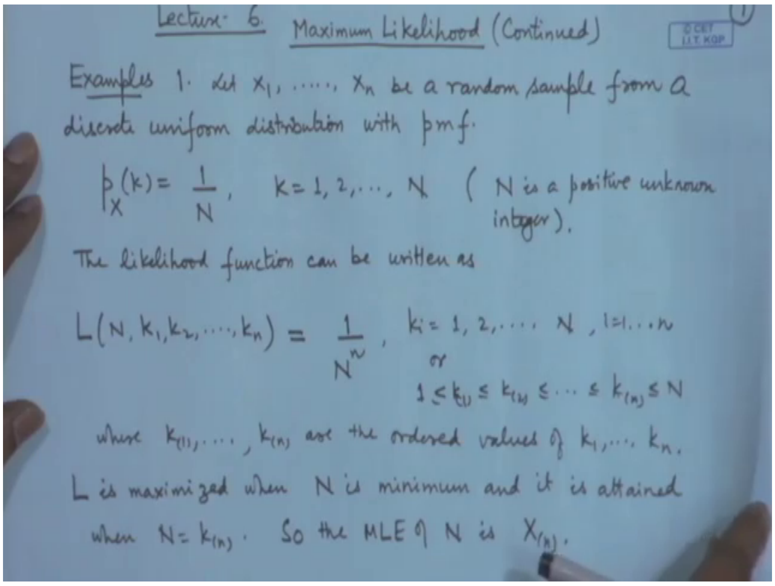
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 11
Finding Estimators – [vocalized-noise]

Yesterday, we have discussed in detail various models; we have various probability models, and how to find out the maximum likelihood estimator for that. We have seen here that the effect of changing the parameter space or effect of the prior information on the parameter space plays an important role in the maximum likelihood estimation, which makes it different from other methods such as unbiased estimation or the method of moment's estimation. So, today I will explain this method with the help of several other examples, and we will discuss certain important large sample properties of the maximum likelihood estimators.

(Refer Slide Time: 01:05)



Lecture 6. Maximum Likelihood (Continued)

Examples 1. Let $X_1, \dots, X_n$ be a random sample from a discrete uniform distribution with pmf.

$$p_X(k) = \frac{1}{N}, \quad k = 1, 2, \dots, N \quad (\text{N is a positive unknown integer}).$$

The likelihood function can be written as

$$L(N, k_1, k_2, \dots, k_n) = \frac{1}{N^n}, \quad k_i = 1, 2, \dots N, \ i=1 \dots n$$
or
$$1 \le k_{(1)} \le k_{(2)} \le \dots \le k_{(n)} \le N$$

where $k_{(1)}, \dots, k_{(n)}$ are the ordered values of $k_1, \dots, k_n$.

L is maximized when N is minimum and it is attained when $N = k_{(n)}$. So the MLE of N is $X_{(n)}$.

Let me start with the a couple of examples on discrete distributions . So, let us consider say a discrete uniform distribution. Let X 1, X 2, X n be a random sample from a discrete uniform distribution. So, a discrete uniform distribution is usually concentrated on n points, and normally we take the points from 1 to n, and each one will be equal probability. So, we can consider the probability mass function as follows with probability mass function given by, so we write p X k is equal to 1 by n, where k can take values 1 to

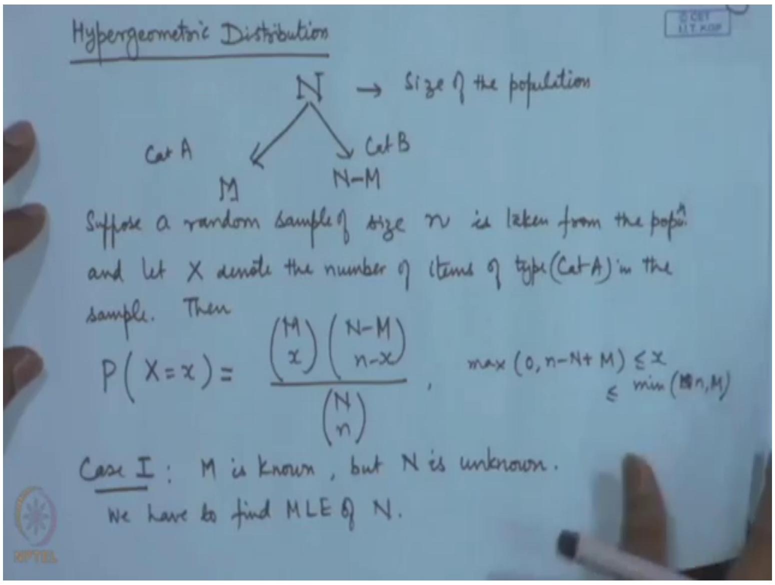
n. Now, in this case, there may not be any inference problem if we know on how many points the distribution is concentrated. The inference problem arises if we do not know how many points are there. So, this type of situation may arise where we know that each possibilities with equal probability, but how many possibilities are there that may not be known. So, in that case we may be interested in estimating that number.

So, we are assuming here that n is a positive unknown integer. So, we proceed as before we write down the likelihood function which is the glint distribution of X 1, X 2, X n. So, we consider points, X 1 is equal to K 1; X 2 is equal to K 2; X n is equal to k n. So, we can write it in the following fashion the likelihood function can be written as L N. And as I mentioned we are considering the points K 1, K 2, K n which are the observed values of the random variables X 1, X 2, X n respectively, so that is equal to 1 by N to the power n where each of the i s can take values 1 to n for i is equal to 1 to n.

Now, the problem here is to maximize this function with respect to N as n is appearing in the denominator it will be the minimum value of n. So, this will be maximized when N is taking the minimum value. Now, what is the minimum value of n that is possible here? So, this region we can write it in a more appropriate fashion that K 1 suppose I order them K 2 up to K n, then the region can be written as 1 less than or equal to K 1 less than or equal to K 2 less than or equal to K N less than or equal to N. So, from here it is clear that the minimum value of N that is possible is the maximum value of K 1, K 2, K n, where K 1, K 2, K n are the ordered values of K 1, K 2, K n.

So, L is maximized, when N is minimum and it is attained when N is equal to K n. Now, K n corresponds to the largest order statistics here. So, we conclude that the maximum likelihood estimator of N is X n. You can notice the analogy with the continuous uniform distribution which we discussed in the previous class. In the continuous uniform distribution on the interval 0 to theta, the maximum likelihood estimator for theta was also the largest order statistics that is X n. So, in the discrete uniform case also the same thing is happening; the only different here is that here Xi's are taking positive integral values here .

(Refer Slide Time: 06:38)



Hypergeometric Distribution

N → Size of the population

Cat A

M

Cat B

N−M

Suppose a random sample of size n is taken from the popn and let X denote the number of items of type (Cat A) in the sample. Then

$$P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad \max(0, n-N+M) \le x \le \min(n, M)$$

Case I: M is known, but N is unknown.

We have to find MLE of N.

Let us take another important discrete distribution that is hypergeometric distribution . Now, a hypergeometric distribution is usually considered in the following fashion that there is a large population of size N; this is the size of the population. Now, this population is divided into two parts let us say category A and category B. The entire population for example, we may divide a employees in a of an organization by two categories that is those who are in the supervisory position and those who are in the working conditions in the that is there they are the lower level employees and the higher level employees. We may divide the patients into two groups; say those who are having communicable diseases, those who do not have communicable diseases. We may divide a section of a students into the students who are following engineering discipline, and the others who are studying say medical discipline.

So, we have a large population and the population size of one category is N, and therefore, the other category population has N minus M numbers. Suppose, we random we take a random sample; a random sample of size small n is taken from the population. And let X denote the number of items, items may this could be persons anything of type category A in the sample. Then the probability distribution of X is given by M c x N minus c M minus M c n minus x divided by N c n.

Now, obviously, this random variable X, it can take values from 0, 1 to n, because in a random sample of size n you may have none of this category, and all of the other category, one of one category, n minus 1 of another category and so on. However, this is also subject to the restrictions of the total elements of each type, and therefore, we may write the restrictions in a more strict sense as that is x is a integer between maximum of 0 and n minus N plus M to minimum of n, M.

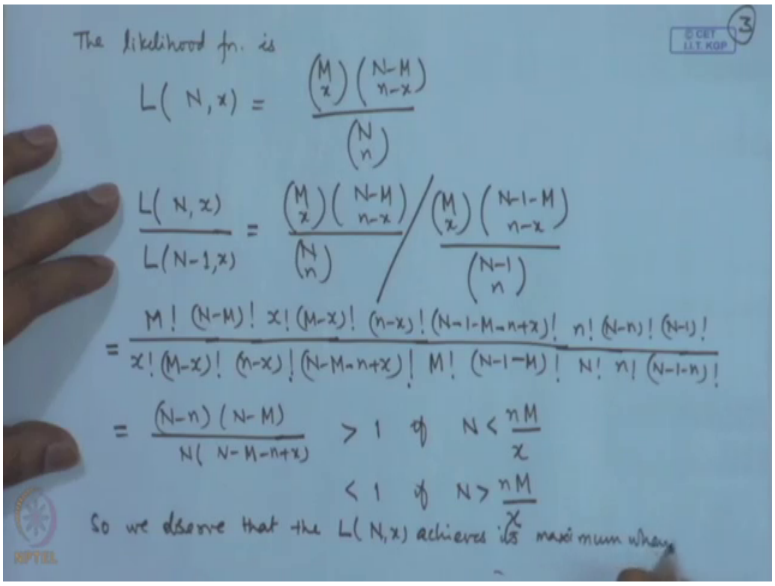
Now, when we look at this probability model, there can be two different cases one case could be that the total population size is unknown. Now, this type of situation arises for example, in estimating a say we have a lake and a company which is involved in the fishing, it may like to estimate that how much of fish amount will be available in the lake if they start the fishing operations. Now, obviously, one cannot take out the water from the lake and count the how many fish will be there.

So, we assume that the size of the population that is capital N is unknown. Now, one may conduct a the following experiment which is known as capture recapture technique we take a random sample of size capital N from the lake the fish that are taken out they are tagged, that means, they are mart with something then they are shifted back to the lake. So, that they get mixed up with the entire population of the fish. Later on we consider a random sample of size N from the fish once again from the lake we again take a random sample of size N. Now, out of that you look at how many of them are tagged and how many of them are untagged. So, now this capital M is known to you and capital N is not known to us. And the problem will come how to estimate capital N.

Similarly, there can be another problem where the total population size is known. We may like to estimate how many people are suffering from a different disease or a certain virus for example, how many people are infected with HIV virus. In that case, we again take a sample of size N. And in that sample X will denote the number of people who are actually infected with the virus and then on the basis of that we estimate N. So, in this case capital N may be known, but capital M is unknown.

So, when we consider this hypergeometric model, there are two cases. So, case one is that M is known, but N is unknown. So, in this case we have to find the maximum likelihood estimator of N. In order to do that, we write the likelihood function, now in this case the observation is the sample of size N has been taken and X is the number of items of type category A. So, this is the recorded item. So, this function itself denotes the likelihood function in this particular case, because this is the probability mass function of the observation here.

(Refer Slide Time: 13:50)



The likelihood fn. is

$$L(N,x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\frac{L(N,x)}{L(N-1,x)} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \Bigg/ \frac{\binom{M}{x}\binom{N-1-M}{n-x}}{\binom{N-1}{n}}$$

$$= \frac{M!\,(N-M)!\,x!\,(M-x)!\,(n-x)!\,(N-1-M-n+x)!\,n!\,(N-n)!\,(N-1)!}{x!\,(M-x)!\,(n-x)!\,(N-M-n+x)!\,M!\,(N-1-M)!\,N!\,n!\,(N-1-n)!}$$

$$= \frac{(N-n)(N-M)}{N(N-M-n+x)} \quad > 1 \text{ if } N < \frac{nM}{x}$$

$$< 1 \text{ if } N > \frac{nM}{x}$$

So we observe that the $L(N,x)$ achieves its maximum when

So, the likelihood function is let me call it L N here, so that is equal to M c x N minus M c N minus x N c n. And we need to maximize this with respect to capital N. Now, the methods that I mentioned in the previous examples cannot be directly implemented here. The main reason is that here N is an integer. So, we cannot apply differentiation procedure, taking log etcetera. So, we carry out a different analysis. Let us write down we try to see the increasing or decreasing nature of this function in a straight forward fashion.

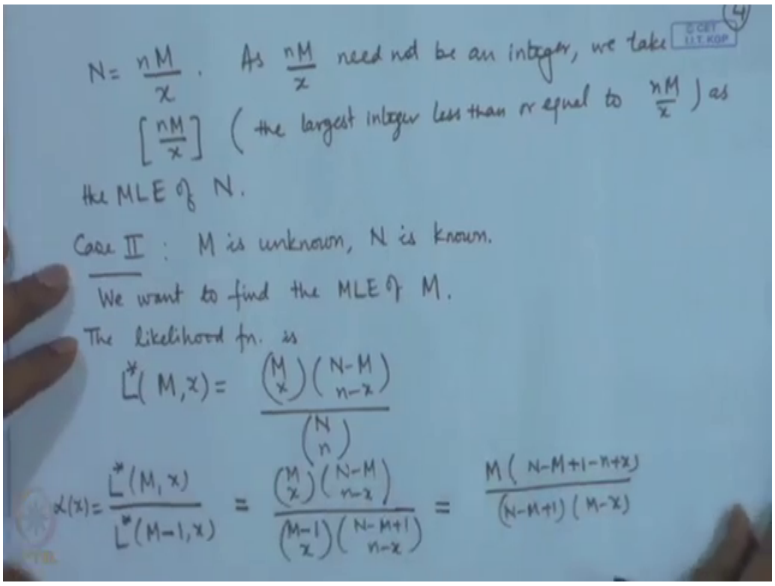
Let us consider for example, the value of the likelihood function at N, and the value of the likelihood function of N minus 1 . So, this is M c x c N minus M c n minus x divided by N c n, and then this whole thing is divided by M c x N minus 1 minus M c n minus x divided by N minus 1 c n. Ah We may expand the factorials here. So, you will get M factorial divided by x factorial. So, this entire thing comes out to be like M factorial divided by x factorial M minus x factorial, then we have N minus M factorial divided by n minus x factorial, then N minus M minus n plus x factorial .

This whole thing is then divided by these terms. So, M factorial x factorial into M minus x factorial, then we have N minus 1 minus M factorial n minus x factorial, and then n minus 1 minus M minus n plus x factorial . Then further we have n minus and then we had this N c n and n minus 1 c n. So, we write that also N factorial n factorial N minus n factorial. And in a similar way this will be n minus 1 factorial n factorial N minus 1 minus n factorial. So, it is easy that one can simplify these terms and we get it as N minus n into N minus M divided by N into N minus M minus n plus x.

Now, you notice that this is greater than 1, if N is less than n M by x; and it is less than 1, if N is greater than n M by x. Now, obviously, you can see N is taking integer values from 1, 2 and so on. Now, this ratio that is L N x divided by L N minus 1 x. So, what we are observing here is that if I increase N, if I from N minus 1 to N if I go, then this ratio is greater; that means, it is an increasing function of N when N is less than n M by x.

And when N is bigger than n M by x, then this value starts decreasing therefore, you can say that this function increases till this and then decreases. Therefore, the maximum of L N function is achieved when N is equal to n M by x. Now, naturally n M by x need not be an integer although x, N and M are integers, but this expression need not be an integer. So, we may take the integral portion of n M by x as the maximum likelihood estimator for N.

(Refer Slide Time: 19:11)



$N = \frac{nM}{x}$. As $\frac{nM}{x}$ need not be an integer, we take $\left[\frac{nM}{x}\right]$ (the largest integer less than or equal to $\frac{nM}{x}$) as the MLE of N.

Case II : M is unknown, N is known.

We want to find the MLE of M.

The likelihood fn. is

$$\overset{*}{L}(M,x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\alpha(x) = \frac{\overset{*}{L}(M,x)}{\overset{*}{L}(M-1,x)} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{M-1}{x}\binom{N-M+1}{n-x}} = \frac{M(N-M+1-n+x)}{(N-M+1)(M-x)}$$

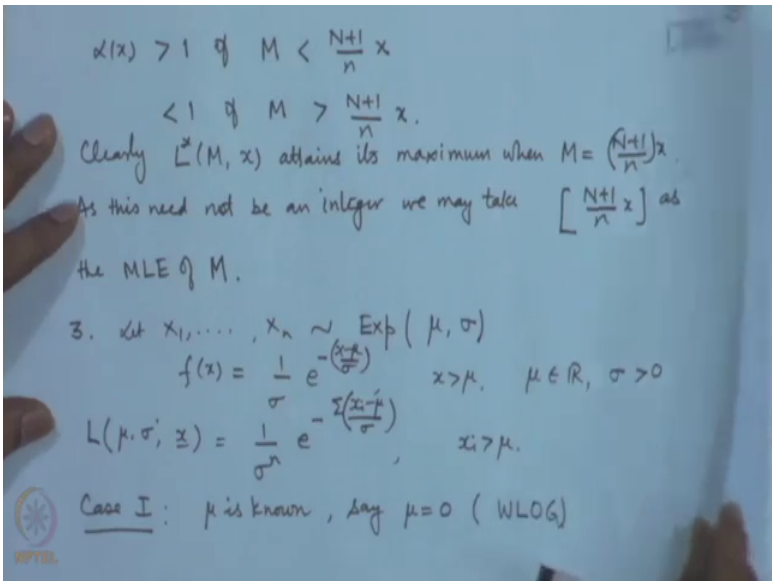
So, we observe that the L function achieves its maximum when N is equal to n M by x. As n M by x need not be an integer, we take n M by x integral portion that is the largest integer less than or equal to n M by x as the maximum likelihood estimator of N. Now, let us take up the other case when M is unknown M is unknown and N is known, so here we want to find out the maximum likelihood estimator of M.

Now, once again if you consider this likelihood function here, I wrote here it as a function of N, because this is coming from the probability mass function of x, here M and N both are involved. Now, if N is known and M is unknown, I will consider the likelihood function as a function of N, so the likelihood function will become although it will be the same expression, it will be written as L M, x that we call it L star, so this is M c x N minus M c n minus x.

Now, as before we have to consider the maximization of this with respect to M. Now, M is an integer and the factorials are involved here. Therefore, one cannot apply the usual methods of analysis such as differentiation etcetera. Rather, we try to see the behavior of this in a straight forward fashion.

So, once again we write L star M, x divided by L star M minus 1 x. Now, that is equal to M c x N minus M c n minus x, then we write this ratio N c n will be same, so that will cancelled out, and we will get M minus 1 c x N minus M plus 1 c n minus x. Now, as before we can simplify this and the terms turns out to be M into N minus M plus 1 minus n plus x divided by N minus M plus 1 into M minus x. Now, once again we observe that this ratio let me call it say alpha x.

(Refer Slide Time: 22:50)



$\alpha(x) > 1$ if $M < \frac{N+1}{n} x$

< 1 if $M > \frac{N+1}{n} x.$

Clearly $L^*(M, x)$ attains its maximum when $M = \left(\frac{N+1}{n}\right)x$.

As this need not be an integer we may take $\left[\frac{N+1}{n} x\right]$ as the MLE of M.

3. Let $X_1, \ldots, X_n \sim \text{Exp}(\mu, \sigma)$

$$f(x) = \frac{1}{\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)} \quad x > \mu. \quad \mu \in \mathbb{R}, \sigma > 0$$

$$L(\mu, \sigma; \underset{\sim}{x}) = \frac{1}{\sigma^n} e^{-\Sigma\left(\frac{x_i-\mu}{\sigma}\right)}, \quad x_i > \mu.$$

<u>Case I</u>: μ is known, say $\mu = 0$ (WLOG)

So, if we observe this ratio, alpha x is greater than 1 if M is less than N plus 1 by n x, and it is less than 1 if M is greater than N plus 1 by n x. So, we can easily see that the L star function, it is increasing for M less than N plus 1 by n x, and it will start decreasing for M greater than this. Therefore, the maximum will be attained at N plus 1 by n x. And therefore, we can consider the integral part of this as the maximum likelihood estimator for n.

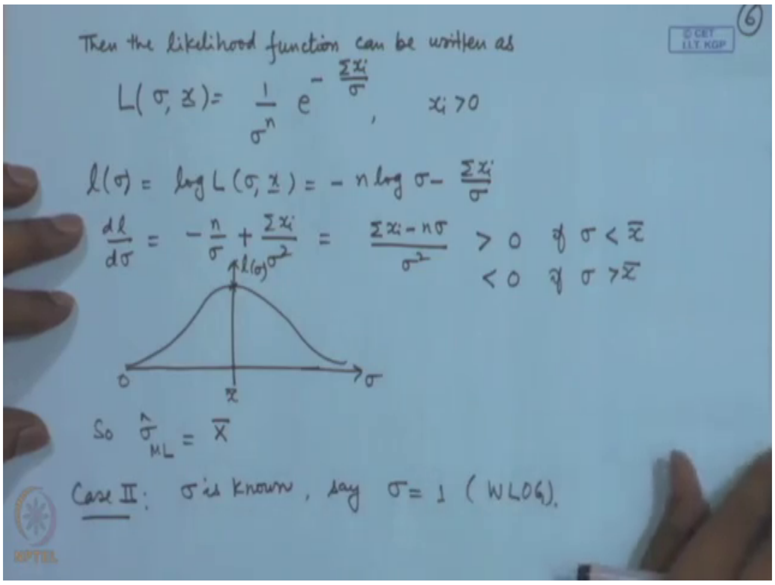
Clearly, L is the M attains its maximum when M is equal to N plus 1 by n x. As this need not be an integer, we may take the integral portion of this as the MLE of M. So, here we have seen that in the discrete case the method of obtaining the maximum likelihood estimators differs little bit.

We have not considered another important distribution, which arises quite often in statistical modeling that there is a exponential distribution. Now, the exponential distribution once again has two parameters, it may have a scale parameter, it may have a location parameter. So, I will consider a general model, and then we look at the solution here.

Let X 1, X 2, X n follow exponential mu, sigma estimation, when I say this we are writing down the density function as 1 by sigma e to the power minus x minus mu by sigma, where x is greater than mu. Here mu can be any real number, and sigma is positive. In the usual study, which are related to reliability and life testing. There mu is considered as the minimum guarantee time, and there mu will be positive, but in many other applications it need not be so. So, I am taking the general case here mu can take any real value, and sigma of course is associated with the average, therefore sigma is greater than 0.

So, we consider the likelihood function here 1 by sigma to the power n e to the power minus sigma x i minus mu by. Now, when we are dealing with the two parameter situation, one may have different cases, it may happen that the minimum guarantee time is fixed, and therefore we may take it to be 0. It may happen that sigma is fixed and therefore we may take it to be 1. So, we consider these cases. So, case 1 let us consider say mu is known, so we may take without laws of generality this to be 0 if that is so, then we may write the likelihood function. If we substitute mu is equal to 0, the form of this function becomes much simple.

(Refer Slide Time: 27:09)



Then the likelihood function can be written as

$$L(\sigma, \underset{\sim}{x}) = \frac{1}{\sigma^n} e^{-\frac{\sum x_i}{\sigma}}, \quad x_i > 0$$

$$l(\sigma) = \log L(\sigma, \underset{\sim}{x}) = -n \log \sigma - \frac{\sum x_i}{\sigma}$$

$$\frac{dl}{d\sigma} = -\frac{n}{\sigma} + \frac{\sum x_i}{\sigma^2} = \frac{\sum x_i - n\sigma}{\sigma^2} \begin{array}{l} > 0 \text{ if } \sigma < \bar{x} \\ < 0 \text{ if } \sigma > \bar{x} \end{array}$$

So $\hat{\sigma}_{ML} = \bar{X}$

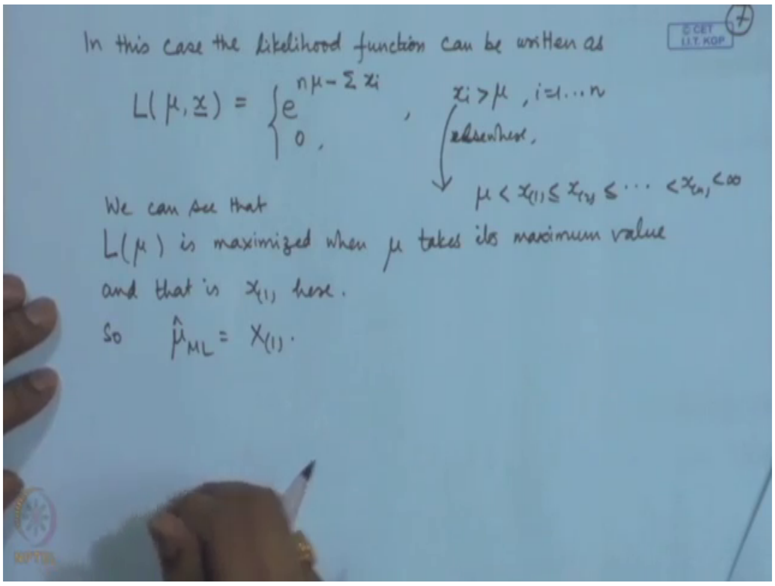
Case II: σ is known, say $\sigma = 1$ (WLOG).

And we get it as then the likelihood function can be written as L sigma, x as 1 by sigma to the power n e to the power minus sigma x i by sigma where each x i will be greater than 0. So, we write down the large likelihood function that is equal to minus n log of sigma minus sigma x i by sigma. So, now, this is a straight forward function for sigma we can consider the derivative with respect to sigma and we get minus n by sigma minus sigma x i. So, this will become plus sigma x i by sigma square which gives us sigma x i minus n sigma by sigma square. Obviously you can study its behavior, it will be greater than 0 if sigma is less than x bar; it will be less than 0 if sigma is greater than x bar.

So, if we consider the plotting of the curve as a function of sigma if we plot L sigma, now sigma if of course, positive, so this is starting from 0. So, this is increasing till x bar and thereafter it is decreasing, because our derivative is positive for sigma less than x bar, and it is less than 0 for sigma greater than x bar. Therefore, easily you can see that the maximum occurs at x bar. So, the maximum likelihood estimator of sigma turns out to be the mean of the distribution.

Now, here as before like we have considered in the normal distribution, one may have additional information about sigma. For example, sigma may be having an upper bound such as sigma less than or equal to sigma naught or sigma greater than or equal to sigma naught or sigma li[e] may lie in an interval. In that case the solutions will for the maximum likelihood estimator will get modified accordingly as we have discussed in the case of normal distribution. So, I will be skipping those descriptions here. Let us take up the second case when sigma is known when sigma is known we can take it to be 1 without loss of generality.

(Refer Slide Time: 30:41)



In this case the likelihood function can be written as

$$L(\mu, \underline{x}) = \begin{cases} e^{n\mu - \sum x_i}, & x_i > \mu,\ i=1\ldots n \\ 0, & \text{elsewhere.} \end{cases}$$

$$\downarrow \quad \mu < x_{(1)} \leq x_{(2)} \leq \cdots < x_{(n)} < \infty$$

We can see that

$L(\mu)$ is maximized when μ takes its maximum value and that is $x_{(1)}$ here.

So $\hat{\mu}_{ML} = x_{(1)}$.

Now, in this case the likelihood function can be written as so this is now a function of sigma mu because sigma is known. So, if you look at the form that I discussed here 1 by sigma to the power n e to the power minus sigma x i minus mu by sigma. So, here if I put sigma is equal to 1 this term vanishes, and you are left with only the exponent term which I can simply write as e to the power n mu minus sigma x i, so e to the power n mu minus sigma x i. And of course, each x i is greater than mu. And obviously, this is 0 if let me say elsewhere, each of x i has to be greater than mu in this particular case.

Now, if you look at this function, we have to maximize this with respect to mu here. And this n mu is occurring in the exponent without any multiplication or any other involvement of any other term. So, naturally you can easily see that the maximization will occur for the maximum value of mu. Now, what is the maximum possible value of mu? Now, mu is less than each of the xi s therefore, this region can be written as mu less than x 1 less than x 2 and so on. Therefore, the maximum value of mu can be only x 1. So, the maximum likelihood estimator of mu is x 1 in this case. We can see that L mu is maximized when mu takes its maximum value and that is x 1 here. Once again here this x 1, x 2, x n denote the order statistics of the original observations. So, mu hat ML is equal to the minimum of the observations here. So, you have seen in the uniform distribution, we got the maximum of the observations. And in this particular case, we are getting the minimum of the observations.