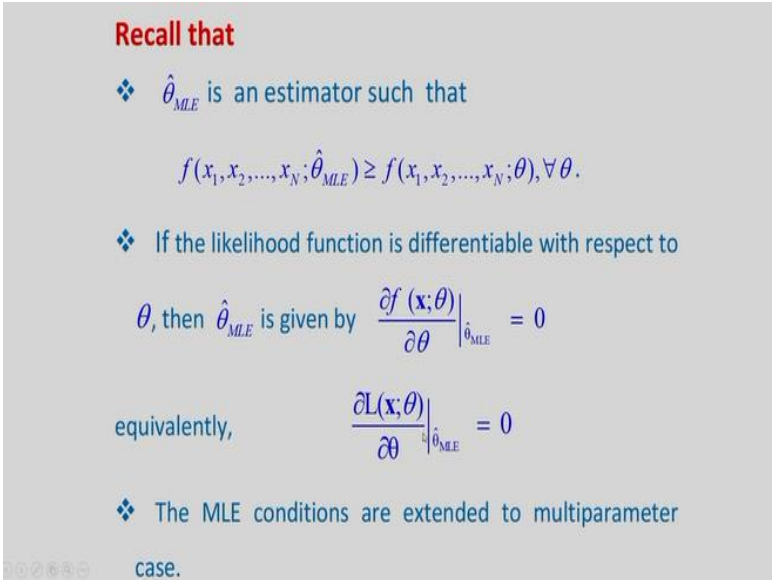
Prof. Prabin Kumar Bora
Department of Electronics & Electrical Engineering
Indian Institute of Technology, Guwahati

Lecture 14
Properties of Maximum Likelihood Estimators (MLE)

Hello students welcome to lecture 14, in this lecture we will discuss properties of maximum likelihood estimators.

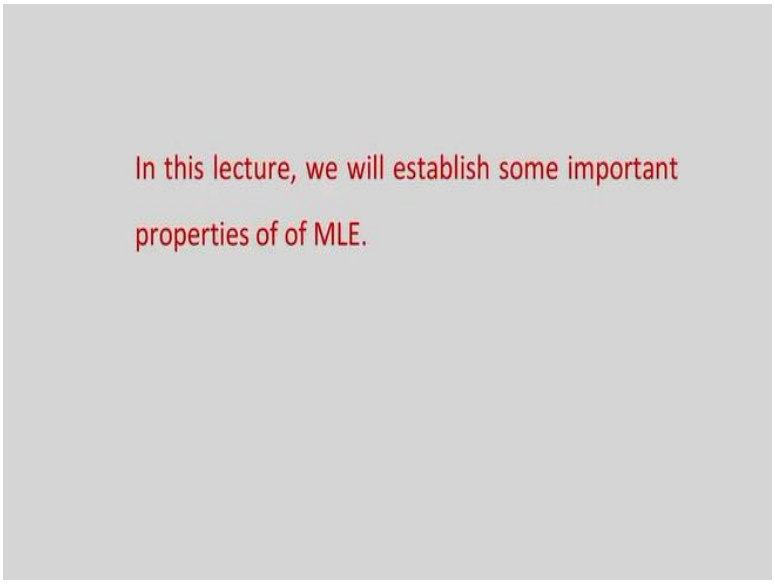
(Refer Slide Time: 00:45)



Recall that, theta hat MLE maximum likelihood estimator is an estimator such that density function we are considering as a function of theta hat MLE is greater than equal to the density function as a function of any other theta, so this is the basic definition and if you likelihood function is differentiable with respect to theta. Then theta hat MLE is given by this expression, del f del theta at theta hat MLE is equal to 0.

Equivalently, we can write in terms of the log likelihood function del L del theta, theta hat MLE is equal to 0. These MLE conditions can be extended to multi-parameter case.

(Refer Slide Time: 01:43)



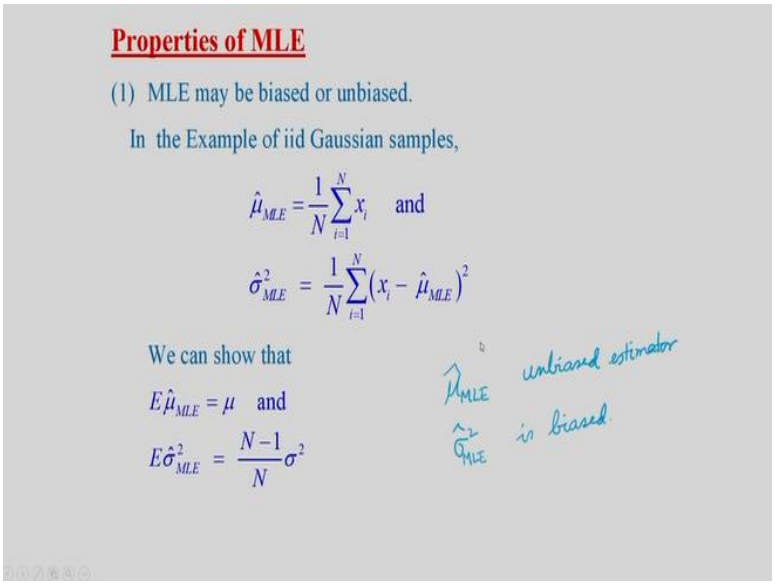
We will now discuss some important properties of MLE.

(Refer Slide Time: 01:46)

MLEs are backed by elegant mathematical theories. They possess desirable properties of good estimators, particularly for large samples. In many situations, MLE turns out to be MVUE minimum variance unbiased estimator. For some distributions, closed form solutions of likelihood equations may not exist. In such cases MLEs may be constructed numerically through iterative algorithms and their properties may be studied.

For example, we made determine the bias-variance etcetera. Some properties of MLEs are described next.

(Refer Slide Time: 02:42)

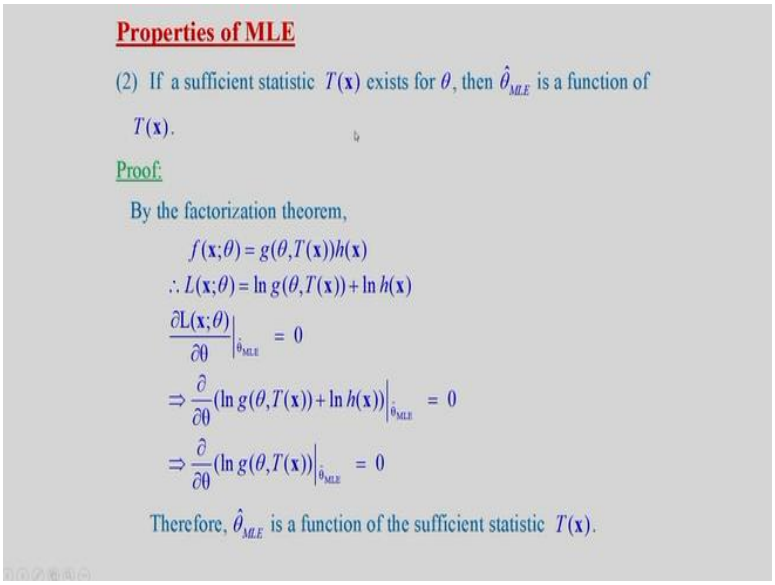


The first property is an MLE maximum likelihood estimator maybe biased or unbiased. In the example of iid Gaussian samples, we establish that the MLE form Mu that is Mu hat MLE is equal to summation xi, i going from 1 to N divided by N and similarly MLE for Sigma square, Sigma hat square MLE that is equal to 1 by N summation i going from 1 to N of xi - Mu hat MLE whole Square.

And now we can show that if I take the expected value of this then right hand side also I have to take the expected value of its xi and xi are iid, so it will have the same mean so therefore we can show that expected value of Mu hat MLE is equal to Mu, that is the true parameter. So that way Mu hat MLE is an unbiased estimator. Now, if I take the expected value of Sigma hat square MLE.

Then I will get same way I can expand this and take the expected value then I can show that this expected value is equal to N - 1 divided by N times Sigma square. So this is not equal to Sigma square, therefore Sigma hat square MLE is a biased estimator. So Mu hat MLE unbiased estimator and Sigma hat squad MLE is biased. So that way N MLE may be biased or unbiased.

(Refer Slide Time: 04:51)

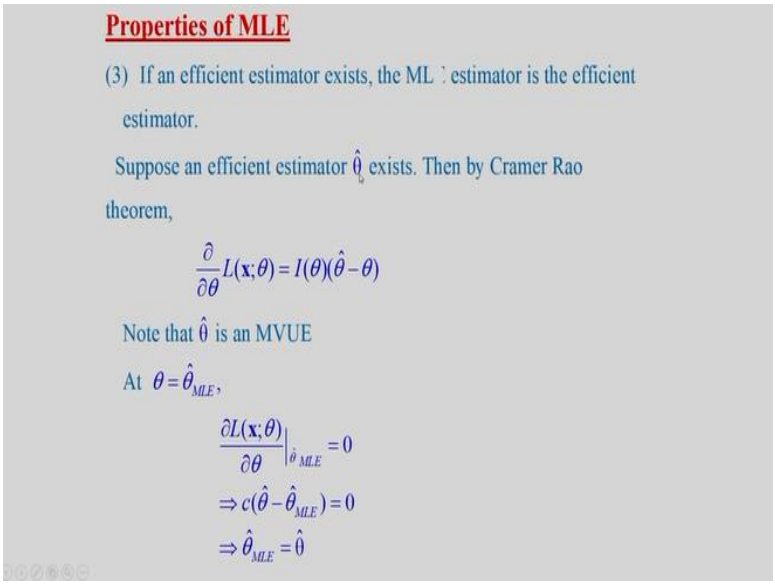


Next, one important property if a sufficient statistic Tx exists for theta, then theta hat MLE is a function of Tx. So if we know that, if sufficient statistic Tx is there then theta hat MLE must be a function of Tx. This proof is also easy now, by factorization theorem, that likelihood function is product of g of theta Tx into h x, this is the factorization theorem. One part that involved this statistic and theta other part simply functions of x.

Therefore if we take the logarithm, log likelihood function will be log of this part g theta Tx + log of h x. Now, since del L del theta, theta hat MLE is equal to 0, we will get del, del theta of log of g theta Tx + log of h x theta hat MLE, it must be equal to 0 and this part does not involve theta, therefore we will simply get del, del theta of log of g theta Tx at theta hat MLE must be equal to 0.

Since theta hat MLE is the solution of this equation. Therefore theta hat MLE must be a function of sufficient statistic Tx. So this property we established if a sufficient statistic Tx exists for theta, then theta hat MLE is a function of Tx.

(Refer Slide Time: 06:44)



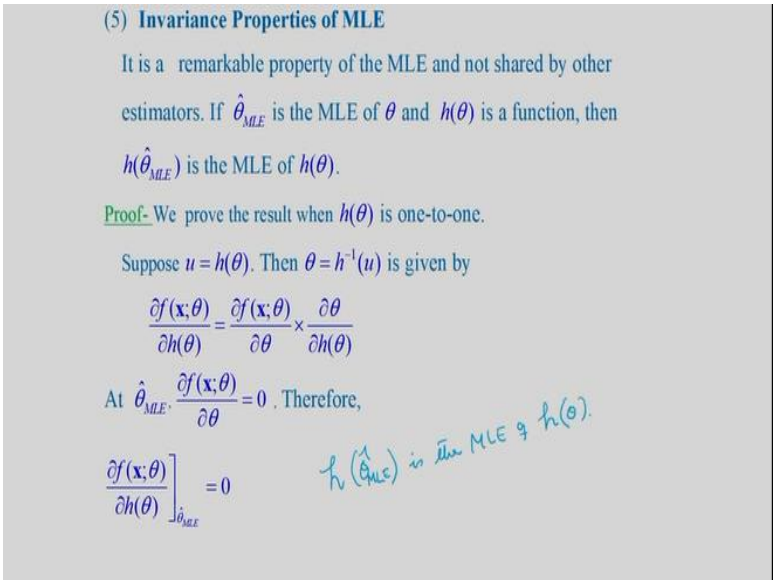
Next property is if an efficient estimator exists, the ML estimator is the efficient one. So this is an important property because we always look for an efficient estimator and if we know that it exists then ML estimator will be the efficient estimator. Now, suppose an efficient estimator theta hat exists, we know that efficient estimator satisfies the Cramer Rao lower bound with equality.

And if we apply the Cramer Rao theorem here, then the condition for efficient estimator is that the partial derivative of the log likelihood function del L del theta must be product of I theta and theta hat - theta. Where theta hat is the efficient estimator, this is the condition we get from the Cramer Rao theorem and in that case theta hat will be d MVUE minimum varience unbiased estimator.

And now let us examine this case suppose one efficient estimator theta hat exists, now that theta is equal to theta hat MLE del L del theta is equal to 0. So it implies that this is a constant C into theta hat - theta hat MLE must be equal to 0 and this implies that theta hat MLE is equal to theta hat. So we have established that, if an efficient estimator theta hat exists then theta hat must be equal to theta hat MLE.

So this is an important property but here also we presume that the efficient estimator theta hat exists.

(Refer Slide Time: 08:44)

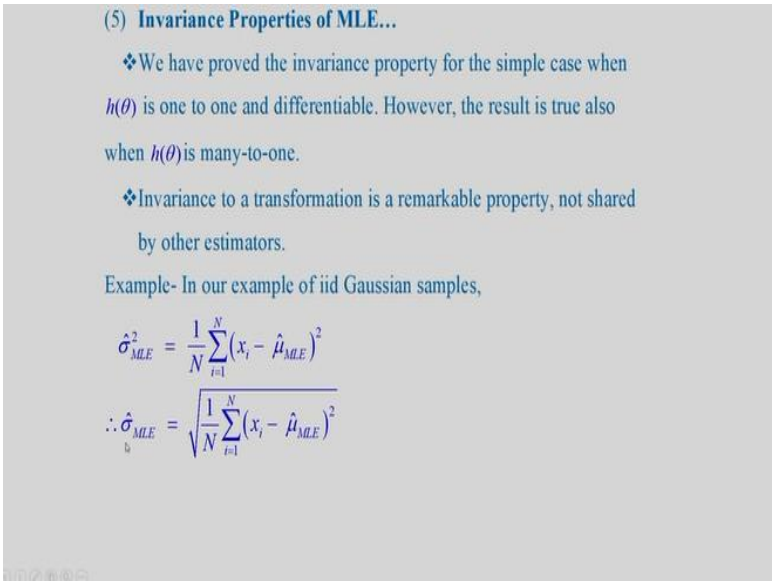


Next property is the invariance properties of MLE. It is a remarkable property of the MLE and not shared by other estimators, so if theta hat MLE is the MLE of theta and h theta is a function, then h theta hat MLE is the MLE of h theta. So here we know suppose theta hat MLE is the MLE of theta. Now, if theta is another function and for this is the maximum likelihood estimator will be a h of theta hat MLE, so that is the invariance.

The under transformation that maximum likelihood point is invariant so whatever theta hat MLE is there at that point corresponding function is theta hat MLE will be the MLE for h theta. We proof to result when h theta is 1 to 1 this is a simple case, suppose u is equal to h theta then theta is equal to h inverse of u, because 1 to 1 it is invertible. Now del f del h theta that will be equal to del f del theta multiplied by del theta del h theta, now this quantity exists.

And this is a invertible function, so we can find out this f theta hat MLE this quantity del f del theta will be equal to 0 and this is usually nonzero. Therefore we will get del f, x, theta del h theta at theta hat MLE is equal to 0. So this will imply that this maximum likelihood estimator for h theta will be at h theta hat MLE, so h theta hat MLE is the MLE of h theta.

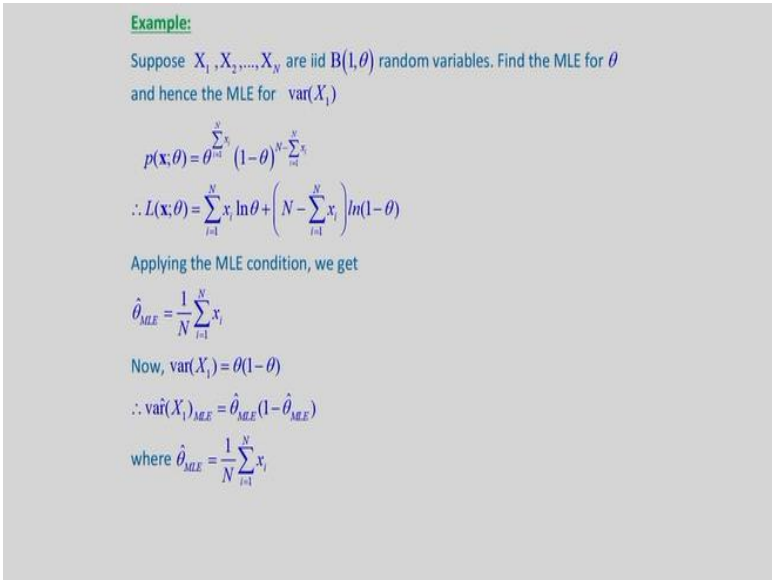
(Refer Slide Time: 10:57)



We have proved the invariance property for the simple case when h theta is one-to-one and differentiable. However the result is true also when h theta is many-to-one. Invariance to a transformation is a remarkable property, not shared by any other estimators. Example in our example of iid Gaussian sample, suppose Sigma hat square MLE that is the MLE for variance is equal to 1 by N summation i going from 1 to N of xi - Mu hat MLE whole square.

Therefore what will be the Sigma MLE, so that is the MLE for standard deviation that is Sigma hat MLE will be given by square root of this MLE expression, therefore this will be the square root of 1 by N summation xi - Mu hat MLE whole square, i going from 1 to N. So this is the MLE for standard deviation. Once we know the MLE for variance we can find out the MLE for standard deviation using the invariance property.

(Refer Slide Time: 12:15)



We will consider another example suppose Xi’s are iid we wanted this is the symbol for Bernoulli random variable with success parameter theta, so Xi’s are B, 1, theta random variables. Find the MLE for theta and hence the MLE for variance of X1. So here Xi's are iid Bernoulli. Now p of x theta, x is the vector x1, x2 up to xN. So P of x theta will be theta to the power the number of success is summation xi, i going from 1 to N, because xi takes value either want for success or 0 for failure.

So that way summation xi will be the number of success. so theta to the power of summation xi, i going from 1 to N into 1 - theta to the power N - summation xi, i going from 1 to N, this is the likelihood function. So if we take the logarithm then we will get the log likelihood function and this will be given by summation xi, i going from 1 to N times log of theta + N - summation xi, i going from 1 to N times log of 1 - theta.

So now we can take the partial derivative with respect to theta and make it equal to 0 and we will get theta hat MLE is equal to 1 by N into summation xi, i going from 1 to N. So this is the theta hat MLE for this parameter theta in the Bernoulli random variable. Now variance of X1 is equal to theta into 1 - theta, which is a function of theta. You can consider any variable X1, X2 or any Xi because they are iid.

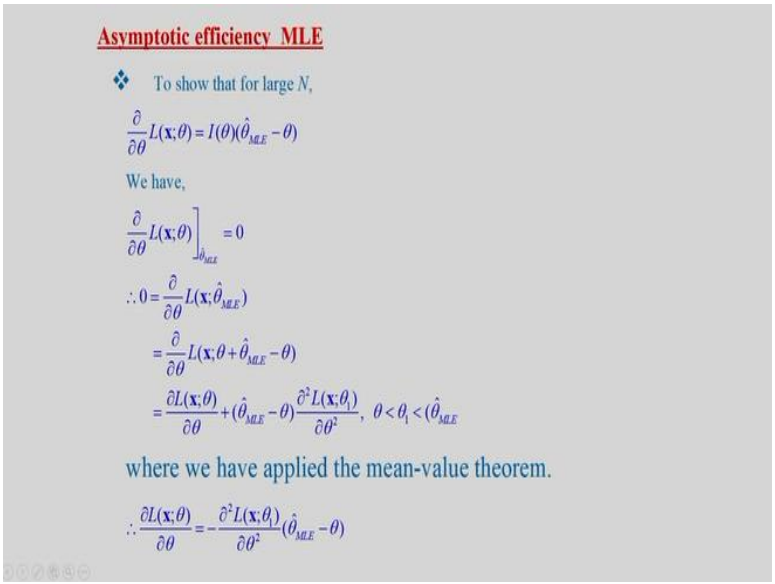
Therefore we can find out the MLE for variance of X1, that is variance of X1 MLE, that is equal to we will substitute theta hat MLE, so theta hat MLE into 1 - theta hat MLE were theta hat MLE is given by the likelihood estimator of theta given by this expression. So that way we can change the application of invariance property of maximum likelihood estimator and this have many uses in signal processing.

(Refer Slide Time: 14:56)

Now we will discuss large simple properties of MLE; we saw that if an efficient estimator exists, the MLE is the efficient estimator. Another attractive feature of MLE is that its behavior becomes better and better with more number of samples. So that will establish now. The asymptotic properties of MLE holds under the regularity conditions like those applied in deriving Cramer Rao theorem, we will not discuss about those regularity conditions but these conditions are usually satisfied in practice.

The MLE is asymptotically unbiased and efficient this is a remarkable property, thus for large and MLE is approximately efficient. asymptotic property means as N tends to infinity therefore when we have large N, the MLE is approximately efficient. We will explain the asymptotic efficiency property.

(Refer Slide Time: 16:03)

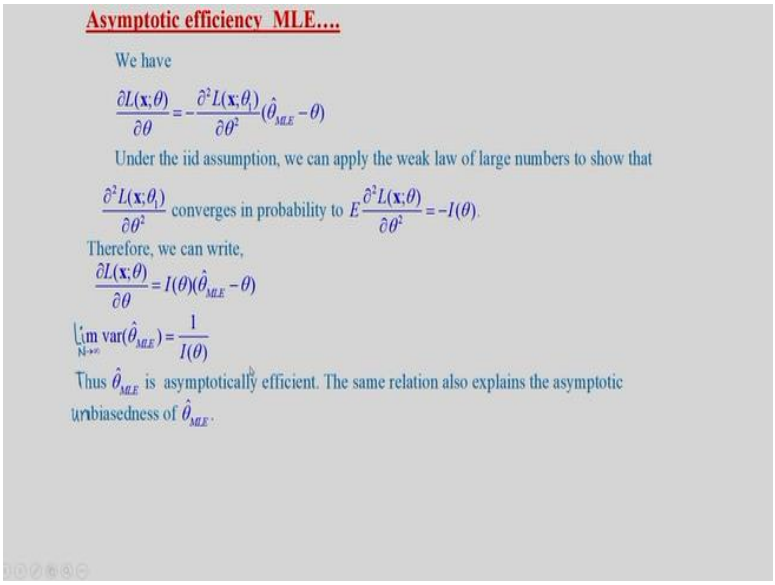


We have to show that for large N, del del theta of L x, theta is equal to I theta into theta hat MLE - theta, this is the one we want to solve. If we can show this, then this theta MLE will be asymptotically efficient. Now we will start with the MLE condition del L del theta at theta hat MLE is equal to 0, therefore 0 will be equal to del del theta of Lx theta hat MLE. Now this theta hat MLE, we can write like this theta + theta hat MLE - theta where it is the original value of the parameter.

So we will write this theta hat MLE by this manipulation theta + theta hat MLE - theta. Now we can apply the mean value theorem to this expression, so we can write this as del L del theta + theta hat MLE - theta into because we are considering this function so derivative of this function will be del 2 L del theta square a point x, theta1 where theta1 lies between theta, theta1, theta hat MLE.

So this expression we get by applying the mean value theorem, so this quantity del L del theta, x, theta hat MLE is same as del x del theta + theta hat MLE - theta into del 2 L del theta square upon theta1, for some theta 1 line between theta, theta1, theta hat MLE. Now a since left-hand side is 0, so we can write del L del theta is equal to minus, if we take it to the other side - del2 L del theta square into theta hat MLE - theta. So we got one expression like this.

(Refer Slide Time: 18:03)



We have established that, del L del theta is equal to - del L del theta squared times theta hat MLE - theta. Now under the iid assumption because we are assuming that Xi's are iid, we can apply the weak law of large numbers to this quantity, actually this quantity can be expanded

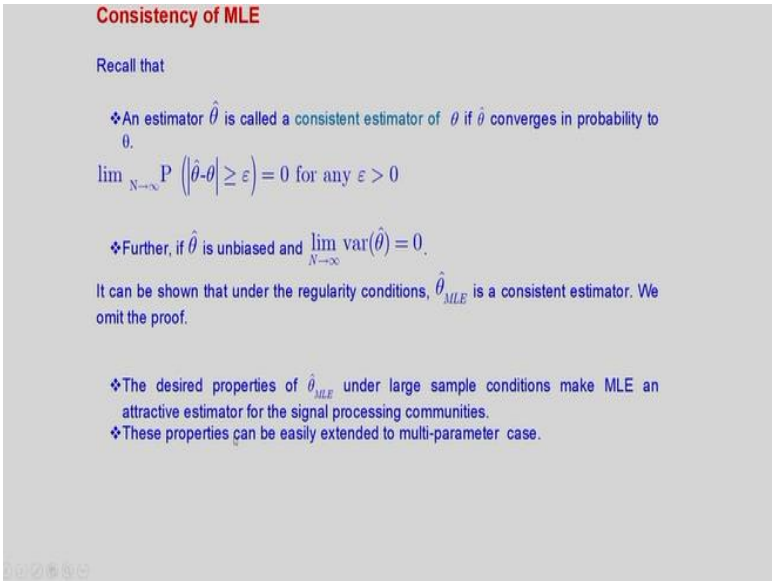
at the sum of individual log likelihood functions. So that way it is a sum of some large number of log likelihood functions.

Therefore we can applied it weak law of large number which means that this quantity del2 L x theta1 del theta square will converge in probability to the corresponding expected value of del2 L x theta del theta square and which is equal to - of I theta. Therefore we can write, del L del theta is equal to I theta into theta hat MLE - theta. So we have established that for large N, when N tends to infinity del L del theta can be expressed as I theta into theta hat MLE - theta, which means that theta hat MLE is the efficient estimator.

And if I consider the Cramer Rao, therefore limit for the variance of theta hat MLE as N tends to infinity will be given by 1 by I theta, where I theta is the Fisher information statistic and it is given by this expression E of del2 L del theta square is equal to - of I theta. Thus theta hat MLE is asymptotically efficient. The same relation also explains the asymptotic unbiasedness of theta ahat MLE, because you see that when this condition is satisfied in that situation theta-hat MLE is the MVUE, minimum variance unbiased estimator.

So this condition implies that this theta hat MLE will be also asymptotically unbiased. So we have established that theta hat MLE is asymptotically efficient and it is asymptotically unbiased.

(Refer Slide Time: 20:25)



Last property we shall consider the consistency of MLE; recall that an estimator theta hat is consistent estimator of theta, if theta hat converges in probability to theta in other words limit

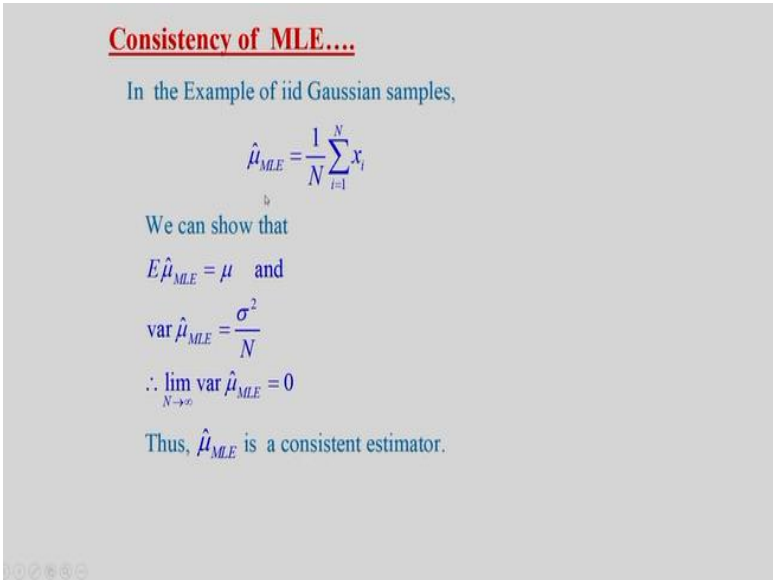
N tends to infinity of the probability that mode of theta hat - theta is greater than equal to epsilon will be always equal to 0 for any epsilon greater than 0. So this is the condition for convergence in probability, therefore with this definition we can take whether theta hat MLE is consistent.

Further, if theta hat is an unbiased estimator than the condition for consistency is this limit of variance of theta hat N tends to infinity is equal to 0. If this happens then theta hat will be a consistent estimator. Now it can be shown that under the regularity conditions theta hat MLE is consistent estimator, we can show that it satisfies this property limit of probability of theta hat - theta greater than equal to any epsilon as n tends to infinity will be equal to 0.

But we will omit this proof but what we claim is that theta hat MLE is a consistent estimator under some regularity conditions but those regularity conditions are satisfied in practice. The desired properties of theta hat MLE under large sample condition make MLE an attractive estimator for the signal processing communities because we have seen that if we consider general properties of theta-hat MLE then they are not very attractive because we have to presume something.

But here these large sample properties assume that the theta-hat MLE is asymptotically efficient and it is a consistent estimator. So these properties also can be extended to multi-parameter cases.

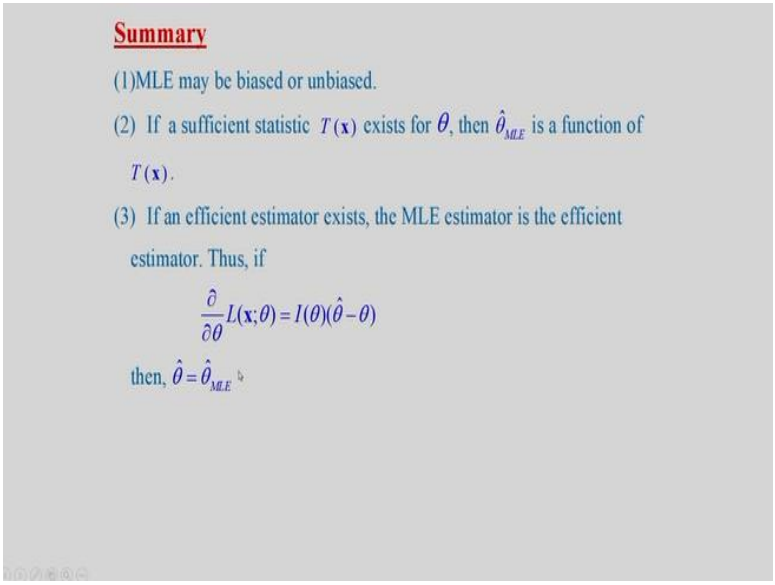
(Refer Slide Time: 22:44)



let us give an example, in the example of iid Gaussian samples; theta Mu hat MLE is given by 1 by N summation xi, i going from 1 to N, this is the MLE for Mu. Now we can show that E of Mu hat MLE is equal to Mu, it is a unbiased estimator and variance of Mu hat MLE is equal to Sigma square by and using the iid property. We can establish this therefore limit of billions of Mu hat MLE will be equal to 0.

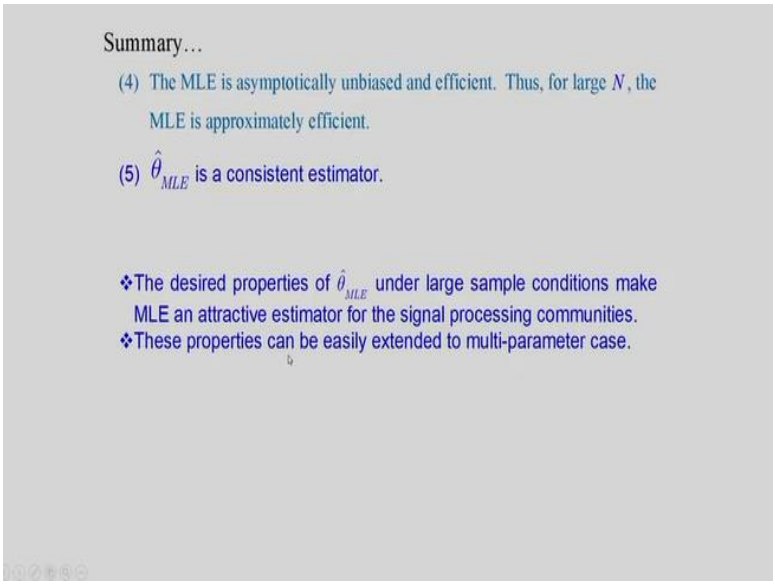
We simply that Mu hat MLE is a consistent estimator, so we have shown that maximum likelihood estimator for Mu is a consistent estimator. This is a general property.

(Refer Slide Time: 23:34)



To summarize MLE may be biased or unbiased. If a sufficient statistic Tx exists for theta, then theta hat MLE is a function of T x. If an efficient estimator exists, then MLE estimator is the efficient estimator. Thus if we can write that del L del theta x, theta equal to I theta into theta hat - theta, then this theta hat must be equal to theta hat MLE.

(Refer Slide Time: 24:08)



Then we establish the large sample properties of MLE, MLE is asymptotically unbiased and efficient. Thus, for a large number of samples MLE satisfy the Cramer Rao lower bound with equality. Thus, for large N, MLE is approximately efficient theta-hat MLE is also a consistent estimate, that we did not proof but this is also a general property of maximum likelihood estimator.

The desired properties of theta hat MLE under large sample conditions make MLE an attractive estimator for signal processing communities. Many signal processing application we use for example there are applications of mean, variance, median, etcetera, where the MLE estimators are used. These properties of MLE can be extended to multi parameter case. Thank you.