certainly negative and it is bigger. So, naturally I think this will become negative, whereas theta 2 head is going to be positive. The function is decreasing before theta less than theta 1 head. So, something like this, and then it will between theta 1 head to theta 2 head it increases, and from theta 2 onwards again it will start decreasing. So, you can see that, at theta is equal to theta 2 head we get a maximizing value. So, we can see that, l theta is maximized at theta is equal to theta 2 head. 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 confirms 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 square roots. So, once again taking expectation 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

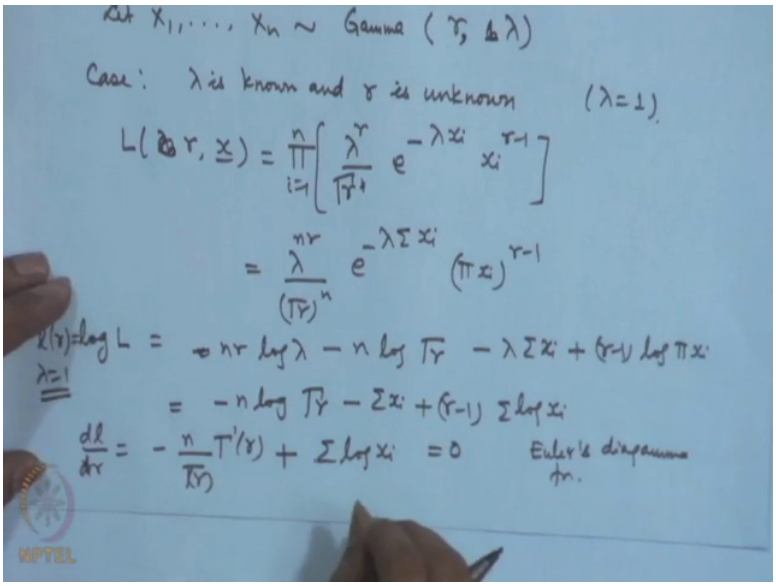
the likelihood equation. In this case although solution is coming it is not in a good form, but there may be a situation, where we may not be able to solve it explicitly. So, let me give an example of that situation also.

(Refer Slide Time: 15:43)



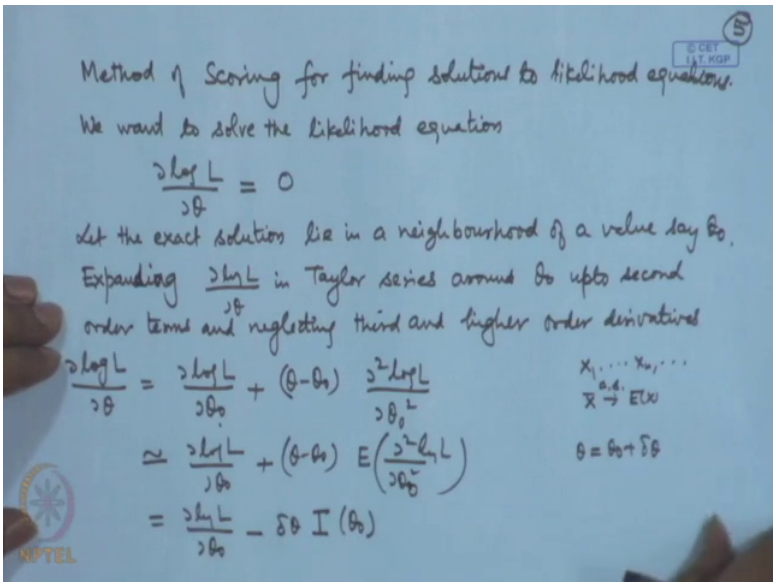
M L E may not be in a closed form. Let us consider, say a random sample from a gamma distribution, with parameter say r and lambda. Now there can be two cases as we have seen earlier. r could be known lambda may be unknown, and r could be unknown and lambda may be known or both may be unknown. Let us consider the case, lambda is known, and r is unknown. If lambda is known since it is occurring as 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 n r by gamma r to the power n, e to the power minus lambda, sigma x i product x i 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 one log of product x i.

(Refer Slide Time: 15:43)



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 x i. 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 digamma function, minus this will become 0 plus sigma log of x i. So, this if we put 0, gamma prime r by gamma r, this is known as Euler's Digamma 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.

(Refer Slide Time: 19:22)



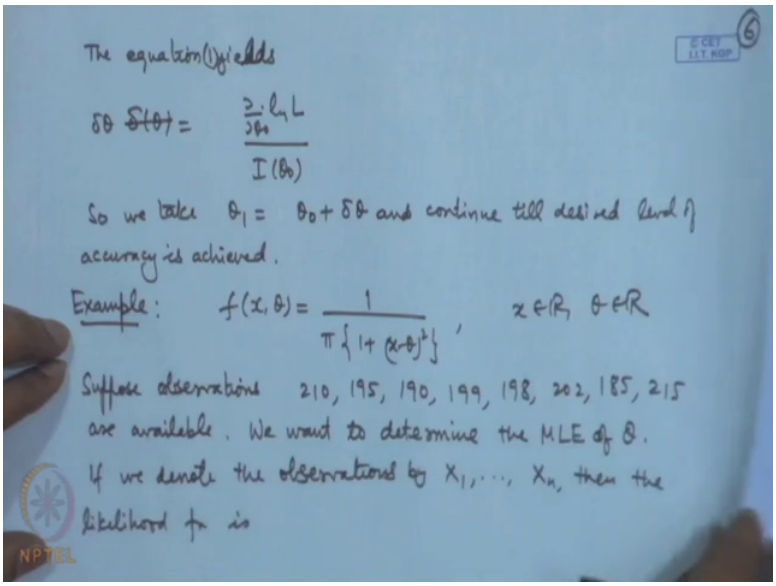
The method of scoring for finding solutions to likelihood equations. So, the method is briefly as follows; we are actually looking at, del by del theta log of l is equal to 0, we are trying to solve this equation. So, in case the solution is existing there is no problem. However, there may be cases when 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, it could be theta naught or 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 a 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, 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 roots, say theta naught. And let us consider 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 i i d random variables then x bar converges to expectation x bar almost surely, that is with probability one; that means, if n is large enough this approximation is all right. So, we have replaced this term by its expectation and this 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.

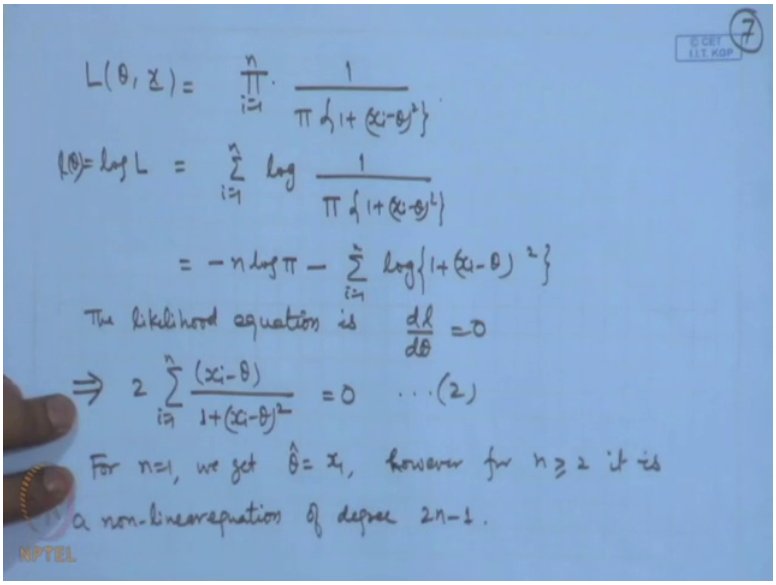
(Refer Slide Time: 25:03)



So, now from here what we get, the equation yields; let me call it equation number one. The equation one 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 0, then we can simplify this and we get the first approximation delta theta, as del by del theta naught log l by i theta

naught. So, we take the next, it rate 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's Newton Raphson method or fisher's 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, 191, 99, 198, 202, 195 and 2 1 5 8 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.

(Refer Slide Time: 28:13)

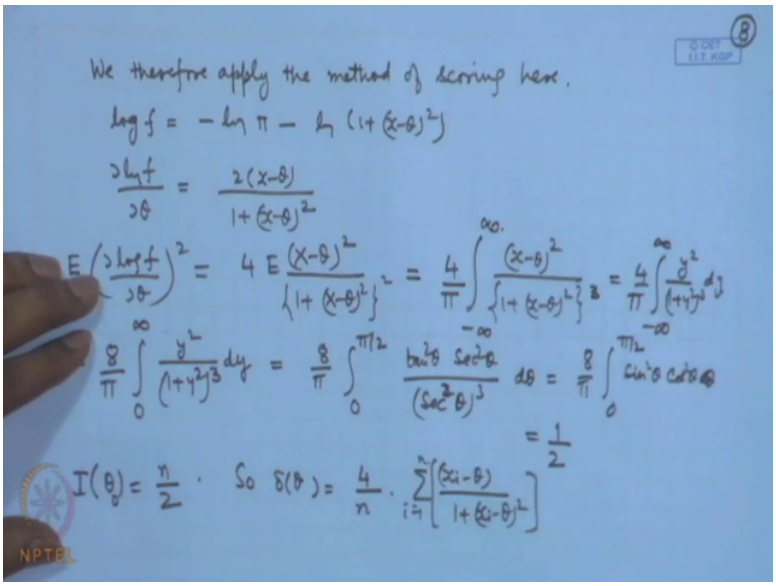


if we denote the observations by say x 1 x 2 x n, then the likelihood function is, product i is equal to 1 to n, 1 by pi 1 plus x i minus theta square. So, if we take log of this, we will get sigma log of 1 by pi, 1 plus x i minus theta square, which we can write as minus n log pi, minus sigma log of 1 plus x 1 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 two times x i minus theta with a minus sign. So, we will get twice sigma x i minus theta, divided by 1 plus x i 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 head 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 one plus x 1 minus theta square, plus x 2 minus theta divided by 1 plus x 2 minus theta square. And; obviously, that equation we 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 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 two log l by del theta naught square. 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: 32:10)



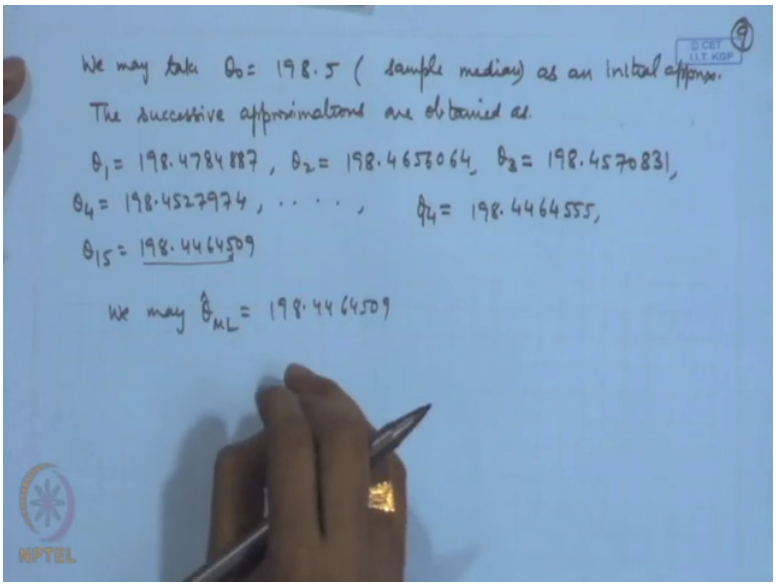
So, we therefore, apply the method of scoring here. Now for cauchy distribution, log of f is equal to minus log of pi, minus log of 1 plus x minus theta square. So, del log f by del theta, is equal to twice x minus theta divided by 1 plus x minus theta square. So, expectation of del log f by del theta whole square; that is equal to four 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 four 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 three 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 four 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 eight 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 eight 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, 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 four 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 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: 36:37)

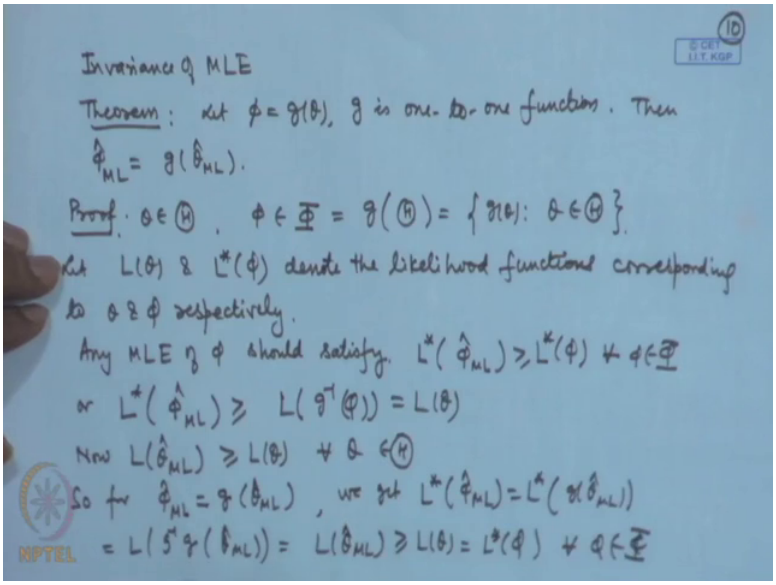


So, we may take theta naught is equal to 198 point 5. The sample median as an initial approximation. Now, for each theta. So, if I take theta naught here 198 point 5, here we substitute n is equal to 8 and we have the data available to us, x i's in the form of these values 210, 195 etcetera. So, if we 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 point 4784887, you have done up to seven decimal places. Next approximation give198 point 4656064, theta 3 is198 point 4570831, theta 4 is equal to 198 point 4527974 and so on. If we look at 14 that is 198 point 4464555, and theta 15 is equal to 198 point 4464509 So, it is accurate up to five decimal places, so we may take the solution as 198 point 4464509. Of course, you can see that, this is not much different from the sample median, because the sample median was 198 point 5.

So, this method of 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. Now there are certain other properties of the maximum likelihood estimators like; invariance which make it very attractive. What is the meaning of invariance. Suppose we are able to obtain as a natural parameters theta, theta one theta two etcetera, say suppose we have a one parameter problem, and we have theta. So, we obtain the M L E of theta, however suppose in the given problem it may be required that, theta square is the quantity of interest, one by theta is quantity of interest, log of theta is a quantity of interest. In that case we can substitute the maximum

likelihood estimator in that function. In general, if we are considering function g theta, then g theta head m l will be the actual M L E of g theta.

(Refer Slide Time: 39:51)

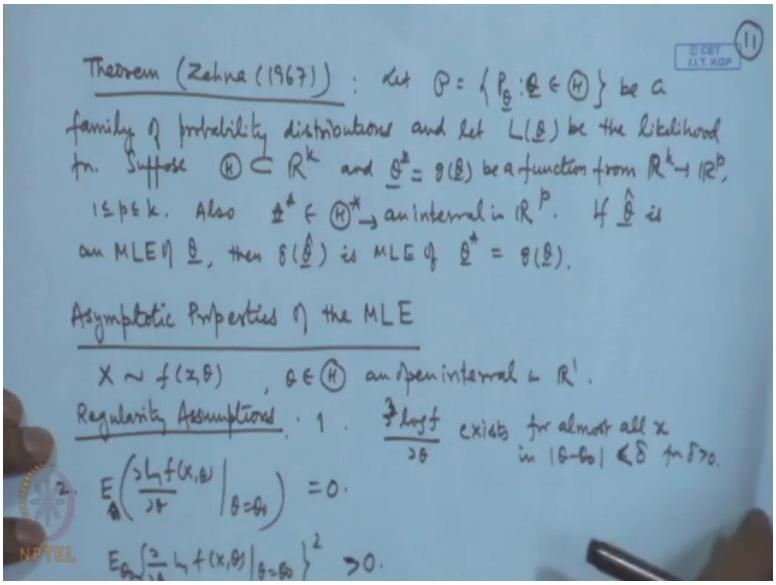


Now, this property I will be proving in two forms; invariance of M L E. Firstly, I will prove it for the one one functions, and then actually I will give the general proof, which is true for any function. Let phi be g theta, where g is a one-to-one function. Then phi head M L is equal to g of theta head M L. Let us look at the proof of this, so suppose my parameter space for theta is capital script theta, and phi we wrote as capital phi and this is actually the g theta space, that is the set of all g theta values as theta varies over script theta. Let L theta and L star phi denote, the likelihood functions corresponding to theta and phi respectively. Essentially they are that same function, because L theta is obtained as the joint distribution, written at the point theta. Now in that you substitute, because g theta is equal to phi. So, if you substitute in terms of phi here in that function, then that will be a function of phi and we denote it by L star. So, they are actually same functions, but written in as functions of different variables.

Now, any maximum likelihood estimator of phi should satisfy, L star pi head M L greater than or equal to L star phi, for all phi belonging to script phi, or L star phi head M M greater than or equal to L g inverse phi, that is equal to L theta. Now l theta head M L, it is greater than or equal to L theta, for all theta belonging to theta script theta. So, for phi head M L equal to g of theta head M L, we get L star phi head M L, is equal to L star g of theta head M L, that is equal to L g inverse g of theta head M L, that is equal to L of theta head M L, which

is greater than or equal to L theta, that is equal to L star phi for all phi. So, we can say that phi head M L is equal to g of theta head M L, is the M L E of phi is equal to g theta. Now naturally this result is true, we have proved for g being a one one function. However, even if we have any function the same invariance property can be used. A justification for this was provided by zehna in 1967.

(Refer Slide Time: 44:04)



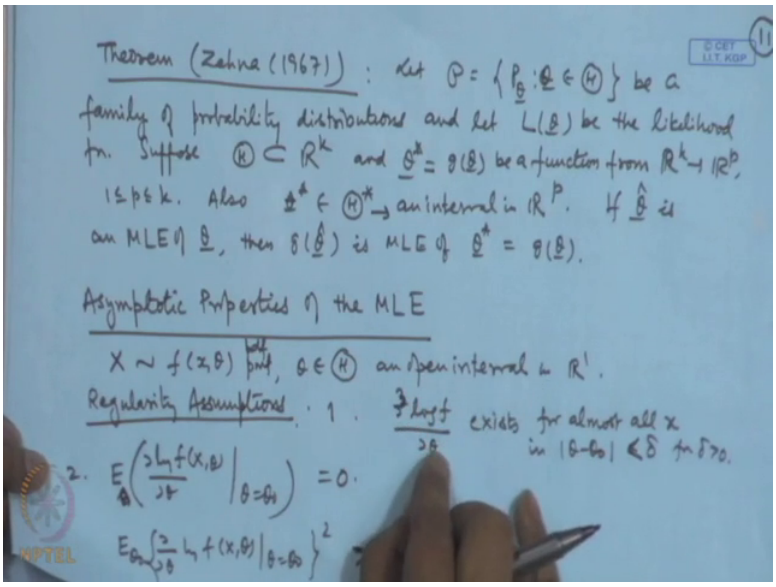
I will state the result without proof here, let p theta belonging to script theta, be a family of probability distributions, and let L theta be the likelihood function. Suppose theta is a subset of k dimensional space, and theta star be a function from r k to say r p, where p will be less than or equal to k. Also we assume that the range of theta star. This is an interval in r p. So, if theta head is an M L E of theta, then g of theta head is M L E of. The proof of this requires that for every value of g theta. If there are several value which lead to the same value, because now the function can be a many one function. Then the likelihood function for theta star is defined as the maximum of all those values. In the case of one-to-one transformation, what we have done is, L theta is L log g inverse phi which we call L star phi. Whereas in the case of many one function, there can be several values corresponding to one value of phi, in this case theta star there can be several values.

So, what we will do, that for all those values we take the maximum of the likelihood function, and then we maximize that. So, when we associate maximum for each inverse image, what will happen is that we are actually creating a one-to-one function, and therefore this theorem

is once again applicable. However, I am skipping the proof here, for the details we can look at the paper by zehna in 1967. I will now give some asymptotic properties of the, maximum likelihood estimators. So, let me call it large sample properties are asymptotic properties of the maximum likelihood estimators. Now these properties are true under certain conditions, which we usually call regularity conditions. Now these conditions were initially given by Cramer, and these are the usually call Cramer–Rao or Frechet-Cramer–Rao regularity conditions.

So, I will just call it regularity conditions. So, in general we are considering a class of probability distributions, now they may have probability densities or probability mass functions. So, let me write, that the density function or the mass function. This is a general notation I am using. Theta belongs to script theta. This is an open interval in real line. Then we have the following regularity assumptions. The assumptions are as follows; that the up to the third order derivative exists, and this should be for all theta. However, it is enough if we assume it in a neighborhood of the solution. Suppose, we know that the solution exist around theta naught, then if we assume this derivative existing in an interval or in a neighborhood of theta naught then it is enough, less than delta for some delta positive.

(Refer Slide Time: 49:55)

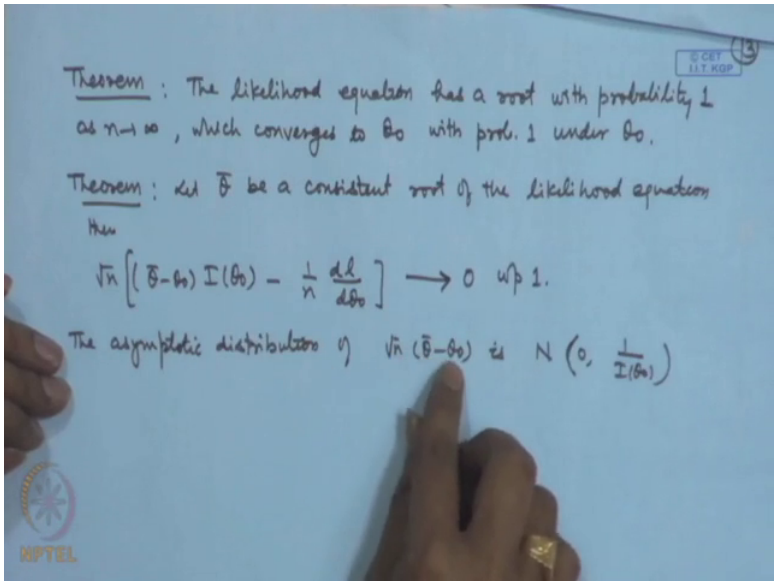


The second condition is that expectation of del log f by del theta, at theta is equal to theta naught, that is equal to 0. Expectation of theta naught del by del theta log of f x theta, at theta is equal to theta naught, whole square that is positive. The third condition is that the third

derivative of the density or the mass function is bounded in a neighborhood of theta naught, and this function itself is having finite expectation, a bounded expectation at any point in the interval theta naught minus delta 2, theta naught plus delta. So, if we are considering x 1 x 2 x n, independent and identically distributed as x, and observed values are. Then the likelihood equation is the log likelihood is. And the likelihood equation is d l by d theta is equal to sigma f prime x i theta. Now, here prime means derivative with respect to theta, this is the likelihood equation. Let us look at the conditions once again, whatever assumptions we have made, here f can be p d f or p m f.

Theta belongs to the parameter space, which is an open interval in the real line. We are assuming the derivative up to the third order exist, and the assumption is at least for an interval in the neighborhood of the solution, and then expectation of the first order derivative at theta is equal to theta naught was vanished. The expectation of first derivative square, that should be positive. In fact, we have defined earlier this as the information function. If we look at this one expectation of del log l by del theta whole square. This is the called Fisher's Information Measure. We will talk more about it somewhat later. Third assumption is that, the third order derivative is bounded by an integrable function. Now, under these conditions we have, under these assumptions we have the following large sample results for the maximum likelihood estimator. I will state it in the form of theorem without proof.

(Refer Slide Time: 53:26)



The first result says, that the likelihood equation has a root with probability one as n tends to infinity, and the root converges to theta naught with probability one under theta naught. So, this says that the likelihood equation has a root with probability one, and the root converges to theta naught with probability one, so this is a very important result. And the second one says that the, let theta bar be a consistent root of the likelihood equation, then square root n theta bar minus theta naught into I theta naught. The information measure we have defined, minus 1 by n d l by d theta naught. This converges to 0 with probability one. Now what does it mean, that square root n theta bar minus theta naught I theta naught minus 1 by n d l by d theta naught. This result is something like, that the root of the likelihood equation that is a maximum likelihood estimator is asymptotically efficient, because what term you are getting theta bar minus theta naught.

See if you take it to this side one by n d l by d theta naught divided by I theta naught. And the asymptotic distribution of root n theta bar minus theta naught is normal 0, 1 by I theta; that is as n becomes large, the distribution of root n theta bar minus theta naught converges to that of a normal distribution with mean 0, and variance one by i theta naught, where i theta naught is the Fisher's Information Measure. The proofs of these results are not difficult. Actually they use the laws of large numbers and the central level theorem at various points. I am skipping the proof here, and what is more important is that this is true under fairly general conditions. For example, the assumptions that I have stated. Now, these assumptions are true for say binomial distribution, say for poisson distribution, say for normal distribution, say for gamma distribution; that means, there is a large class of distributions, particularly the distributions in the exponential family which satisfy these conditions.

Cauchy distribution is not in the exponential family, but even that also satisfies this condition. So, there is a fairly large class of distributions and densities which will actually satisfy this property. So, under fairly general conditions we can say that, the likelihood equation has a solution. The solution is consistent with probability one; that means, it converges to the true value with probability one. Moreover the asymptotic distribution is normal, and it is also second order efficience in the sense of Rao. So, that makes the use of maximum likelihood estimators a fairly important practice in the statistical theory. See this property which I have written at the end, that is square root n theta bar minus theta naught, or say square root n theta bar minus theta naught is asymptotically normal. This is also known as consistent asymptotic normal property or Chain estimator.

So, if an estimator is consistent as well as its asymptotic distribution is normal. So, naturally these are having some desirable properties. We also say best asymptotically normal estimator that is Bian estimators, so Chian Bian. So, under certain conditions such estimator exists. And maximum likelihood estimators are more likely to satisfy these properties. I will be completing this discussion now, and we move over to, once again the unbiased estimators. We have defined unbiased estimators, but I had mentioned that there can be many unbiased estimators. So, which one we should choose. We will introduce that criteria in the next lecture.