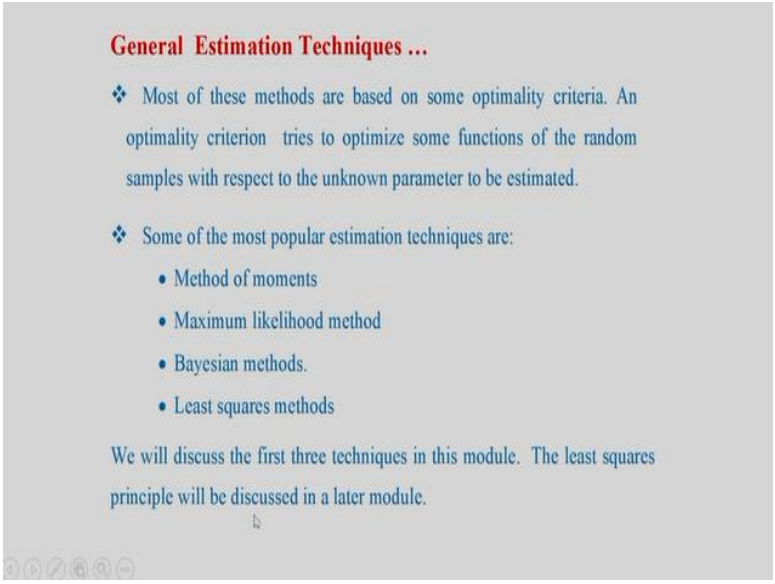
Statistical Signal Processing
Prof. Prabin Kumar Bora
Department of Electronics & Electrical Engineering
Indian Institute of Technology, Guwahati

Lecture 13
Method of Moments and Maximum Likelihood Estimators

Hello students welcome to lecture 13 on methods of moments and maximum likelihood estimators.

(Refer Slide Time: 00:55)



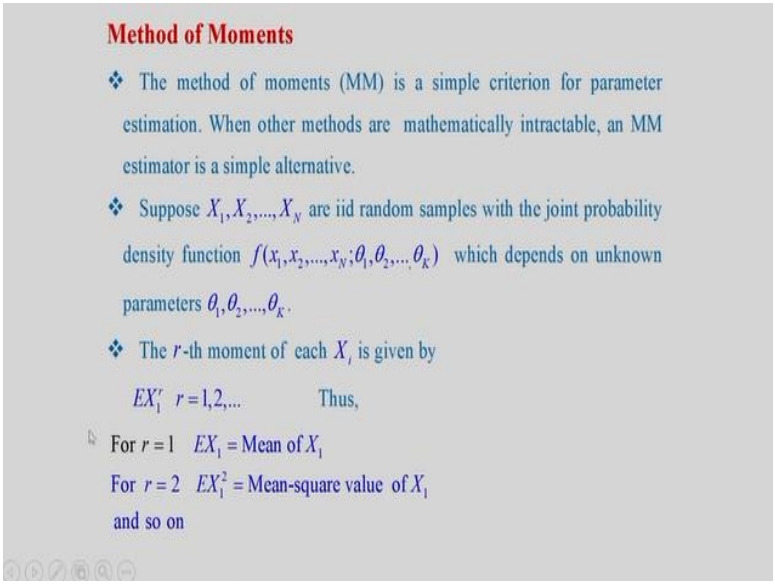
In this lecture we will introduce some general methods for parameter estimation. We saw that MVUE minimum variance unbiased estimator is the most desirable estimator. We discussed two approaches to find the MVUE through CRLB and MVUE through a complete sufficient statistic. The above approaches may not be feasible for practical models of random data. In a practical situation, we have to apply some general techniques to construct a good estimator.

The same technique applied on different probability models may produce different estimation rules. The goodness of an estimator is measured in terms of the desired properties of an estimator like unbiasedness, consistence and efficiency. Most of these methods are based on some optimality criteria. An optimality criterion tries to optimize some functions of the random sample with respect to the unknown parameter to be estimated.

Some of the popular estimation techniques are method of moments, maximum likelihood method, Bayesian methods, least squares methods. We will discuss the first three techniques in this module. We will discuss method of moments, maximum likelihood method and

Bayesian methods. The least squares principle will be discussed in a later module. When we discuss about adaptive filters the time we will introduce least squares estimation method.

(Refer Slide Time: 02:52)

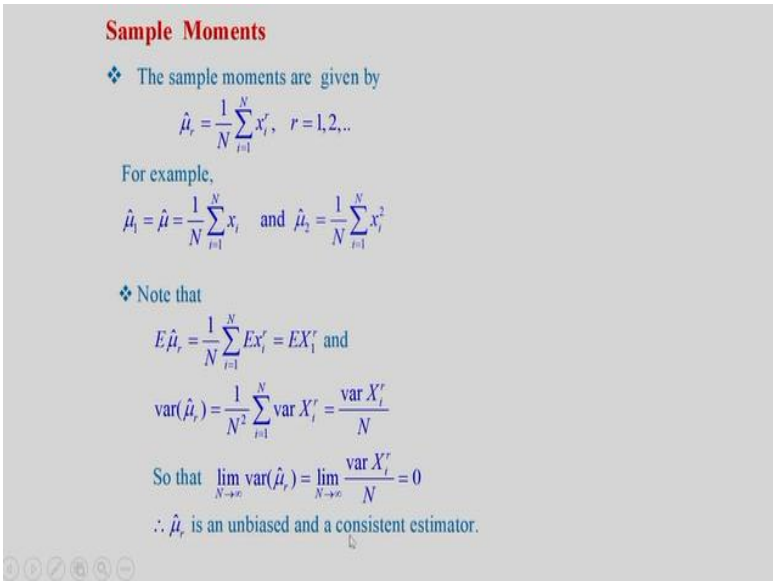


We will start with method of moments; the method of moments abbreviated as MM is a simple criterion for parameter estimation, it is the simplest criterion. When other methods are mathematically intractable, an MM estimator is a simple alternative. First we will discuss what are moments, suppose X1, X2 up to XN are iid random samples with the joint probability density function f x1, x2 up to xN as a function of theta1, theta2 up to theta K.

This density function depends on unknown parameters theta1, theta2 up to theta K, now the r-th movement of Xi is given by E of X1 to the power r, because they are iid any random variable we can consider E of X1 to the power r that is the r-th movement and we can define for r is equal to 1, 2 etcetera any positive integer the loop. So for r is equal to 1, we have E of X1 that is the mean of X1.

Similarly for r is equal to 2 we have E of X1 square that is the mean square value of X1 and so on. So that way we can define the moment of a random variable Xi

(Refer Slide Time: 04:21)

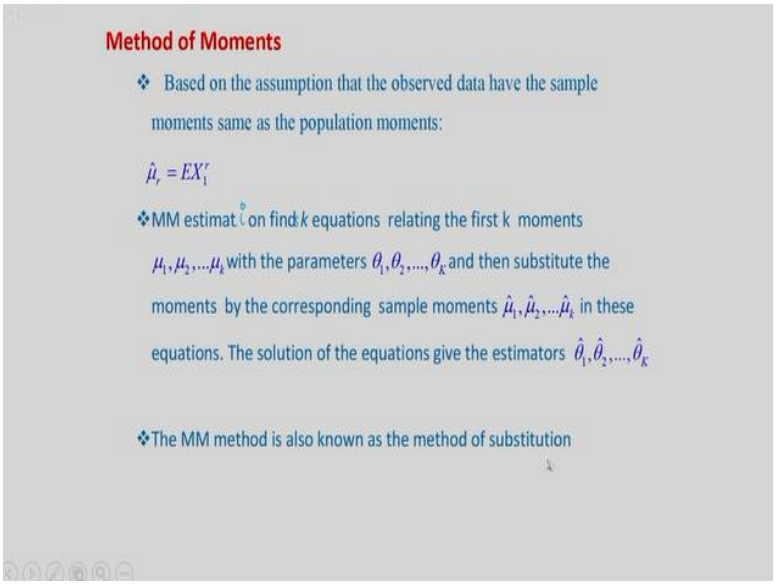


Now sample moments are the moments estimated from data. The sample moments are given by suppose r-th moment, its sample version is Mu r hat and this is given by summation Xi to the power R, i going from 1 to N divided by N and for r equal to 1, 2 etc, so this is the definition of sample moments. For example; when we put r is equal to 1, we have Mu 1 hat that is equal to Mu hat that is the sample average and given by 1 by N summation Xi, i going from 1 to N.

Similarly Mu2 hat that is the estimator for mean square value and it is given by 1 by N summation Xi square, i going from 1 to N. So we have defined moments and the sample moments, moments are also known as population moments. Note that E of Mu r hat that is if I take the expectation, because expectation is a linear person we can take insight. So that way this will be 1 by N into summation i going from 1 to N, E of Xi to the power r.

And all are iid, so all will have the same moment therefore we will have N such terms and divided by N and N will get cancelled, we will get E of Xi to the power r. So this means that Mu r hat is an unbiased estimator. Similarly for this unbiased estimator variance of Mu r hat is given by 1 by N square, summation variance of Xi to the power r, because of the independent property, all cross terms will be 0.

So that way we will have simply variance of Xi to the power r divided by N, so that N tends to infinity this quantity will go down to 0, therefore Mu r hat is also consistent. So what we conclude that Mu r hat that is simple moment is an unbiased and a consistent estimator. **(Refer Slide Time: 07:00)**

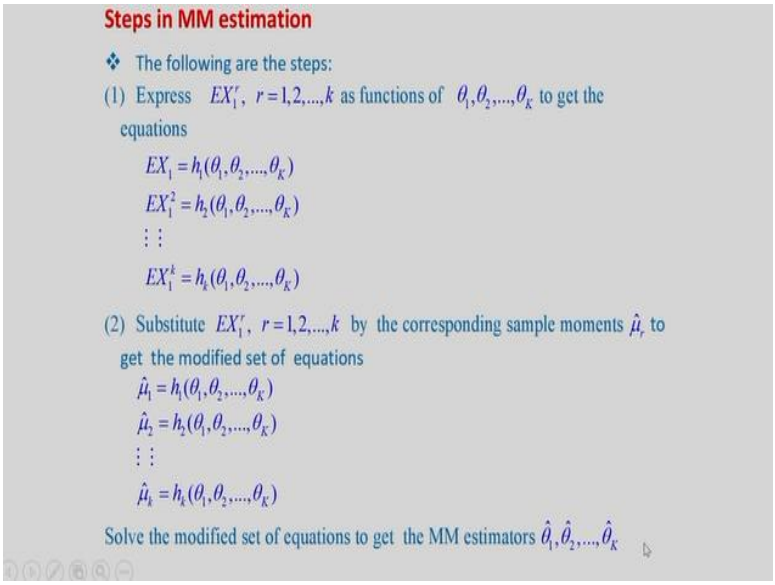


Now we will discuss what is method of moments; it is based on the assumption that the observed data have the sample moments same as the population moments. So whatever sample moments we have that is same as the population moment. So that way Mu r hat is equal to E of X1 to the power r, this is the example we use in method of moments. Now MM estimation finds k equations relating the first k moments Mu1, Mu2 up to Mu k, with the parameters theta1, theta2 up to theta k.

Then it is substitute the moments, these moments by the corresponding sample moments Mu1 hat, Mu 2 hat up to Mu k hat, then we find the solution of the equation the solution of this equation give the estimators theta1 hat, theta 2 hat up to theta k hat. So this is the principal part we will find out the first k moments in terms of the parameters, then substitute the moments by the corresponding sample moments and then solve the equations.

The MM method is also known as the method of substitution because you are substituting the true moments by the corresponding sample moments.

(Refer Slide Time: 08:32)

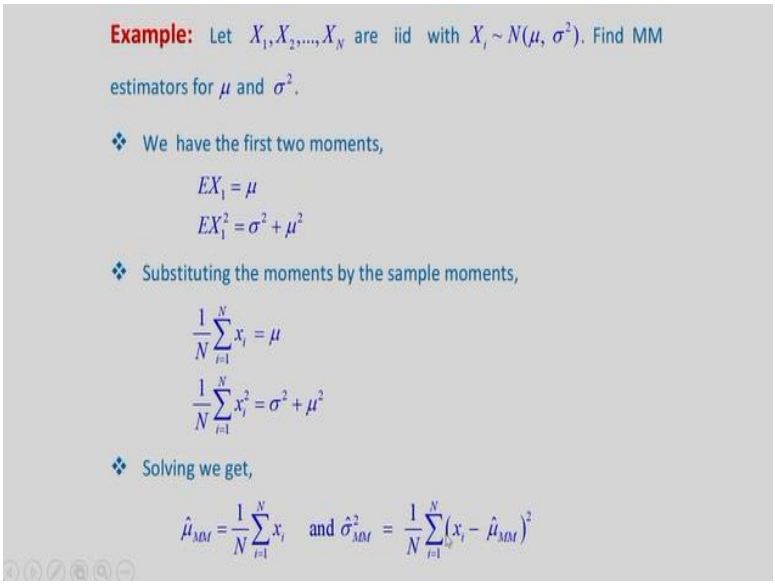


We will describe the method in steps, the following are the steps; first you express, the E of X1 to the power r, r is equal to 1 to up to k as the function of theta1, theta2 up to theta k to get the equation. So these are the equation E of X1 is equal to some function of theta1, theta2 up to theta k. So that way we write E of X1 is equal to h1 theta1, theta 2 up to theta k. Similarly E of X1 square is equal to h2 theta1, theta 2 up to theta k and so on up to E of X1 to the power k that is a function of theta1, theta 2 up to theta k and this function will call as hk.

So that way we have k equations relating the moments with the parameters. Now substitute E of X1 to the power r, that the r-th moment by the corresponding r-th sample moment to get the modified set of equation. Now we will substitute E of X1 by Mu1 hat, E of X1 square by Mu2 hat and so on up to knee up X1 k we will substitute by Mu k hat. Now again we have k equation but this time it is in terms of the sample moments.

Solve the modified set of equation to get the MM estimators theta1 hat, theta2 hat up to theta k hat. So this set of equation there are k unknowns, this k unknowns are the theta1 hat, theta2 hat up to theta k hat. So we can solve this set of equations to get the MM estimators.

(Refer Slide Time: 10:28)

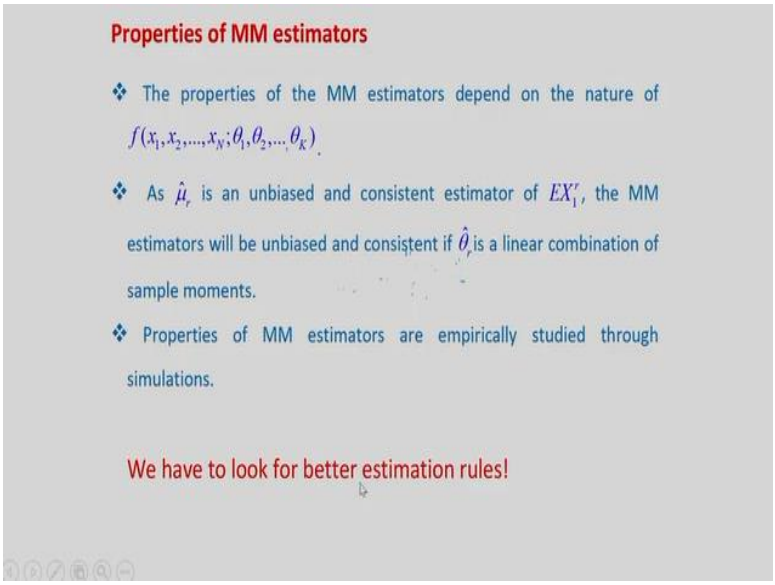


We shall consider one example let X1, X2 up to XN are iid independent and identically distributed with each Xi is normally distributed with mean Mu and variance Sigma square. Find MM estimators for Mu and Sigma square. We have the first two moments for normal distribution, E of X1 is equal to Mu and E of X1 square is equal to Sigma square + Mu square. Now we substitute the moments E of X1 and E of X1 square by the corresponding sample moments.

Therefore what we will get 1 by N summation xi, i going from 1 to N that will be equal to Mu, 1 by N summation Xi square, i going from 1 to N that is equal to Sigma square + Mu square. So if we solve this 2 equation in terms of Mu and Sigma square we will get, Mu hat MM is equal to 1 by N summation Xi, i going from 1 to N and Sigma hat square MM that will be equal to 1 by N summation Xi - Mu hat MM whole square, i going from 1 to N.

So that way we got the estimators for mean that is Mu and the Sigma square. So this is the estimator for mean, this is the estimator for Sigma square.

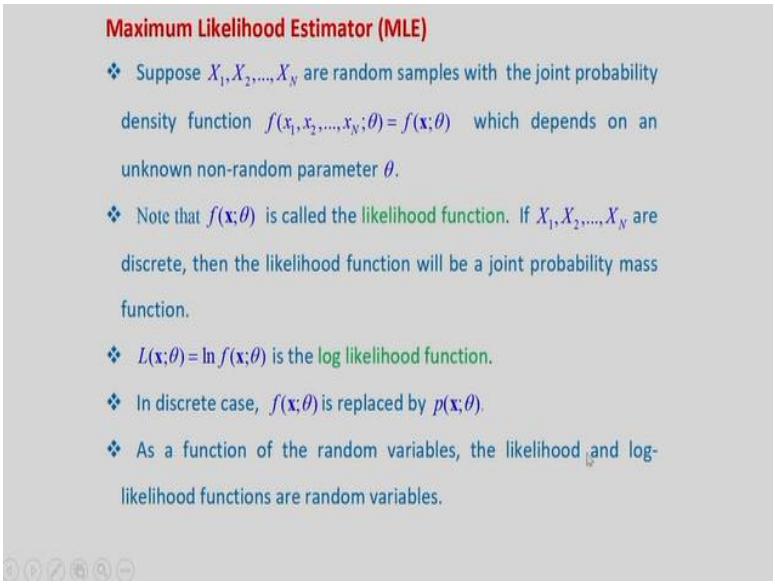
(Refer Slide Time: 12:00)



Let us tell briefly about the properties of MM estimators; the properties of MM estimators depend on the measure of this distribution. So for each distribution we will have different properties. As Mu r hat is an unbiased and a consistent estimator of E of X1 to the power r that we have already established the MM estimators will be unbiased and consistent if theta r hat is a linear combination of sample moments, this is an observation

Usually properties of MM estimators are empirically studied through simulation because the general properties are not obvious we apply numerical simulation to find out this suppose for example bias variance etc and then establish whether it is a consistent estimator, whether it is an unbiased estimator etc. So MM estimator is a very simple estimator it was introduced by Pearson as Bacchus 1901 and it is widely used also but it may not have the good properties of an estimator. So we have to look for a better estimator

(Refer Slide Time: 13:28)

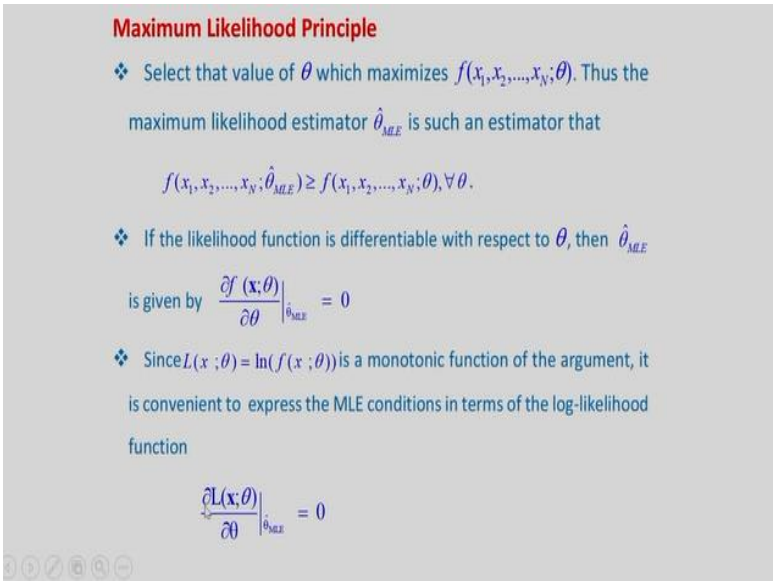


Such a rule is the maximum likelihood estimator; suppose X1, X2 up to XN are random samples with the joint probability density function f of x1, x2 up to xN as a function of theta and this we write as f of x vector as a function of theta. So this is the probability density function and it depends on an unknown non random parameter theta. Suppose we are considering only in the case of simple parameter case, then it depends on an unknown non-random parameter theta.

And note that this f x, theta is called a likelihood function. If x1, x2 up to xN are discrete, then the likelihood function will be a joint probability mass function that also we know. Now L of x, theta that is equal to log of f x, theta is the log likelihood function. In discrete case, f x, theta is replaced by p of x, theta in this expression to find out this a log likelihood function. Now at f x, theta and p x, theta they are function of x and when this x are varying.

We can consider this to be a function of random variable similarly this to be a function of random variables. Therefore the likelihood and log likelihood functions are also random variables, when we consider for varying x.

(Refer Slide Time: 15:06)

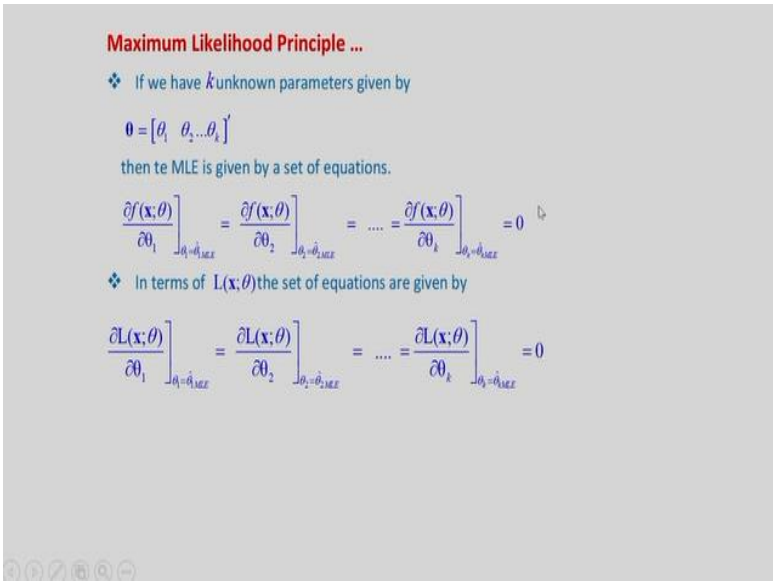


Now we will state the maximum likelihood principle according to this principle select that value of theta which maximizes the likelihood function. Thus maximum likelihood estimator will denoted by theta hat MLE is such an estimator that likelihood at theta hat MLE is greater than equal to the likelihood xN, theta, this is true for all theta. Therefore this is the maximum likelihood principle that likelihood of theta hat MLE is greater than equal to the likelihood of any other theta.

If the likelihood function is differentiable with respect to theta and then theta hat MLE is given by this relation that is partial derivative of likelihood function with respect to theta at theta hat MLE is equal to 0 and now log function is a monotonic function of the argument, therefore this function is also monotonic of this f x, theta. Therefore it is convenient to express the MLE condition in terms of log likelihood function as this.

That is partial derivative of Lx, theta with respect to theta at theta hat MLE is equal to 0. So either we can use this condition or this condition to find out theta hat MLE.

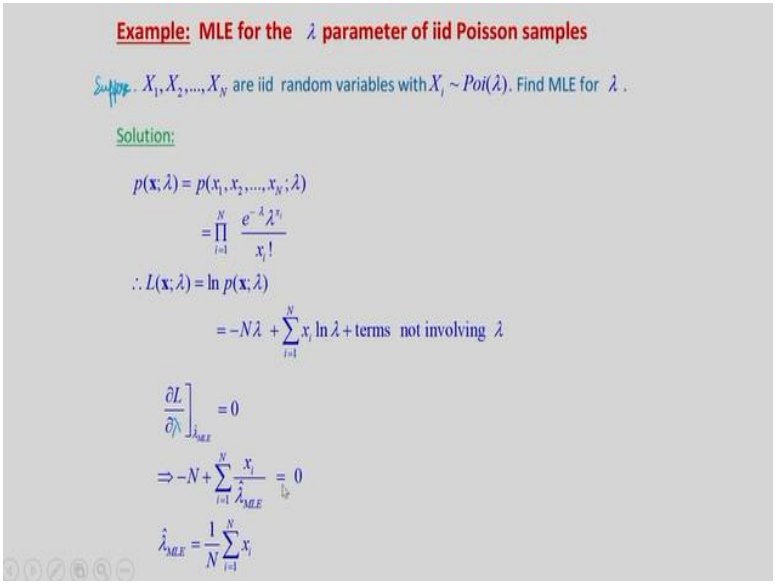
(Refer Slide Time: 16:38)



If we have k unknown parameters given by, theta vector that is equal to theta1, theta 2 up to theta k transpose, this is the parameter vector now. Then the MLE is given by a set of equations and that is del f del theta 1 at theta 1 hat MLE that is equal to 0, similarly del f del theta 2 f theta 2 hat MLE is equal to 0 like that they left del theta k at theta k hat MLE is equal to 0. So these are the conditions for theta hat MLE.

Now in terms of the log-likelihood function also we get a set of equations, so instead of f, now we can use L, del L del theta1 at theta1 hat MLE is equal to 0, del L del theta2 at theta 2 hat MLE is equal to 0 like that up to del L del theta k at theta k hat MLE is equal to 0. So these two conditions if we solve then we will find out the theta hat MLE.

(Refer Slide Time: 17:58)



We will consider some examples first example is MLE for the lambda parameter of the iid Poisson samples. Suppose, X1, X2 up to XN are iid random variables with each Xi is distributed as Poisson with lambda parameter. Find MLE for lambda. So let us solve this likelihood function is P x1, x2 up to xN as a function of lambda, now each one is distributed as Poisson.

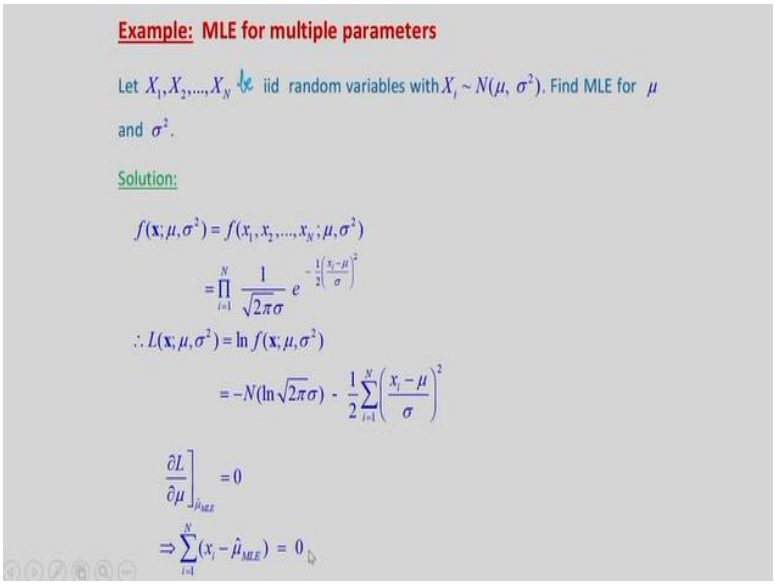
Therefore each distribution is given by e to the power - lambda, lambda to the power xi divided by factorial xi; this is the distribution for xi. So that way we have N distribution because of independent and identically distributed we will get this as product of i going from 1 to N, e to the power of - lambda, lambda to the power xi divided by factorial xi. So this is the likelihood of x as a function of lambda.

Now if we take the logarithm because that exponential terms are there taking logarithm will be easier, so L of x, lambda will be log of p of x, lambda and because entrants are there that we will get as - N lambda + summation Xi, i going from 1 to N log of lambda + terms not involving lambda, because we want to do the partial derivative with respect to lambda.

So we need not consider the terms which do not involve lambda. So if we take the partial derivative with respect to lambda at lambda hat MLE, we will get 0. So we will take the partial derivative of this will be equal to - N and similarly this part will give you 1 by lambda hat MLE. So that way we will get - N + summation i going from 1 to N, xi divided by partial derivative of this is 1 by lambda, so that way it will be lambda hat MLE that will be equal to 0.

So if we solve this we will get that lambda hat MLE is equal to 1 by N summation xi, i going from 1 to N. So that way we can find out the MLE maximum likelihood estimator for the lambda parameter. So this is the sample mean only.

(Refer Slide Time: 20:40)

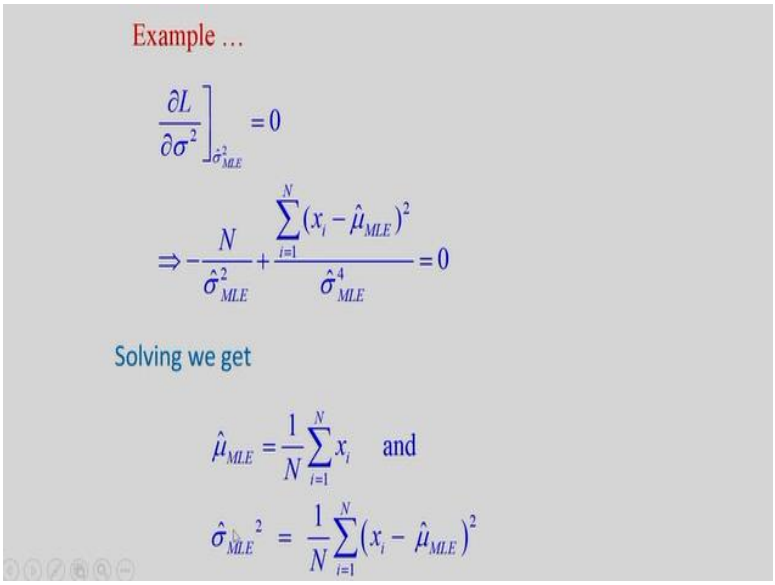


Second example will consider MLE for multiple parameters; let X1, X2 up to XN are iid random variables with this Xi distributed as a normal distribution with mean Mu and variance Sigma square. Find MLE for Mu and Sigma squared. So we have this likelihood function that is f of x1, x2 up to xN as a function of Mu, Sigma square. Now we have iid independent and identically distributed, so we will have a product of and N PDFs.

So product i going from 1 to N, 1 by root over 2 Pi Sigma into e to the power - half of xi - Mu divided by Sigma whole square. So this is the joint PDF of x1, x2 up to xN as a function of Mu and Sigma square, again here exponentially they are so taking logarithm will be beneficial so log likelihood function will be given by - N log of root over 2 Pi Sigma - 1/2 of summation i going from 1 to N of xi - Mu divided by Sigma whole square.

Now this is the expression and this we will take the partial derivative with respect to Mu and Sigma square. So if I take the partial derivative with respect to Mu at Mu hat MLE, we will get because these 2 and 2 will get cancelled summation xi - Mu hat MLE, i going from 1 to N, that must be equal to 0, this is one equation.

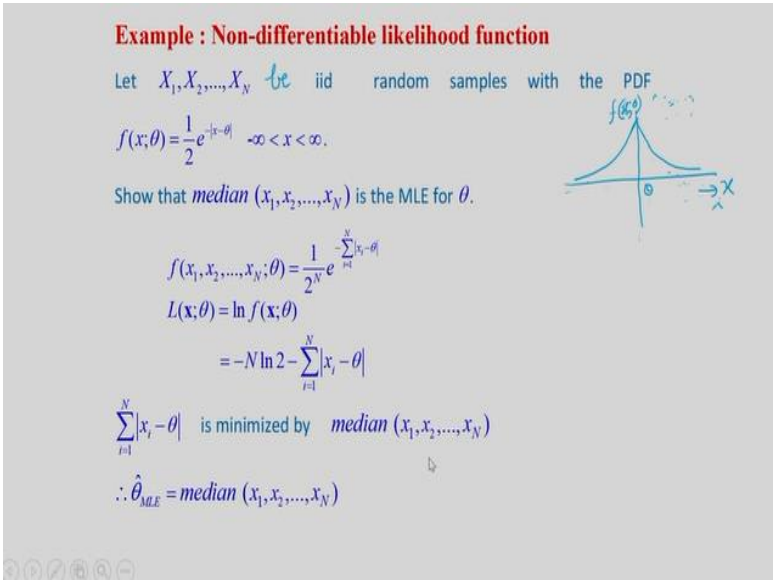
(Refer Slide Time: 22:32)



Similarly del L del Sigma square at Sigma hat square MLE that is also equal to 0, so from depth we will get - N by Sigma hat MLE square + summation x i - Mu hat MLE whole square, i going from 1 to N divided by Sigma hat to the power 4 MLE is equal to 0. So this is the equation given, so if I take the derivative with respect to Sigma square we will get this expression.

If we solve these two equations part equation is this and second equation is this, we will get Mu hat MLE is equal to 1 by N summation xi, i going from 1 to N. This is the sample mean and similarly this is the sample variance and that is Sigma hat MLE squared is equal to 1 by N summation xi - Mu hat MLE whole square, i going from 1 to N and so that way we get the estimators for mean and variance.

(Refer Slide Time: 23:49)

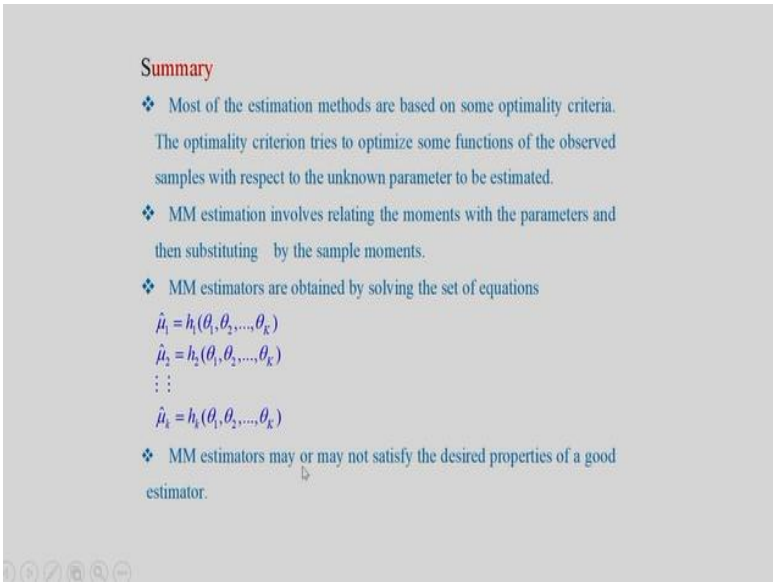


In the third example, we will consider a non differentiable likelihood function. Let X1,X2 up to XN be iid random variables with the PDF given by this f x, theta is equal to 1/2 e to the power - mode of x - theta, so this is symmetric about theta and this distribution is known as the Laplace distribution. So this is the PDF, so it will be distributed like this compared to Gaussian it will have a longer tail and its mean value is that theta.

Now for this distribution will show that median of x1, x2 up to xN is the MLE for theta, this PDF is given by this because it is a iid independent and identically distributed random variable. So if we have to take the joint PDF that is the product, so this product is given by this. So log likelihood function is given by - N log up to - summation mode of xi - theta, i doing from 1 to N.

Because it is a negative sign is here, we have to minimize and this minimization is done by the median of x1, x2 up to xN, this absolute sum will be minimized if we take theta to be the median. Therefore theta hat MLE is the median of x1, x2 up to xN. We see that for the Laplace distribution, the MLE for theta is not the mean but the median of x1, x2 up to xn, so that way we considered the how to find out MLE for a non differentiable likelihood function.

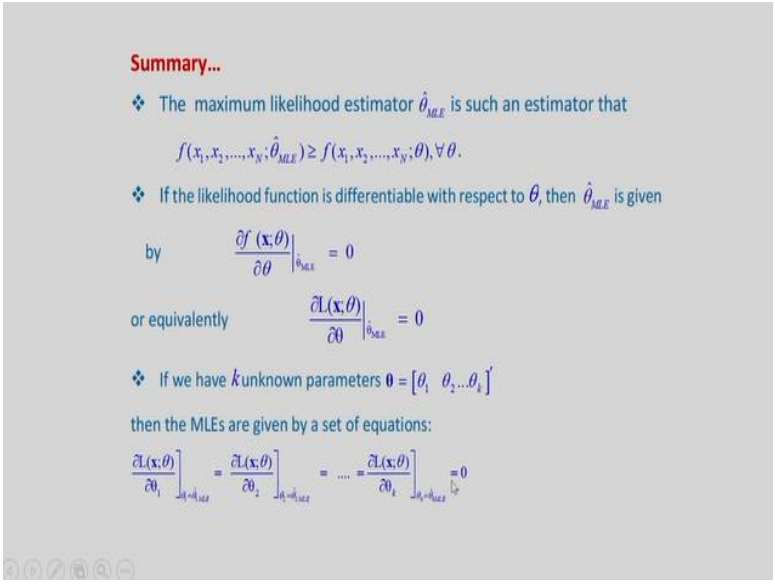
(Refer Slide Time: 25:47)



Let us summarize the lecture today, most of the estimation methods are based on some optimality criteria. The optimality criterion tries to optimize some functions of the observed samples with respect to the unknown parameter to be estimated. Method of moment estimation involves relating the moments with the parameters and then substituting the moments by the corresponding sample moments.

Therefore MM estimators are obtained by solving the following set of equation that is Mu1 hat is equal to suppose h1 theta1, theta 2 up to theta k, Mu2 hat is equal to h2 theta1, theta 2 up to theta k like that, Mu k hat is equal to hk theta1, theta 2 up to theta k. So this set of equation if we solve we will get the MM estimators. MM estimators may or may not satisfy the desired properties of a good estimator to numerical simulation, we can study the properties of MM estimators.

(Refer Slide Time: 27:04)



Then we discussed maximum likelihood estimator and theta hat MLE is such an estimator that this likelihood function at theta hat MLE is greater than equal to the likelihood function at any other theta that is true all theta, this is the likelihood principle. If the likelihood function is differentiable with respect to theta, then theta hat MLE is given by this relationship partial derivative of f with respect to theta f theta at MLE is equal to 0.

Equivalently the terms of log likelihood function del L del theta at theta hat MLE is equal to 0. So this is the equation we solve to find out the value of theta hat MLE. If we have k unknown parameters theta is equal to theta1, theta2 up to theta k transpose, this is the parameter vector. Then the MLEs are given by a set of equations that is del L del theta1, theta 1 equal to theta 1 hat MLE, that must be equal to 0.

del L del theta 2, theta 2 is equal to theta 2 hat MLE, that must be equal to 0, like that; del L del theta k, theta k is equal to theta k hat a valid that must be equal to 0. In the next lecture we will see the properties of ML estimators, thank you.