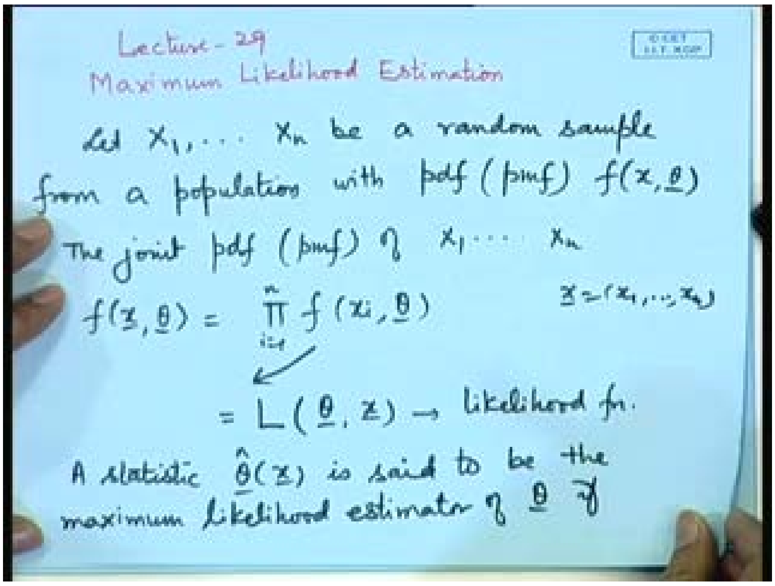
Probability and Statistics
Prof. Dr. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture No. #29
Estimation-III

In the last lecture, I introduced method of maximum likelihood estimation. Besides that, we had also discussed the method of least squares estimation and the method of moments.

(Refer Slide Time: 00:35)



Lecture - 29
Maximum Likelihood Estimation

Let $X_1, \ldots X_n$ be a random sample from a population with pdf (pmf) $f(x, \underline{\theta})$

The joint pdf (pmf) of $X_1 \cdots X_n$

$f(\underline{x}, \underline{\theta}) = \prod_{i=1}^{n} f(x_i, \underline{\theta})$ $\quad \underline{x} = (x_1, \ldots, x_n)$

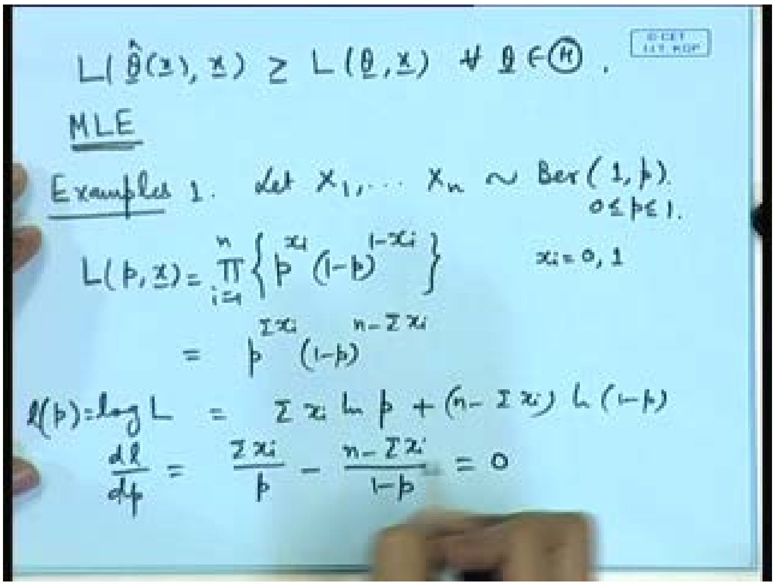
$= L(\underline{\theta}, \underline{x}) \rightarrow$ likelihood fn.

A statistic $\hat{\underline{\theta}}(\underline{x})$ is said to be the maximum likelihood estimator of $\underline{\theta}$ if

Now, I continue the discussion on the method of maximum likelihood estimation which is actually the last among the 3 mentioned methods that I have given. So, let me repeat the definition of the maximum likelihood definition that how it is obtained.

Firstly, we define likelihood function. So, if I have x 1, x 2, x n let x 1, x 2, x n be a random sample from a population with either p d f or p m f say f x theta. So, what we write the joint p d f or p m f say of x 1, x 2, x n. So, we write it as say f x theta equal to product of f x i theta i is equal to 1 to n we call this now see when x 1, x 2, x n is observed to be some value small x 1, small x 2, small x n, then we call this as the likelihood function of theta and we may put x here just to denote its dependence on x also this is called the likelihood function.

(Refer Slide Time: 02:53)



$L(\hat{\theta}(x), x) \geq L(\theta, x) \quad \forall \theta \in \Theta$.

MLE

Examples 1. Let $X_1, \dots X_n \sim Ber(1, p)$. $0 \leq p \leq 1$.

$L(p, x) = \prod_{i=1}^{n} \{ p^{x_i} (1-p)^{1-x_i} \}$ $\quad x_i = 0, 1$

$= p^{\Sigma x_i} (1-p)^{n - \Sigma x_i}$

$l(p) = \log L = \Sigma x_i \ln p + (n - \Sigma x_i) \ln (1-p)$

$\frac{dl}{dp} = \frac{\Sigma x_i}{p} - \frac{n - \Sigma x_i}{1-p} = 0$

What we consider a statistic theta head x is said to be the maximum likelihood estimator of theta if L theta head x is greater than or equal to L theta x for all theta belonging to the parameter space. So, in short we use the terminology MLE for maximum likelihood estimator a popular technique to obtain the optimizing value of theta is to take logarithm of l and then consider the derivative of L with respect to the parametric functions and then equate to 0 and solve those equations.

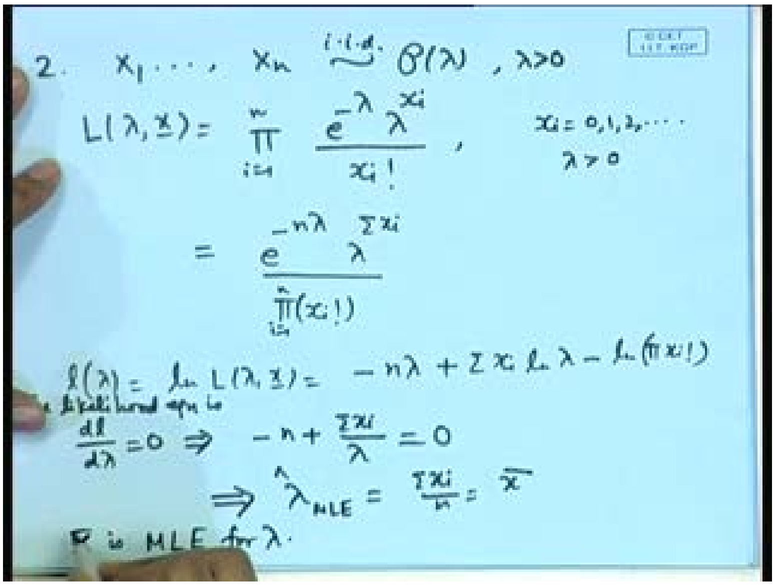
Let me explain the method through some examples let x 1, x 2, x n, follow say, Bernoulli distribution. So, here p is the unknown parameter we want to find out the maximum likelihood estimator of for p. We write down the likelihood function here that will be equal to p to the power x i 1 minus p to the power 1 minus x i product i is equal to 1 to n if you look at the inside quantity this is the probability mass function of x i. So, here each of the x i's can take value 0 or one. So, this can be simplified as p to the power sigma x i 1 minus p to the power n minus sigma x i.

So, we have to differentiate this with respect to p and solve it. So, what we can do we can consider log of L which I call as small L of p that is equal to sigma x i log p plus n minus sigma x i log of 1 minus p. So, d l by d p is equal to sigma x i by p minus n minus sigma x i by 1 minus p that is equal to 0.

Now, then this gives after some simplification we will get p head is equal to. So, in fact you can solve it let us combine the terms it will give me 1 minus p sigma x i minus n p plus p sigma x i divided by. So, this we can put p is belonging to the interval 0 to 1 we can take the case p is equal to 0 and p is equal to 1 separately. So, this becomes sigma x i minus n p is equal to 0; that means, p head is equal to sigma x i by n which i can call say x by n where x is the sigma of x i.

If you compare it with the method of moments estimator actually it is the same thing here; that means, the maximum likelihood estimator of the probability of success is actually the number of successes in n trials divided by the number of trial that is the proportion of the number of success which is a very logical estimator and it is coming through the method of maximum likelihood estimation.

(Refer Slide Time: 06:40)



2. $X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathcal{P}(\lambda)$, $\lambda > 0$

$$L(\lambda, \underline{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, \dots \quad \lambda > 0$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod_{i=1}^{n} (x_i!)}$$

$$l(\lambda) = \ln L(\lambda, \underline{x}) = -n\lambda + \sum x_i \ln \lambda - \ln(\prod x_i!)$$

likelihood eqn is

$$\frac{dl}{d\lambda} = 0 \Rightarrow -n + \frac{\sum x_i}{\lambda} = 0$$

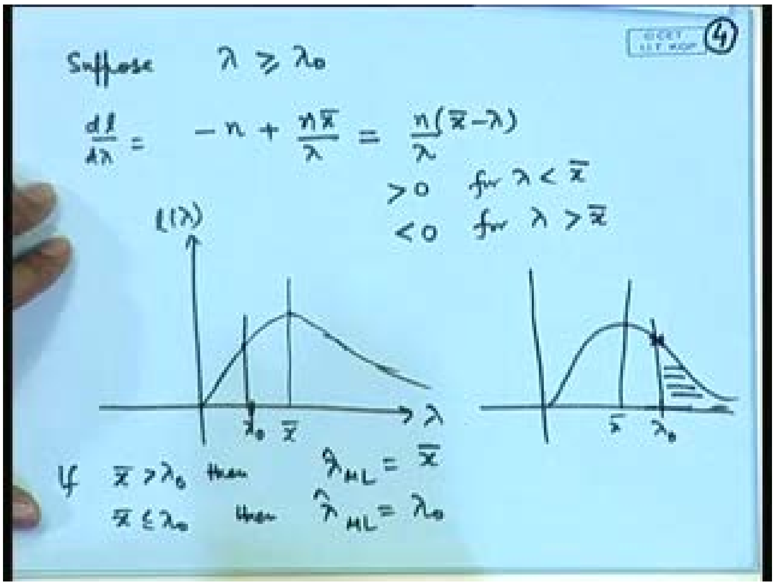
$$\Rightarrow \hat{\lambda}_{MLE} = \frac{\sum x_i}{n} = \bar{x}$$

$\bar{x}$ is MLE for λ.

Let us take say x 1, x 2, x n to be a random sample from poison distribution with parameter lambda let us write down the likelihood function that is equal to e to the power minus lambda to the power x i divided by x i factorial product I is equal to 1 to n here each of the x i’s can take values 0, 1, 2 and. So, on and lambda is positive. So, this is equal to e to the power minus n lambda to the power sigma x i divided by product x i factorial I is equal to 1 to n.

So, we can write the log likelihood function as minus n lambda plus sigma x i log of lambda minus log of product x i factorial. The likelihood equation d l by d lambda is equal to 0 that is minus n plus sigma x i by lambda is equal to 0. The solution of this gives lambda head m l e is equal to sigma x i by n that is x bar. So, x bar is the maximum likelihood estimator for lambda note here that in these 2 cases the maximum likelihood estimator and the method of moments estimators are same. However, as we will see it is not a rule in many cases, they will not be the same.

(Refer Slide Time: 08:48)



Suppose $\lambda \geq \lambda_0$

$$\frac{dl}{d\lambda} = -n + \frac{n\bar{x}}{\lambda} = \frac{n(\bar{x}-\lambda)}{\lambda}$$

> 0 for $\lambda < \bar{x}$

< 0 for $\lambda > \bar{x}$

If $\bar{x} > \lambda_0$ then $\hat{\lambda}_{ML} = \bar{x}$

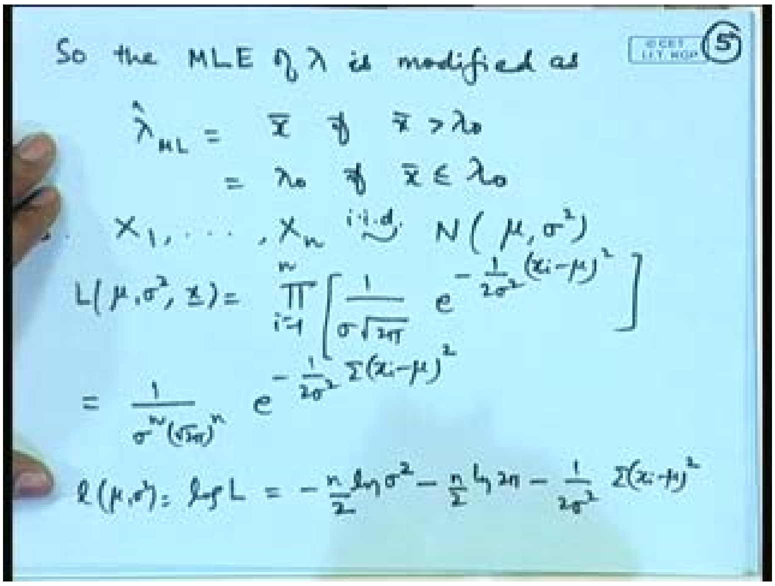
$\bar{x} \leq \lambda_0$ then $\hat{\lambda}_{ML} = \lambda_0$

Let me take an example of that kind. Another thing that we can observe here the maximization is done over the parameter space. So, in case there is a modification for example, in the previous exercise suppose, we know that lambda is say greater than or equal to lambda not; that means, we know that the rate of arrival in a Poisson process is bigger than a prescribed quantity. In that case, if we look at x bar this is actually the maximization over the full parameter space and we may have a situation where x bar is actually less than lambda naught. In that case, this does not satisfy the property that likelihood function is maximized over the parameter space.

What we do then? Consider the behavior of the likelihood function what we are getting is d l by d lambda is equal to. So, we write it in minus n plus n x bar by lambda that we can write as x bar minus lambda n by lambda that is n x bar by lambda minus lambda minus n. So, you can see here that if lambda is less than x bar then this is positive it is less than 0 for lambda greater than x bar.

We look at the behavior of the function it is increasing up to x bar and then decreasing after x bar. So, let us plot it as a function of lambda the likelihood function; that means, on this side I have log likelihood it is increasing and then it is decreasing this is the point x bar now suppose lambda not is here if lambda not is here then you can see that the maximum value x bar is satisfying this condition and therefore, x bar remains the m l e that is if x bar is greater than lambda not then lambda head m l is x bar; however, we may have a situation where this is x bar and lambda not is say here in that case you see the parameter space is this; that means, the maximum value that is occurring is at actually lambda not; that means, if x bar is less than or equal to lambda not then lambda head m l is equal to lambda not. So, our estimator is then modified.

(Refer Slide Time: 11:41)



So the MLE of λ is modified as

$$\hat{\lambda}_{ML} = \bar{x} \text{ if } \bar{x} > \lambda_0$$
$$= \lambda_0 \text{ if } \bar{x} \le \lambda_0$$

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$$

$$L(\mu, \sigma^2, \underline{x}) = \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right]$$

$$= \frac{1}{\sigma^n (\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2}$$

$$l(\mu, \sigma^2) = \log L = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2$$

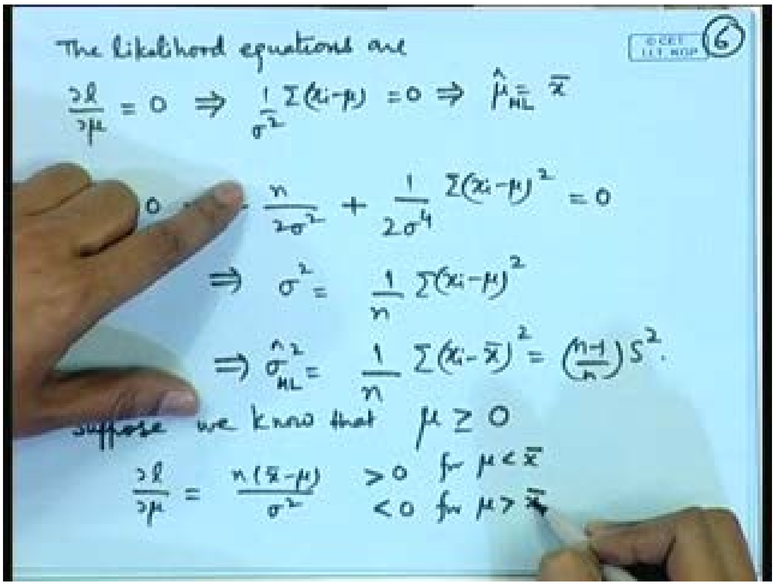
The maximum likelihood estimator of lambda is modified as let me write it as, lambda head m l this is equal to x bar if x bar is greater than lambda not it is equal to lambda not if x bar is less than or equal to lambda not.

Now, this brings into focus another important property or you can say another important aspect of the maximum likelihood estimator which would have been missed by the method of moments because, in the method of moments the restriction on the parameter space does not play any role. There we simply look at the moment which is not affected by the restriction on the parameter space and. So, the method of moments estimator remains as x bar. Whereas, here you see the effect of the restriction on the parameter space is getting affected on the maximum likelihood estimator which is actually a reasonable thing because if x bar is less than or equal to lambda not which is going outside the parameter space we should not take x bar as an estimate for lambda.

Let us take another popular example x 1, x 2, x n follows normal mu sigma square. If we consider the likelihood function here then it is a function of 2 parameters mu and sigma square. So, l mu sigma square x bar x that is equal to product of 1 by sigma root 2 pi e to the power minus 1 by 2 sigma square x i minus mu square now this term we will simplify this can be written as 1 by sigma to the power n root 2 pi to the power n e to the power minus 1 by 2 sigma square sigma x i minus mu square. So, the log likelihood function minus n log sigma minus n log n by 2 log 2 pi minus 1 by 2 sigma square sigma x i minus mu square.

Now, note here that I have written it as minus n log sigma now 1 may ask that if parameter is sigma square then whether we should write it as sigma the answer is that both are because maximum likelihood estimation is invariant under the transformation of the parameters suppose I obtain the MLE of sigma square in place of sigma and then I want for sigma then I should simply take the square root of that. So, I may write it like this also it does not make any difference.

(Refer Slide Time: 14:49)



© CET I.I.T. KGP 6

The likelihood equations are

$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2}\Sigma(x_i-\mu) = 0 \Rightarrow \hat{\mu}_{ML} = \bar{x}$$

$$0 \quad \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(x_i-\mu)^2 = 0$$

$$\Rightarrow \sigma^2 = \frac{1}{n}\Sigma(x_i-\mu)^2$$

$$\Rightarrow \hat{\sigma}^2_{ML} = \frac{1}{n}\Sigma(x_i-\bar{x})^2 = \left(\frac{n-1}{n}\right)s^2.$$

suppose we know that $\mu \geq 0$

$$\frac{\partial l}{\partial \mu} = \frac{n(\bar{x}-\mu)}{\sigma^2} \quad \begin{matrix} > 0 \text{ for } \mu < \bar{x} \\ < 0 \text{ for } \mu > \bar{x} \end{matrix}$$

The likelihood equations in this case will be del l by del mu is equal to 0; that means, sigma x i minus mu 1 by sigma square is equal to 0 which will give mu head is equal to x bar and del l by del sigma square is equal to 0 that gives minus n by 2 sigma square plus 1 by 2 sigma to the power four sigma x i minus mu square this gives sigma square is equal to 1 by n sigma x i minus mu square; that means, the equation for sigma square involves mu.

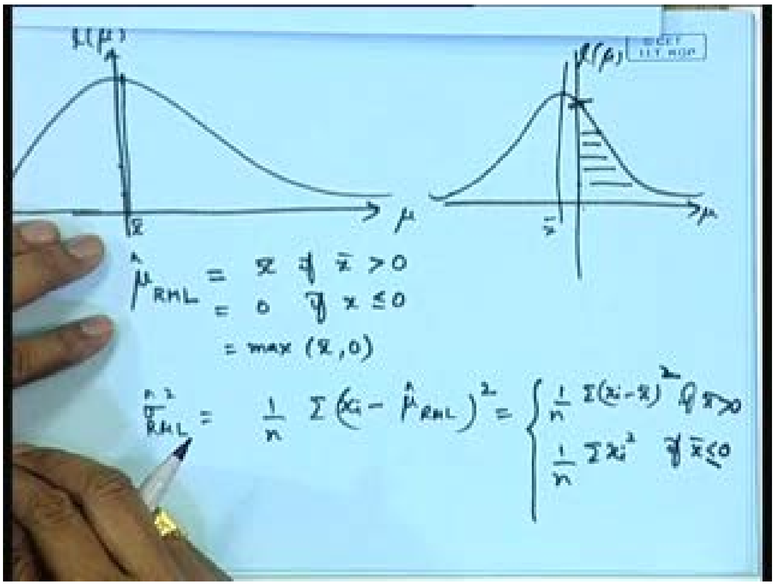
Since we have already got the solution for mu we can substitute. So, sigma head square that is the maximum likelihood estimator becomes that is n minus 1 by n s square.

You can note here that it is similar to the method of moment's estimator in this case. In many situations, the method of moments estimators and the maximum likelihood estimates coincide, but there are other cases where they may not coincide we have already seen the example that if the parameter space gets modified then the maximum likelihood estimator gets modified. For example, in this particular situation, suppose I consider, suppose we know that mu is greater than or equal to 0. Now, this type of situation occurs when through some experience we already know that the mean is actually non negative although, the variable may be normally distributed. But, because of the way the experiment has been framed or any other reason the parameter space is restricted; that means, we know that the mean is greater than or equal to 0.

Now, you see here we have the maximum likelihood estimator for mu as x bar and for sigma square it is n minus 1 by n s square now if we see that x bar by observation gives us a negative value then it will become an unreasonable estimator for mu which is taken to be greater than or equal to 0. So, we can analyze it in a proper way by looking at the behavior of the log likelihood function which we have actually maximized.

So, if we look at with respect to mu we have del l by del mu as equal to. So, del l by del mu that function we got it as n times x bar minus mu by sigma square. So, you can see that it is greater than 0 for mu less than x bar and it is less than 0 for mu greater than x bar; that means, the function is increasing up to x bar and decreasing beyond x bar; that means, the shape of the likelihood function as a function of mu is something like. So, this could be the maximizing choice.

(Refer Slide Time: 18:08)



$$\hat{\mu}_{RML} = \bar{x} \text{ if } \bar{x} > 0$$
$$= 0 \text{ if } x \le 0$$
$$= \max(\bar{x}, 0)$$

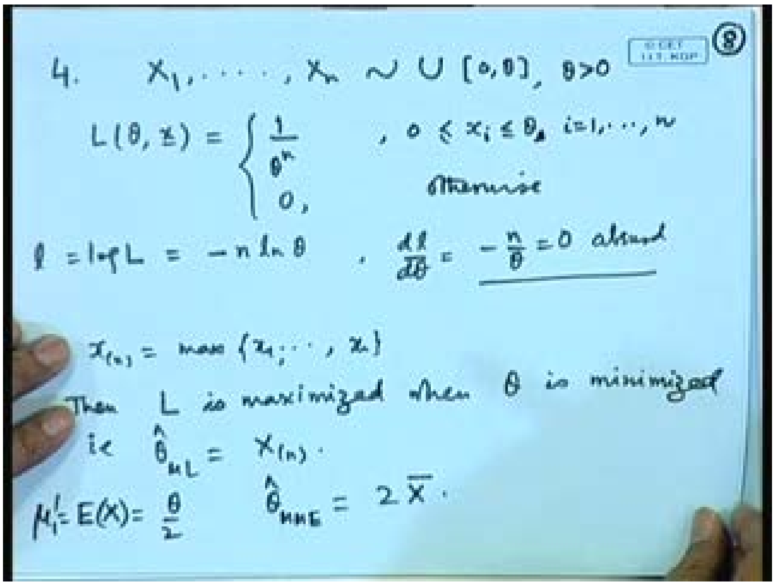
$$\hat{\sigma}^2_{RML} = \frac{1}{n} \Sigma (x_i - \hat{\mu}_{RML})^2 = \begin{cases} \frac{1}{n} \Sigma (x_i - \bar{x})^2 & \text{if } \bar{x} > 0 \\ \frac{1}{n} \Sigma x_i^2 & \text{if } \bar{x} \le 0 \end{cases}$$

Now, if x bar is bigger than 0 then we may take x bar, but we may have a situation where x bar may be less than 0. So, if the parameter space is this then the maximum is occurring at 0 itself. So, the modified or you can say the restricted maximum likelihood estimator becomes x bar if x bar is greater than 0 it is equal to 0 if x bar is less than or equal to 0 which we can actually call as maximum of x bar and 0.

Now, is there any effect on the estimator for sigma square the answer is yes because the maximum likelihood estimator for sigma square was obtained by substituting the estimator for mu in this second likelihood equation. So, we will get sigma head square m l as 1 by n sigma x i minus mu head RML square. So, this we can write as when x bar is positive then this is the odd 1 that is 1 by n sigma x i minus x bar whole square if x bar is positive and it is equal to 1 by n sigma x i square if x bar is negative or less than or equal to 0.

We will show later on that; when the parameter space is restricted this restricted maximum likelihood estimator has a better performance as compared to the usual maximum likelihood estimator that we obtained before. We will define the criteria of better a little later.

(Refer Slide Time: 20:36)



4. $X_1, \cdots, X_n \sim U[0, \theta], \ \theta > 0$

$$L(\theta, \underline{x}) = \begin{cases} \frac{1}{\theta^n} & , 0 \leq x_i \leq \theta, \ i = 1, \cdots, n \\ 0, & \text{otherwise} \end{cases}$$

$l = \log L = -n \ln \theta$, $\frac{dl}{d\theta} = \underline{-\frac{n}{\theta} = 0 \text{ absurd}}$

$x_{(n)} = \max\{x_1, \cdots, x_n\}$

Then L is maximized when θ is minimized

ie $\hat{\theta}_{ML} = X_{(n)}$.

$\mu_1' = E(X) = \frac{\theta}{2}$ $\quad \hat{\theta}_{MME} = 2\bar{X}$.

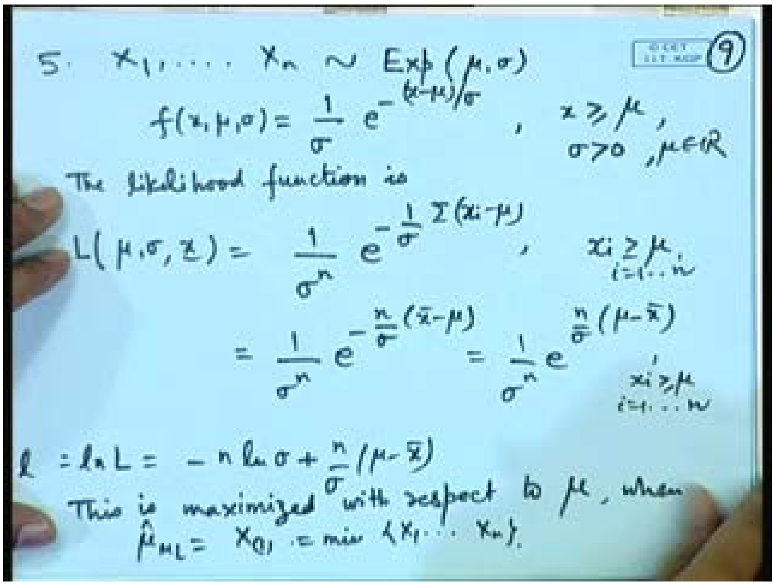
Let us take another example where this process of argument does not seem to work let us consider say a random sample from uniform distribution on the interval say 0 to theta where theta is a positive real number. So, here the likelihood function is simply 1 by theta to the power n for 0 less than or equal to x i less than or equal to theta for I is equal to 1 to n; it is 0 otherwise.

Now, you see if we follow the previous procedure of taking logarithm and differentiating what I will get if I take log of l then that will give me minus n l n theta and if I differentiate let me call it small l. So, d l by d theta that will give me minus n by theta. So, if I put this equal to 0 this is actually absurd. So, why this is happening this is happening because we are not taking care of the reason when we are differentiating and putting equal to 0 basically we are trying to find out the minimum and maximum over a range of parameter here the range of the parameter value is dependent upon x i's. So, that is not being taken care by this kind of process.

So, let us look at a direct argument our aim is to maximize the likelihood function 1 by theta to the power n with respect to theta since theta is in the denominator it corresponds to the minimization with respect to theta what is the minimum value of theta the minimum value of theta will be actually since e theta is greater than each of the x i's the minimum value that theta can take will be the maximum of x 1, x 2, x n. So, if I call x n as the maximum of x 1, x 2, x n that is the largest order statistic; then l is maximized when theta is minimized that is theta head m l is equal to x n. So, this is the maximum likelihood estimator for theta.

Suppose I want to find out the method of moments estimator here what is that consider the mean of this distribution that is the first moment that is theta by 2. So, theta head method of moments estimator that will be twice and in place of mu 1 prime we will put x bar. So, you can see the situation here the method of moments estimator and the maximum likelihood estimator are totally different in fact they are here it is the maximum of the observations and here it is average 2 times the average of the observations further this example illustrates that the method of taking logarithm and differentiating does not always work.

(Refer Slide Time: 24:03)



5. $x_1, \ldots x_n \sim \text{Exp}(\mu, \sigma)$

$$f(x, \mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \quad x \geq \mu, \; \sigma > 0, \; \mu \in \mathbb{R}$$

The likelihood function is

$$L(\mu, \sigma, \underline{x}) = \frac{1}{\sigma^n} e^{-\frac{1}{\sigma}\Sigma(x_i - \mu)}, \quad x_i \geq \mu, \; i = 1 \ldots n$$

$$= \frac{1}{\sigma^n} e^{-\frac{n}{\sigma}(\bar{x} - \mu)} = \frac{1}{\sigma^n} e^{\frac{n}{\sigma}(\mu - \bar{x})}, \quad x_i \geq \mu, \; i = 1 \ldots n$$

$$l = \ln L = -n \ln \sigma + \frac{n}{\sigma}(\mu - \bar{x})$$

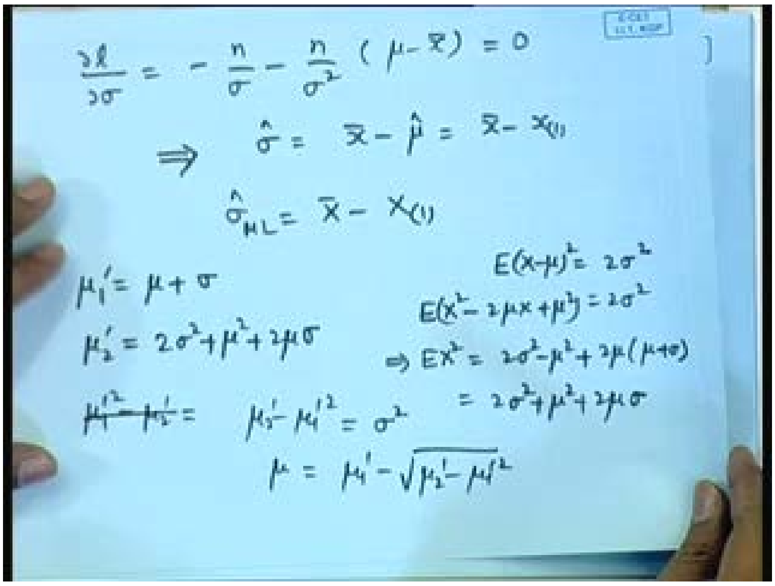
This is maximized with respect to μ, when $\hat{\mu}_{ML} = X_{(1)} = \min\{X_1 \ldots X_n\}$.

Let us take another example say x 1, x 2, x n follow exponential distribution let me consider say 2 parameter exponential distribution mu sigma; that means, the density function is 1 by sigma e to the power x minus mu by sigma that is the density of here x is greater than or equal to mu sigma is positive and mu is any real number the likelihood function will be 1 by sigma to the power n e to the power minus 1 by sigma x i minus mu each x i is greater than or equal to mu which we can write as 1 by sigma to the power n e to the power minus n by sigma x bar minus mu or 1 by sigma to the power n e to the power n by sigma mu minus x bar.

Now, if we want to maximize this function with respect to mu then likelihood equation and then derivative will not give a result like in the uniform case because the log likelihood function minus n l n sigma plus n by sigma mu minus x bar you can see here if I differentiate with respect to mu and put equal to 0; I get an absurd result.

However, with respect to sigma we can do that, but for finding out the maximum likelihood estimator with respect to mu we can use the direct argument this is a increasing function of mu. So, the maximum value will be attained when mu attains its maximum value since mu is always less than or equal to x i’s the maximization will occur when mu is the minimum of the x i’s. So, this is maximized with respect to mu when mu head m l is equal to x 1 that is the minimum of the observations.

(Refer Slide Time: 27:02)



$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{n}{\sigma^2}(\mu - \bar{x}) = 0$$

$$\Rightarrow \hat{\sigma} = \bar{x} - \hat{\mu} = \bar{x} - x_{(1)}$$

$$\hat{\sigma}_{ML} = \bar{x} - X_{(1)}$$

$$\mu_1' = \mu + \sigma \qquad E(x-\mu)^2 = 2\sigma^2$$

$$\mu_2' = 2\sigma^2 + \mu^2 + 2\mu\sigma \qquad E(x^2 - 2\mu x + \mu^2) = 2\sigma^2$$

$$\Rightarrow Ex^2 = 2\sigma^2 - \mu^2 + 2\mu(\mu+\sigma)$$

$$= 2\sigma^2 + \mu^2 + 2\mu\sigma$$

$$\cancel{\mu_1'^2 - \mu_2'} = \mu_2' - \mu_1'^2 = \sigma^2$$

$$\mu = \mu_1' - \sqrt{\mu_2' - \mu_1'^2}$$

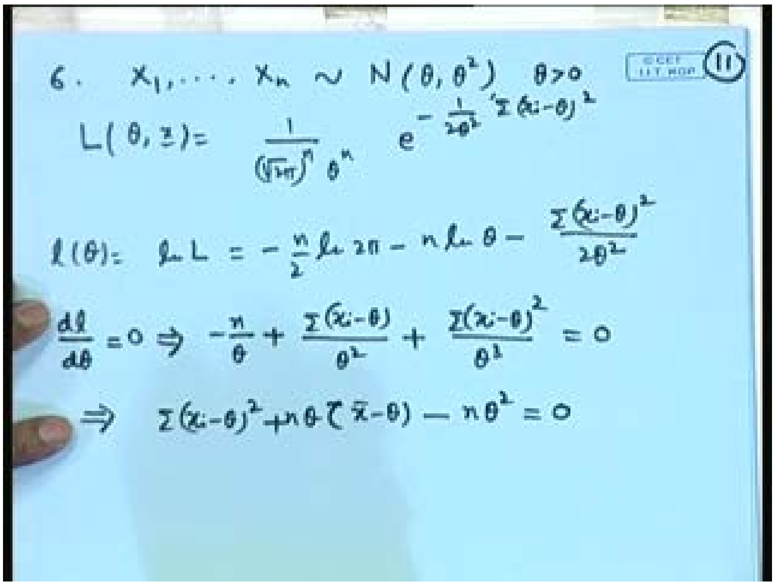
We can substitute this here and get the estimator for sigma also consider the derivative of this with respect to sigma that will give minus n by sigma minus n by sigma square mu minus x bar this is equal to 0. So, this gives sigma head is equal to x bar minus mu head that is equal to x bar minus x 1. So, sigma head m l is equal to x bar minus x 1 let us see what is the method of moments estimator in this case.

To obtain the method of moments estimator we have to consider the moments of this distribution now here it is a 2 parameter distribution I will have to find out first 2 moments. So, mu 1 prime is equal to mu plus sigma mu 2 prime now for that we can consider certain transformation expectation of x minus mu square that is equal to twice sigma square. So, expectation of x square minus 2 mu x plus mu square is equal to twice sigma square. Expectation of x square is equal to 2 sigma square minus mu square plus twice mu expectation of x; that is, mu plus sigma that is equal to twice sigma square plus mu square plus 2 mu sigma. So, the second moment is twice sigma square plus mu square plus 2 mu sigma.

Now, if we consider mu 1 prime square minus mu 2 prime rather I consider mu 2 prime minus mu 1 prime square then that gives me sigma square and therefore, mu becomes mu 1 prime minus square root of mu 2 prime minus mu 1 prime square. So, the method of moments estimators for mu and sigma square will be obtained as sigma head square m m e that is equal to 1 by n sigma x i minus x bar whole square and mu head m m e will be equal to x bar minus square root of 1 by n sigma x i minus x bar whole square. Note here that the maximum likelihood estimators and the method of moment estimators are quite different from each other and again then therefore, the question arises that which of this is better.

Now, here we will discuss also the situations where the form of the maximum likelihood estimator may not be determined explicitly it may not exist or in case of certain situations.

(Refer Slide Time: 30:45)



6. $X_1, \ldots, X_n \sim N(\theta, \theta^2) \quad \theta > 0$

$$L(\theta, \underset{\sim}{x}) = \frac{1}{(\sqrt{2\pi})^n\,\theta^n}\, e^{-\frac{1}{2\theta^2}\Sigma(x_i-\theta)^2}$$

$$l(\theta) = \ln L = -\frac{n}{2}\ln 2\pi - n\ln\theta - \frac{\Sigma(x_i-\theta)^2}{2\theta^2}$$

$$\frac{dl}{d\theta} = 0 \Rightarrow -\frac{n}{\theta} + \frac{\Sigma(x_i-\theta)}{\theta^2} + \frac{\Sigma(x_i-\theta)^2}{\theta^3} = 0$$

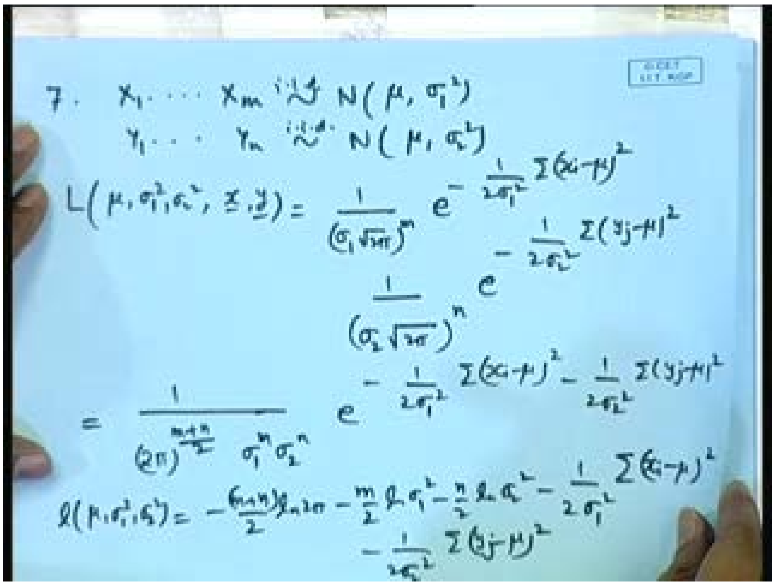
$$\Rightarrow \Sigma(x_i-\theta)^2 + n\theta(\bar{x}-\theta) - n\theta^2 = 0$$

We may have non unique maximum likelihood estimator let us take say x 1, x 2, x n follows a normal theta theta square; that means, I am considering the situation where the mean and the standard deviation are the same. So, naturally theta has to be positive here now here the likelihood function if you write then following the earlier setup it becomes 1 by root 2 pi to the power n theta to the power n e to the power minus 1 by twice theta square sigma x i minus theta whole square.

So, the log of the likelihood function that is minus n by 2 log of 2 pi minus n log theta minus sigma x i minus theta whole square by twice theta square. So, what is the likelihood equation here d l by d theta equal to 0 that gives minus n by theta plus now if i consider here this term this is consisting of theta in the numerator as well as in the denominator. So, the derivative will come in 2 terms sigma x i minus theta by theta square and plus sigma x i minus theta square by theta cube is equal to 0.

So, if theta is taken to be positive i can strike of 1 theta and we can write the equation as. I multiply by theta cube in the full equation and we get it as sigma x i minus theta square plus theta into n x bar minus theta minus n theta square equal to 0 here you can see that the solution of this equation can be obtained in the terms of solution of a quadratic equation the solution can be obtained.

(Refer Slide Time: 33:51)



7. $X_1 \cdots X_m \overset{\text{iid}}{\sim} N(\mu, \sigma_1^2)$

$Y_1 \cdots Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma_2^2)$

$$L(\mu, \sigma_1^2, \sigma_2^2, \underline{x}, \underline{y}) = \frac{1}{(\sigma_1\sqrt{2\pi})^m} e^{-\frac{1}{2\sigma_1^2}\Sigma(x_i-\mu)^2} \frac{1}{(\sigma_2\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma_2^2}\Sigma(y_j-\mu)^2}$$

$$= \frac{1}{(2\pi)^{\frac{m+n}{2}} \sigma_1^m \sigma_2^n} e^{-\frac{1}{2\sigma_1^2}\Sigma(x_i-\mu)^2 - \frac{1}{2\sigma_2^2}\Sigma(y_j-\mu)^2}$$

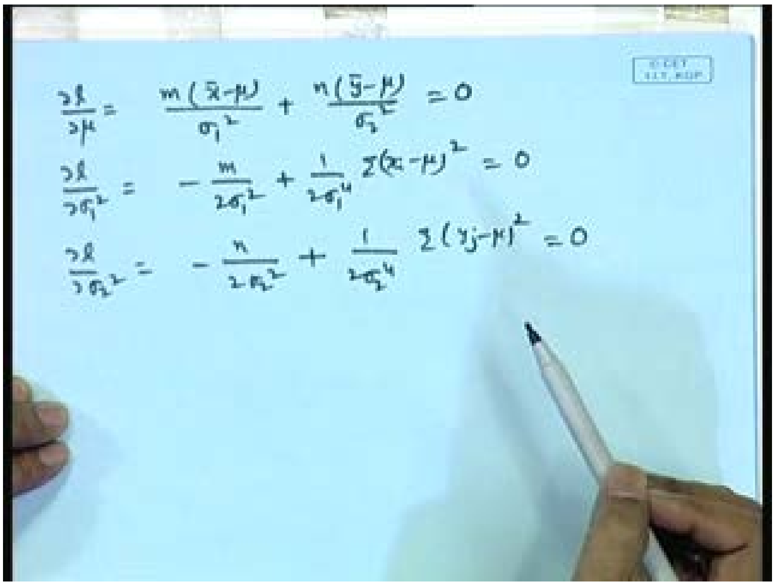
$$l(\mu, \sigma_1^2, \sigma_2^2) = -\frac{(m+n)}{2}\ln 2\pi - \frac{m}{2}\ln\sigma_1^2 - \frac{n}{2}\ln\sigma_2^2 - \frac{1}{2\sigma_1^2}\Sigma(x_i-\mu)^2 - \frac{1}{2\sigma_2^2}\Sigma(y_j-\mu)^2$$

Let me take another example of similar nature where, the form may be even more difficult now this is popularly called the problem of common mean in the statistical inference. So, we have a random sample from a normal population with mean mu and variance sigma 1 square and another random sample y 1 y 2 y n from a normal population with mean mu and variance sigma 2 square. So, in particular sigma 1 square sigma 2 square may be different, but the mean is common.

So, here you write down the likelihood function here 3 parameters are there mu sigma 1 square sigma 2 square and 2 samples x and y are there. So, the likelihood function will involve 1 by sigma 1 root 2 pi to the power m e to the power minus 1 by 2 sigma 1 square sigma x i minus mu square and 1 by sigma 2 root 2 pi to the power n e to the power minus 1 by 2 sigma 2 square sigma y j minus mu whole square. So, the terms can be simplified: 1 by 2 pi to the power m plus n by 2 sigma 1 to the power n sigma 2 to the power n e to the power minus 1 by 2 sigma 1 square sigma x i minus mu square minus 1 by 2 sigma 2 square sigma y j minus mu square.

So, the log likelihood function is equal to minus m plus n by 2 l n 2 pi minus n by 2 l n minus m by 2 l n sigma 1 square minus n by 2 l n sigma 2 square minus 1 by 2 sigma 1 square sigma x i minus mu square minus 1 by 2 sigma 2 square sigma y j minus mu square.

(Refer Slide Time: 35:53)



$$\frac{\partial l}{\partial \mu} = \frac{m(\bar{x}-\mu)}{\sigma_1^2} + \frac{n(\bar{y}-\mu)}{\sigma_2^2} = 0$$

$$\frac{\partial l}{\partial \sigma_1^2} = -\frac{m}{2\sigma_1^2} + \frac{1}{2\sigma_1^4}\Sigma(x_i-\mu)^2 = 0$$

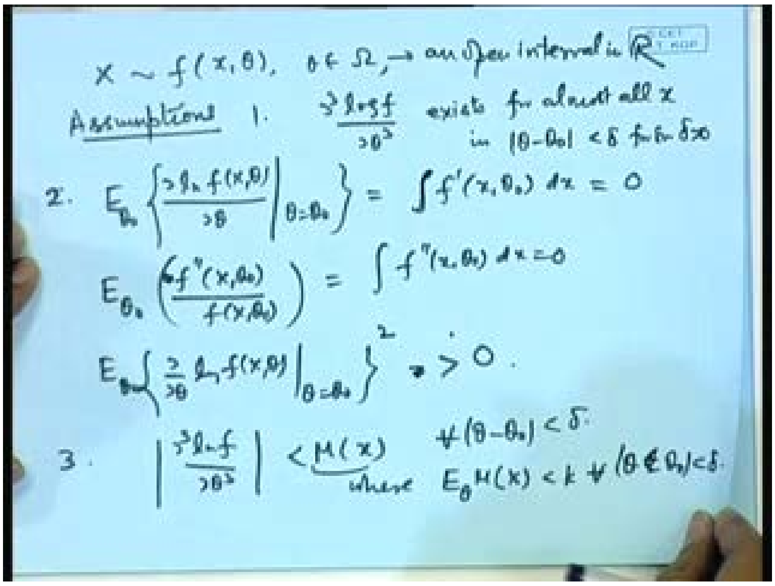
$$\frac{\partial l}{\partial \sigma_2^2} = -\frac{n}{2\sigma_2^2} + \frac{1}{2\sigma_2^4}\Sigma(y_j-\mu)^2 = 0$$

So, if we consider the likelihood equations the equations are del l by del mu is equal to 0 that gives us m x bar minus mu by sigma 1 square plus n y bar minus mu by sigma 2 square is equal to 0 if I consider del l by del sigma 1 square that gives us minus m by 2 sigma 1 square plus 1 by 2 sigma 1 to the power 4 sigma x i minus mu square del l by del sigma 2 square is equal to minus n by 2 sigma 2 square plus 1 by 2 sigma 2 to the power four sigma y j minus mu whole square.

Obviously, you can see that if I obtain the value of mu here from here it involves sigma 1 square and sigma 2 square. So, substituting here we get highly non-linear equations in sigma 1 square and sigma 2 square and the solutions for them cannot be obtained in the explicit form. So, numerical methods can be used to obtain the solutions.

Therefore, the question arises that what are the situations where the maximum likelihood estimator will exist or it will not exist. So, we have certain regularity conditions under which the maximum likelihood estimator always exists. Let me briefly mention about this here.

(Refer Slide Time: 37:44)



$X \sim f(x,\theta)$, $\theta \in \Omega$, → an open interval in $\mathbb{R}$

Assumptions 1. $\frac{\partial^3 \log f}{\partial \theta^3}$ exists for almost all x in $|\theta-\theta_0| < \delta$ for some $\delta > 0$

2. $E_{\theta_0}\left\{\frac{\partial \ln f(x,\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right\} = \int f'(x,\theta_0)\,dx = 0$

$E_{\theta_0}\left(\frac{f''(x,\theta_0)}{f(x,\theta_0)}\right) = \int f''(x,\theta_0)\,dx = 0$

$E_{\theta_0}\left\{\frac{\partial}{\partial \theta} \ln f(x,\theta)\Big|_{\theta=\theta_0}\right\}^2 > 0$.

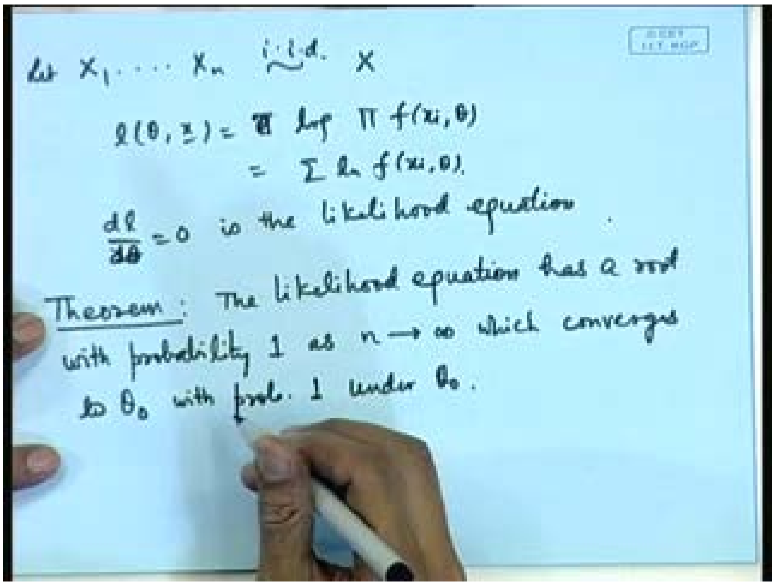
3. $\left|\frac{\partial^3 \ln f}{\partial \theta^3}\right| < \underbrace{M(x)}_{\text{where}}$ $\forall |\theta-\theta_0| < \delta$. $E_\theta M(x) < k \ \forall |\theta - \theta_0| < \delta$.

The likelihood equations we state in the following form let us have the following assumptions. So, we consider x has a distribution f x theta where theta belongs to omega and this omega is an open interval in the real line.

That means, I am considering 1 dimensional case. The assumptions are the third order derivative with respect to theta exists for almost all x in for some delta greater than 0. So, around some neighborhood of a point theta not second assumption is that at a point theta not this expectation becomes 0. Basically, it means that the density can be integrated differentiated under the integral sign now this integral is a generic notation this could be summation also in case we are dealing with the discrete distributions.

So, in particular we assume up to higher order; that means, if we consider f double prime x theta not by f x theta not where this derivative is respect to theta. Then, this should also be 0 and this square is greater than 0 and the third order derivative is bounded in a neighborhood of where m x is also integral function.

(Refer Slide Time: 40:42)



let $X_1, \cdots X_n \overset{i.i.d.}{\sim} X$

$$l(\theta, \underset{\sim}{x}) = \log \prod f(x_i, \theta)$$

$$= \sum \ln f(x_i, \theta).$$

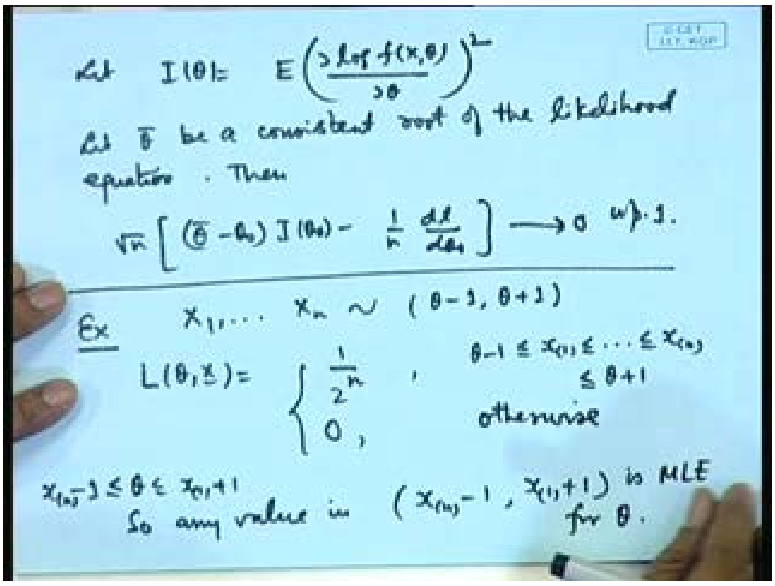
$\frac{dl}{d\theta} = 0$ is the likelihood equation .

Theorem: The likelihood equation has a root with probability 1 as $n \to \infty$ which converges to θ_0 with prob. 1 under θ_0.

So, under these assumptions on the distribution, now let us consider x 1, x 2, x n to be i i d as x and we define the likelihood equation as the log of product f x i theta that is actually sigma then d l by d theta is equal to 0 is the likelihood equation . So, we have the following result.

The likelihood equation has a root with probability 1 as n tends to infinity which converges to theta not with probability 1 under theta not. So, this is an important result that is under certain regularity conditions the maximum likelihood estimator can always be found and it always converges to a the parameter with probability one; that means, it is also consistent.

(Refer Slide Time: 42:17)



Let $I(\theta) = E\left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2$

Let $\bar{\theta}$ be a consistent root of the likelihood equation. Then

$$\sqrt{n}\left[(\bar{\theta} - \theta_0) I(\theta_0) - \frac{1}{n}\frac{dl}{d\theta_0}\right] \longrightarrow 0 \text{ w.p. } 1.$$

Ex $X_1, \ldots X_n \sim (\theta - 1, \theta + 1)$

$$L(\theta, \underset{\sim}{x}) = \begin{cases} \frac{1}{2^n}, & \theta - 1 \le x_{(1)} \le \cdots \le x_{(n)} \le \theta + 1 \\ 0, & \text{otherwise} \end{cases}$$

$x_{(n)} - 1 \le \theta \le x_{(1)} + 1$

So any value in $(x_{(n)} - 1, x_{(1)} + 1)$ is MLE for θ.

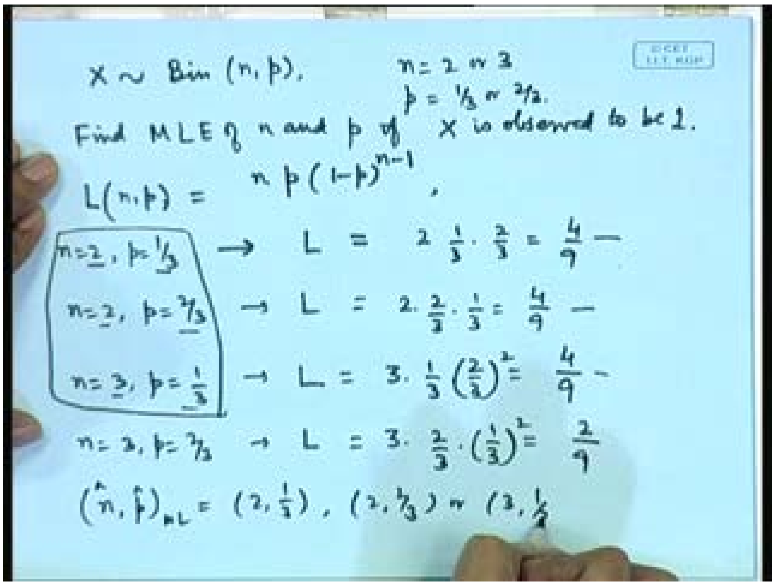
The second thing is that the asymptotic distribution is also normal. So, if I define say I theta as expectation del log f x theta by del theta whole square which is actually called the information let theta bar be a consistent root of the likelihood equation then. So, we are continuing with those assumptions the asymptotic distribution of is 0 with probability one.

We further consider the case when the maximum likelihood estimators may not be unique let us take say x 1, x 2, x n follow a uniform distribution on the intervals say theta minus 1 to theta plus 1. So, the likelihood function in this particular case will be equal to 1 by 2 to the power n for theta minus 1 less than or equal to x 1 and. So, 1 less than or equal to x n less than or equal to theta plus 1; you can see here that theta is less than or equal to x 1 plus 1 and it is also greater than or equal to x n minus 1. So, any value of theta between these 2 limits will be maximum likelihood estimator.

So, any value in the interval x n minus 1 to x 1 plus 1 is maximum likelihood estimate for theta. So, we have a situation here where the maximum likelihood estimator is not unique; however, for convenience 1 may take the average of the n points that is x 1 plus x n by 2 as the maximum likelihood estimator in this case.

Now, we consider the case where the techniques of direct differentiation or even considering like that the behavior of the likelihood function as an increasing function or decreasing function may not be appropriate. I am talking about the case where we may have to take each value 1 by 1 and then check which 1 will give the maximum likelihood estimator.

(Refer Slide Time: 45:35)



$X \sim \text{Bin}(n, p)$. $\quad n = 2$ or 3

$p = 1/3$ or $2/3$.

Find MLE of n and p if X is observed to be 1.

$L(n, p) = n\, p\, (1-p)^{n-1}$,

$n = 2,\ p = 1/3 \rightarrow L = 2 \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}$ —

$n = 2,\ p = 2/3 \rightarrow L = 2 \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{9}$ —

$n = 3,\ p = \frac{1}{3} \rightarrow L = 3 \cdot \frac{1}{3} \left(\frac{2}{3}\right)^2 = \frac{4}{9}$ -

$n = 3,\ p = 2/3 \rightarrow L = 3 \cdot \frac{2}{3} \cdot \left(\frac{1}{3}\right)^2 = \frac{2}{9}$

$(\hat{n}, \hat{p})_{ML} = (2, \frac{1}{3}),\ (2, 2/3)$ or $(3, \frac{1}{3}$

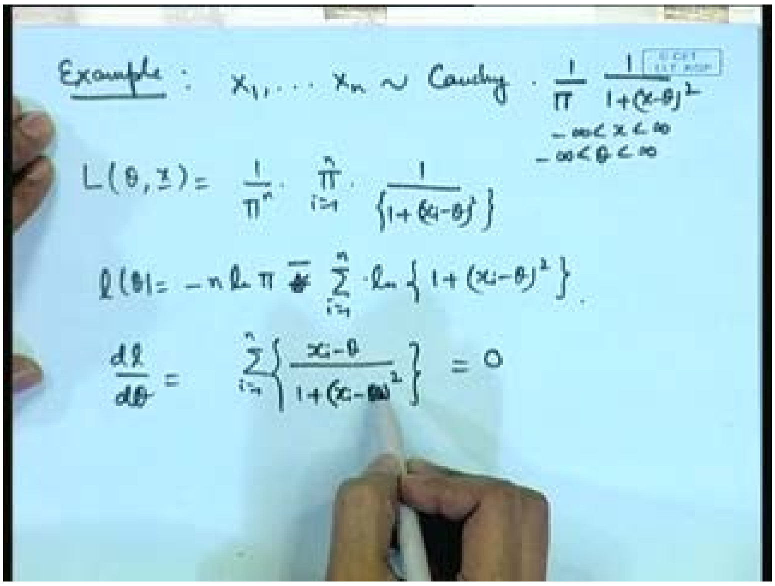
Let me explain through an example. Suppose, I consider binomial n p and n is either 2 or 3 and p is either 1 by 3 or 2 by 3, now we want to find out the maximum likelihood estimator of n and p. If x is observed to be 1 now this is the situation where we actually write down the values of the likelihood function at each of these parameter values that is n is equal to 2 p is equal to 1 by 3 n is equal to 2 p is equal to 2 by 3 etcetera and then see which one is the largest.

Let us write down the likelihood function here the likelihood function here is n c x. So, since x is 1. So, it is n p to the power x means p 1 minus p to the power n minus 1. So, since x is equal to 1 is already observed we are having exactly these values. So, we have the four values here when n is equal to 2 p is equal to 1 by 3 corresponding to this the likelihood function is equal to twice 1 by 3 into 2 by 3. That is equal to 4 by 9 when n is equal to 2 and p is equal to 2 by 3 the likelihood function value turns out to be twice 2 by 3 1 by 3. That is again 4 by 9 when n is equal to 3 and p is equal to 1 by 3 the likelihood function value is equal to 3 1 by 3 2 by 3 square. That is equal to 4 by 9 again and n is equal to 3 p is equal to 2 by 3 here the likelihood function value turns out to be 3 2 by 3 1 by 3 square which is equal to 2 by 9.

If we are looking at the maximization of the likelihood function then, you can observe here that this value this value and this value they are all same and they are the maximum this value is actually smaller; that means, any of the configurations 2 1 by 3 2 2 by 3 3 1 by 3 they are as likely to give the maximum value as any other value.

Therefore, the maximum likelihood estimator for n and p can be considered to be pair 2 1 by 3 2 2 by 3 or 3 1 by 3. Now, this is another case because here four possible configurations are there out of that 3 are corresponding to the maximum likelihood value.

(Refer Slide Time: 48:52)



Example : $x_1, \dots x_n \sim$ Cauchy . $\frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$

$-\infty < x < \infty$
$-\infty < \theta < \infty$

$$L(\theta, \underline{x}) = \frac{1}{\pi^n} \cdot \prod_{i=1}^{n} \cdot \frac{1}{\{1+(x_i-\theta)^2\}}$$

$$l(\theta) = -n \ln \pi - \sum_{i=1}^{n} \ln \{1+(x_i-\theta)^2\}.$$

$$\frac{dl}{d\theta} = \sum_{i=1}^{n} \left\{ \frac{x_i-\theta}{1+(x_i-\theta)^2} \right\} = 0$$

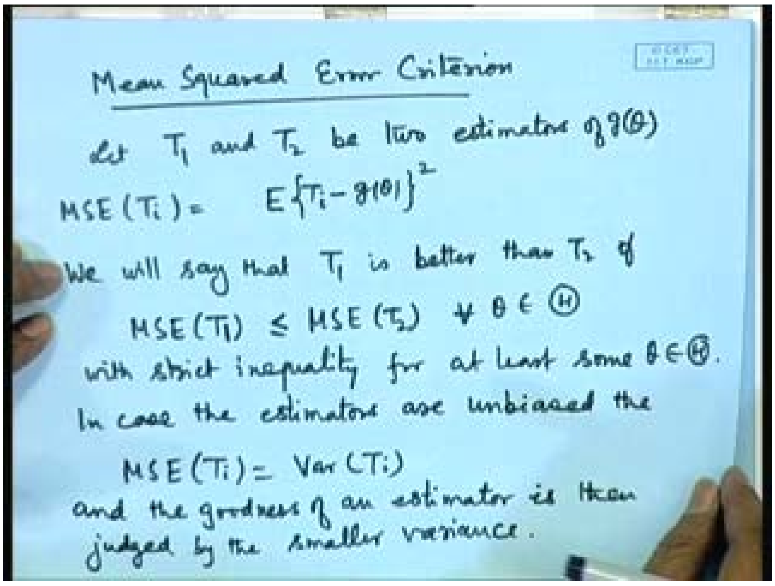
Let us take another example where the argument may follow a different path. Consider say x 1, x 2, x n following say Cauchy distribution with the density function 1 by pi 1 by 1 plus x minus theta square where x is any real number theta is any real number let us look at the likelihood function that is equal to 1 by pi to the power n and product I is equal to 1 to n 1 by 1 plus x i minus theta square.

So, log likelihood function is then equal to minus n plus sigma I is equal to 1 to n. So, log of this term we can write it as a minus 1 plus x i minus theta square. So, d l by d theta is equal to 0 that is the likelihood equation will be equal to i is equal to 1 to n x i minus theta divided by 1 plus x i minus theta square this is equal to 0. You can easily see that it is a non-linear equation and the explicit solution does not exist.

So, 1 had to use numerical methods for finding out the solution of this equation. There is another example where we may not have the solution of the likelihood equation in the direct form; however, if the assumptions that we mentioned earlier are satisfied then we can prove that the solution will exist with probability 1 and it will be consistent estimator.

Now, next we come to the concept of judging the goodness of the estimators. So, for judging the goodness of the estimators 1 may consider the variability aspect for example, unbiased is 1 judgment because, whether the estimator is biased or unbiased. So, if estimator is unbiased it will be considered to be better than the biased estimator if an estimator is consistent another estimator is inconsistent then again we may consider the consistent estimator to be better than the inconsistent estimator; however, if we are having several consistent estimators or several unbiased estimators then how to compare among them.

(Refer Slide Time: 51:44)



Mean Squared Error Criterion

Let T_1 and T_2 be two estimators of $g(\theta)$

$MSE(T_i) = E\{T_i - g(\theta)\}^2$

We will say that T_1 is better than T_2 if

$MSE(T_1) \leq MSE(T_2) \ \forall \ \theta \in \Theta$

with strict inequality for at least some $\theta \in \Theta$.

In case the estimators are unbiased the

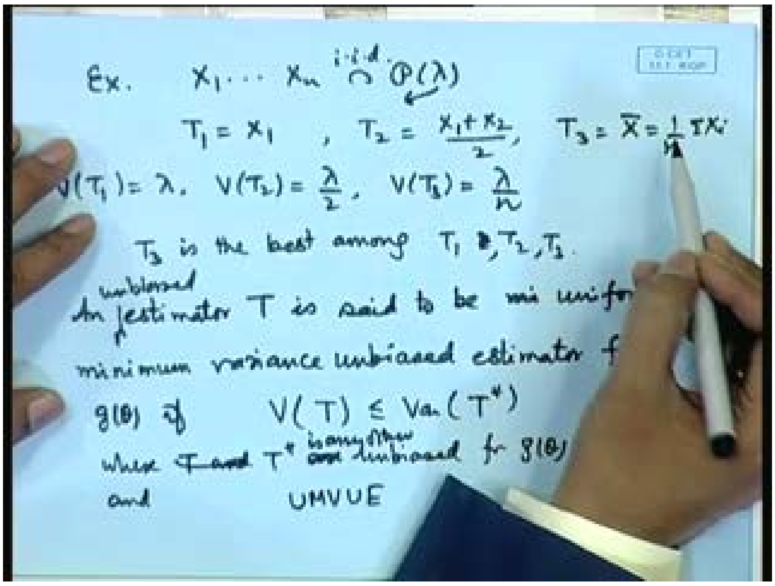
$MSE(T_i) = Var(T_i)$

and the goodness of an estimator is then judged by the smaller variance.

One of the popular criteria is to look at the variability of the estimator. So, we have the. So, called mean squared error criteria let T 1 and T 2 be 2 estimators of theta say g theta then the mean squared error of T i is defined as expectation of T i minus g theta square. So, this is giving a measure of variability of the estimator and we will say that t 1 is better than T 2 if mean squared error of T i is T 1 is less than or equal to the mean squared error of t 2 for all parameters with strict inequality for at least some theta.

Now, if the estimators are unbiased suppose T i is unbiased for g theta then this is reducing to the variance of T i and in that case the criteria can be written as in case the estimators are unbiased the mean squared error of T i simply becomes variance of T i and the goodness of an estimator is then judged by the smaller variance that is the smaller the variance the estimator is better.

(Refer Slide Time: 54:13)



Ex. $X_1 \cdots X_n \overset{i.i.d.}{\sim} \mathcal{P}(\lambda)$

$T_1 = X_1$, $T_2 = \frac{X_1+X_2}{2}$, $T_3 = \bar{X} = \frac{1}{n}\Sigma X_i$

$V(T_1) = \lambda$. $V(T_2) = \frac{\lambda}{2}$. $V(T_3) = \frac{\lambda}{n}$

T_3 is the best among T_1, T_2, T_3.

An unbiased estimator T is said to be ~~mi~~ unifo minimum variance unbiased estimator f $g(\theta)$ if $V(T) \leq Var(T^*)$

where ~~T and~~ T^* is any other ~~are~~ unbiased for $g(\theta)$

and UMVUE

As a simple example, let us consider the exercise considered yesterday. Suppose, I am having x 1, x 2, x n following say, Poisson lambda distribution and I am considering estimation of lambda. So, let me write estimator T 1 as x 1 this is unbiased. Let me write estimator T 2 as x 1 plus x 2 by 2 let me write estimator say T 3 as x bar which is actually the mean of all the observations now let us look at variance of T 1 that is lambda if I look at variance of T 2 that is lambda by 2 if I look at variance of T 3 that is lambda by n. So, clearly here T 3 is the best among T 1 and T 2 among T 1 T 2 and T 3.

This gives a procedure for checking that which estimators are better now among the unbiased estimators the 1 which has the smallest variance we call it minimum variance unbiased estimator. So, an estimator T is said to be uniformly minimum variance unbiased estimator for g theta if variance of T is less than or equal to variance of say T star where T and T star are unbiased for g theta. So, an unbiased estimator T is said to be uniformly minimum variance unbiased estimator for g theta if variance of T is less than or equal to variance T star where T star is any other. For any other estimator, if variance is smaller for T then definitely it is having the smallest. So, we use the technology u m v u e.

For example, in this case of Poisson distribution x bar will be uniformly minimum variance unbiased estimator now the question arises that, how to check that? it is uniformly minimum variance unbiased estimator or how to find the uniformly minimum variance unbiased estimator because here we have already got it and then we can check, but then the total number of estimators are infinite and therefore, how to find that? So, we have to develop a method for finding out the U M V U E.

So, in the next lecture we will be considering these methods there is some additional terminology called sufficiency and completeness which is quite useful also there are methods of obtaining lower bounds which can be found. So, in the next class we will be considering that.