Statistical Methods for Scientists and Engineers

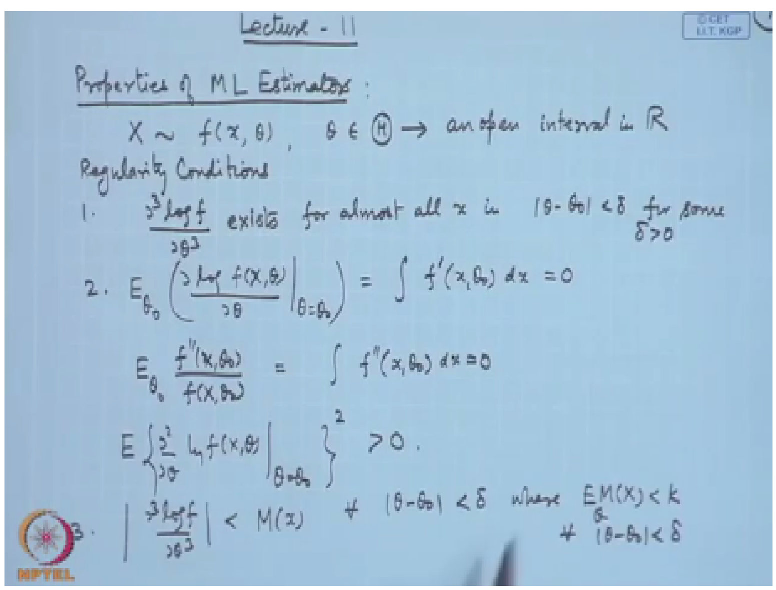
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture – 11
Parametric Methods - III

In the last class, I have introduced methods of estimation and one of them was the method of moments and the other was the method of maximum likelihood estimation. Now, classically speaking in the chronological order; the method of moments was given first by the British statistician Karl Pearson around 1900 and thereafter, it was used for quite some time. However, around 1922 onwards R. A. Fisher, he proposed a new method that is called the maximum likelihood estimation.

And actually, when the popularity of the maximum likelihood estimation is stems from the fact that the estimators, which are obtained by this method, are more efficient and they satisfy certain asymptotic properties also. So, first I will describe a few properties of the maximum likelihood estimators and then we will look at a couple of examples before moving on to other methods here.

(Refer Slide Time: 01:24)



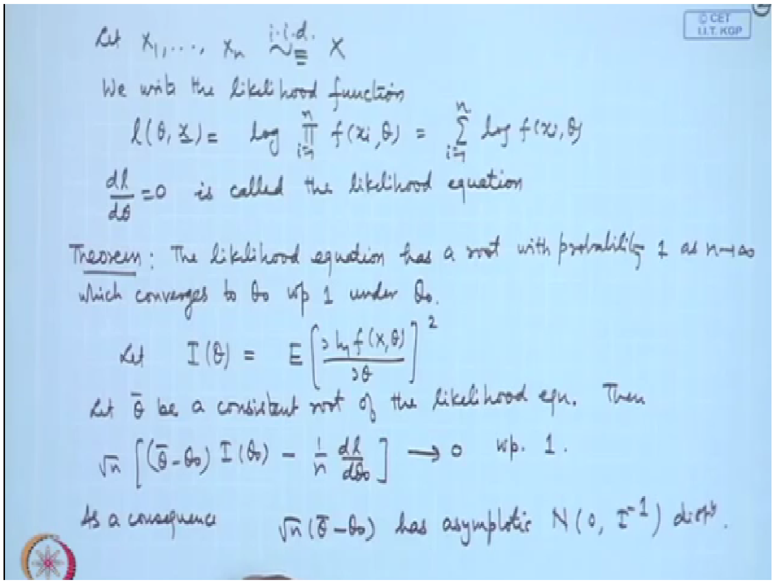
So, let us consider some properties of maximum likelihood estimators. Now, these properties are proved under certain conditions, these are called regularity conditions. So, in general we have the model that we have the observables from a distribution, which may have a probability mass function or a probability density function, which we described by a fx theta and belongs to the parameter space a script theta.

Usually, if I am considering one dimensional parameters, then theta is considered to be an open interval in the real line. For example, if I consider say, Poisson distributions, so parameter lambda is positive, so 0 to infinity as an open interval in R suppose we are considering say, normal description with mean Mu and variance unity, then Mu is considered to be the whole real line. So, in many of the practical problems, this condition is always satisfied.

So, we have some regularity conditions; we assume that the third order partial derivative of the log f exists for almost all x in the interval; theta – theta0 less than delta for some delta positive, so we assume that theta0 is the true value of the maximum likelihood estimator, theta0 is a true value of the parameter and then in the interval; in an interval around that, we assume; we also assume expectation theta0 del log fx theta/del theta give in; add theta = theta0 to be f prime x theta0 dx = 0.

So, when I am writing the integral actually, I am assuming the continuous case, where f is a density, however a similar statement can be written, if I am assuming discrete case and this integral will be replaced by the summation sign and here this f prime denotes the derivative with respect to theta and then the value is taken at theta = theta0, then we further assume second order derivative condition also.

(Refer Slide Time: 05:25)



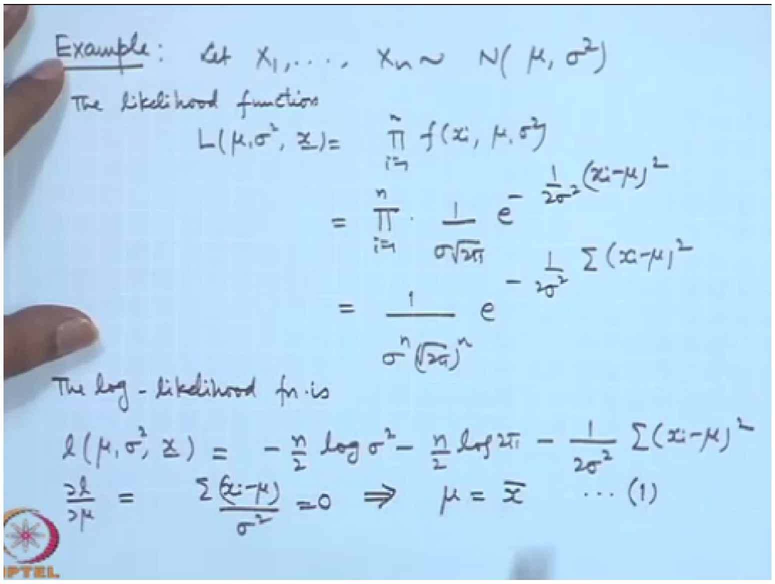
We assume that the second order derivative is strictly positive. We assume a boundedness condition for the third order derivative that it is less than some Mx for all theta in a neverhood of theta0, this is integrable. Basically, we are assuming expectation is bounded again for all theta – theta0 less than some delta.

Let us write down the likelihood equation, so we have a random sample x1, x2, xn, which is having the same distribution as X, we write the likelihood function, we call it say; l theta x, which is nothing but the log of the joint density of x1, x2 and xn, which is actually = sum of log of fxi theta; i = 1 to n.

So, dl/ d theta = 0, this is called the likelihood equation, so we have the following result, which I state in the form of a theorem. The likelihood equation has a root with probability 1 as n becomes large, which converges to theta0 with probability 1 under theta0. So, this is consistently; even say strongly consistent here. Further we have efficiency result here; if I define say, the information I theta to be expectation of del log fx theta / del theta is square. Let theta bar be a consistent root of the likelihood equation.

Then the square root n theta bar – theta0, I theta0 – 1/n dl/ d theta0 goes to 0 with probability 1. As a consequence, square root n theta bar – theta0 has asymptotic normal 0, I inverse distribution that is as asymptotic normality is also satisfied. So, these are some of the; you can say desirable is strong properties of the maximum likelihood estimator and which made it a very popular method of estimation.

(Refer Slide Time: 08:48)

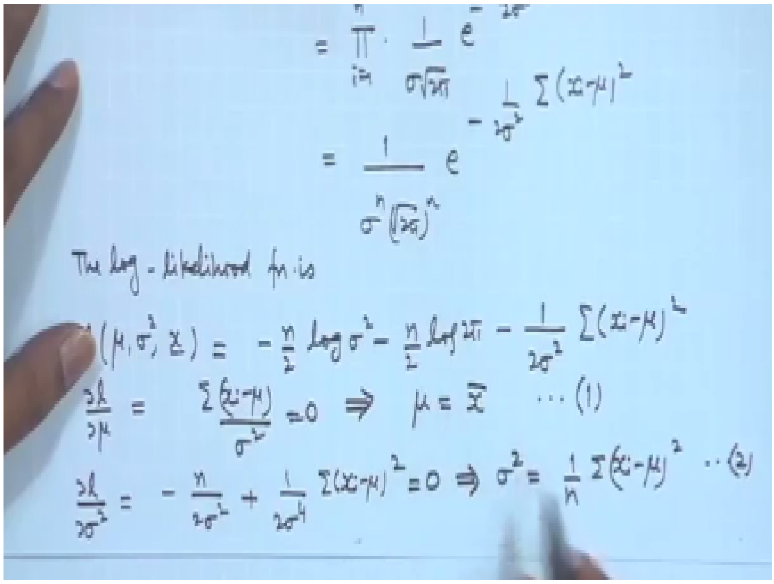


Now, let me give some example here, so let us consider say, x1, x2, xn follows normal Mu sigma square distribution. Now, when we have normal Mu square distribution, the likelihood function; in fact, this was actually the log likelihood function, so we write the likelihood function, which is say L here it is function of Mu and sigma square that is product of the densities that is = product I = 1 to n; 1/ sigma root 2 pi e to the power -1/2 sigma square, xi – Mu square, so that is = 1 by sigma to the power n root 2 pi to the power n, e to the power -1/ 2 sigma square sigma xi – Mu square.

So, the log likelihood; l Mu sigma square x that is equal to –n/ 2 log sigma square – n/2 log 2 pi – ½ sigma square sigma xi – Mu square. The reason for considering log likelihood in place of likelihood function is that; first of all, log is an increasing function of x, therefore the optimisation; that is a maximisation of L is same as the maximisation of the l the problem does not change.

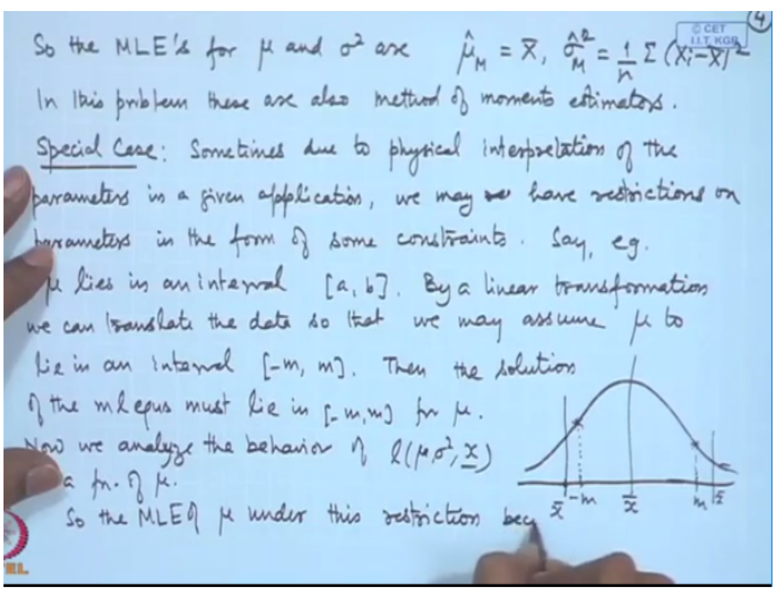
And secondly, because of the distributions nature is in the exponential family when we take the log, then the terms become simplified. So, now here we are considering a 2 parameter case, so the likelihood equation has to be differentiate; the likelihood function has to be differentiated to respect to Mu and sigma square both and we have to check a second order Sn to be a positive definite matrix for the maximisation of this.

(Refer Slide Time: 11:50)



So, we consider here; say del l/ del Mu, then that gives us sigma xi - Mu / sigma square = 0. Now, this can be easily simplified and we get Mu = x bar. Now, if we consider the derivative with respect to sigma square, then we get – n/2 sigma square + 1/2 sigma to the power 4 sigma xi – Mu is square = 0, which gives sigma square = 1/ n sigma xi – Mu is square. It can be check that these are actually the maximising choices of Mu and sigma square.

(Refer Slide Time: 12:45)



However, in this; i just skip this calculations, so now, we can see that the maximum likelihood estimators for Mu and sigma square, so for Mu, it is x bar and for sigma square, the solution consist of Mu here, so we put the solution for Mu as x bar, so we get the maximum likelihood estimators. So, the maximum likelihood estimators for Mu and sigma square are; Mu head; let me call it Mu head Ml = x bar and sigma head square m that is = 1/n sigma xi – x bar is square.

Note here that sigma head m square is not unbiased whereas, Mu head m is unbiased. In fact, in this particular problem these are also the same as the method of moments estimator. In this problem, these are also method of moments estimators. Let us consider a special case, sometimes due to physical interpretation of the parameters in a given application; we may have restrictions on parameters in the form of some constraints.

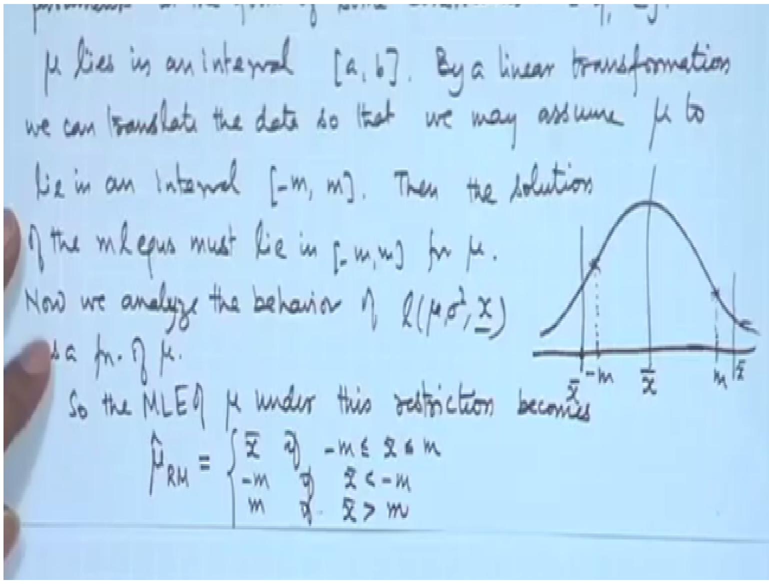
Say for example, Mu lies in an interval say, a to b. Now, by a linear transformation, we can translate the data, so that we may assume Mu to lie in an interval say, -m to m, then the solution here will have to be modified here, that is here we are considering Mu = x bar, actually the solution is coming from the derivative here. Now, if you look at this condition, this is nothing but n x bar – Mu / sigma square.

Since, sigma square is positive, you can concentrate on the numerator part. So, if we are looking at it as the function of Mu, then for Mu < x bar, this is positive that means it is increasing up to x bar, thereafter it is decreasing. So, the nature of the function and the likelihood function with respect to Mu can be considered as that it is increasing up to x bar and thereafter it is decreasing.

Now, if I make the assumption that Mu lies between –m to m, then we have 2 have x bar between this for the solution to be satisfied because in the method of maximum likelihood, we maximise the likelihood function over the given parameter is space. So, now if you put a restrictions say –m to m, then we have to see that the solution also lies in the interval –m to m. So, now if x bar lies between –m to m, we do not have to worry about it.

So, then the solution of the ml equations must lie in –m to m for Mu. Now, we analyse the behaviour of l Mu sigma square x as a function of Mu. So, we observe that it is increasing up to x bar and thereafter it is decreasing and therefore if x bar is between –m to m, we do not bother. However, suppose x bar is here, in that case naturally you can see in this; because Mu is between –m to m and x bar is < -m.

(Refer Slide Time: 18:47)

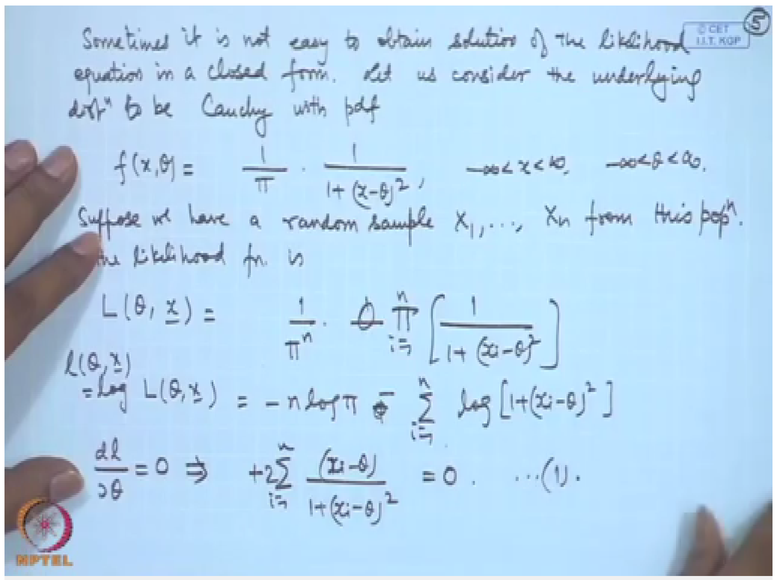


So, our value that will be considered here because x bar has become <-m therefore, we will consider –m here. Similarly, if x bar is >; suppose x bar is here, which is >, say m, then we will consider m here, so the maximum likelihood estimator will become; so the maximum likelihood estimator of Mu under this restriction becomes Mu head, let me call it restricted, so restricted estimator that is = x bar if –m is < or = x bar < or = n and it is -m, if x bar is <-m, it is = +m, if x bar is > +m.

Because we are looking at the behaviour of the function here, since it is increasing here, so the maximum value will be attained at –m, if x bar is < -m, that means it will not go beyond that thing. Similarly, on this side, if we look at; if x bar is > m, then the maximum value that will be considering will be m here because we are not going beyond this value here. Because. the Mu lies between –m to m only.

So, we are looking at the relevant portion of the likelihood function for the maximisation problem. Now, naturally, if Mu head is; Mu is modified, so sigma head square RM, then this will become 1/n sigma xi – Mu head RM square, that means it will be 1/n sigma xi – x bar square, if x bar lies between –m to m, otherwise you are going to replace by –MR to +MR as the case may be. Now, in many practical problems the solution of the likelihood equation and that means, the optimisation problem of the likelihood equation may not come so easily.

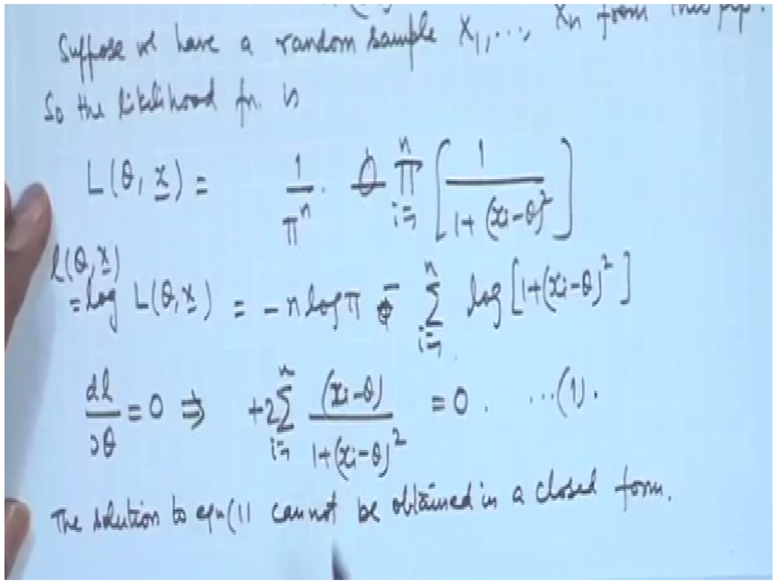
(Refer Slide Time: 20:36)



Let us take one case; sometimes it is not easy to obtain solution of the likelihood equation in a closed form. Let us consider the underlying distribution to be Cauchy, so let us consider with the probability density functions, say fx theta = 1/ pi, 1/ 1+ x – theta square. So, now let us consider here the likelihood function, suppose we have a random sample say, x1, x2, xn from this population, so the likelihood function that is = 1/ pi to the power n, 1/; product i= 1 to n, 1/ 1+xi- theta square.

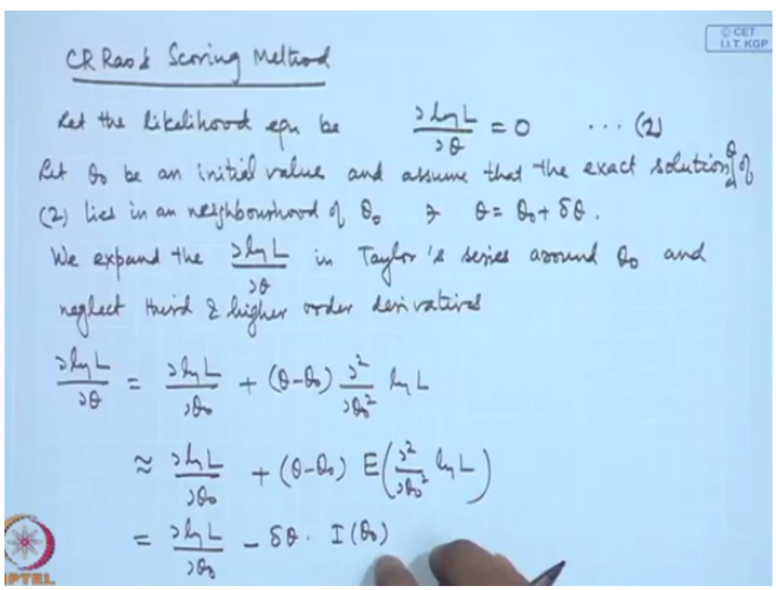
So, you consider log likelihood function, which we are calling a small l theta x that is = - n log pi + sigma i = 1 to n, log of this, which we can write it as –log of 1+ xi – theta square. So, if we consider the derivative with respect to theta, that is this is = 0, that is the likelihood equation, then we get - sigma 1/ 1 + xi – theta square and here we will get derivative of this term, that is = twice xi - theta with a – sign, so this is = 0, i= 1 to n.

(Refer Slide Time: 23:49)



Naturally, you can see that this equation is a polynomial; it is involved in rational functions here, so when you have sum i= 1 to n, then the solution of this is not cannot be obtained in a closed form, okay. So, let me call this equation 1. The solution to equation 1 cannot be obtained in a closed form and even for a moderate value of n, say n = 5 or 8 etc., the equation will be of a high order and therefore it will not be easy to solve this thing.

(Refer Slide Time: 24:25)

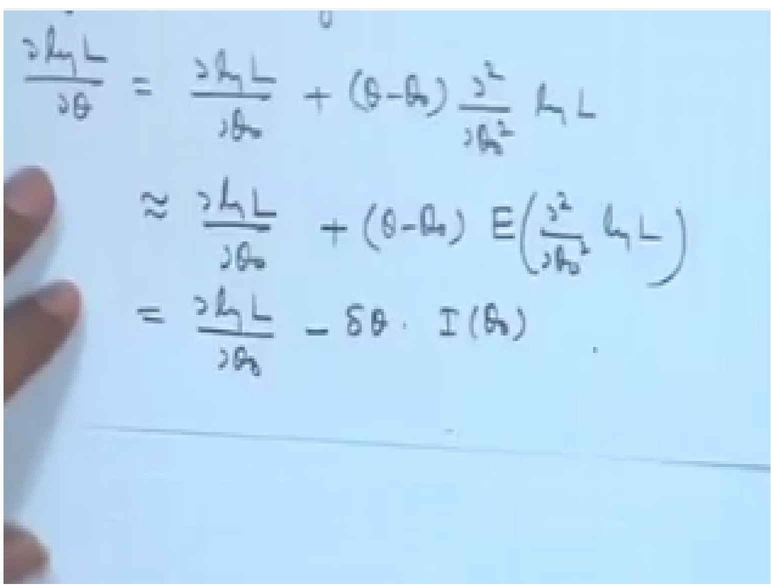


Therefore, some numerical methods are available; CR. Rao, the Indian statistician, he proposed a method which is called the method of scoring or a scoring method. In the method of the scoring, we consider; so let the likelihood equation be written as del log L/ del theta = 0, let me call it equation number 2. Let theta0 be an initial value and assume that the exact solution of 2 lies in an neighbourhood of theta that is; suppose exact solution, say theta, so we are assuming that theta = theta0 + sum delta theta.

So, we consider the term that is del log L/ del theta in Taylor series around theta0 and neglect third and higher order derivatives. So, basically what we are doing; we are writing del log L/ del theta = del log L/ del theta0 + theta – theta0 * del square/ del theta0 square log L, so we have ignored third and higher order terms so, this we can approximately write as del log L/ del theta0 + theta –theta0.

And this term we replaced by expectation, now the reason here is that if you look at this capital L term here; this capital L is actually the product of the density, so log is that sum term here, if we look at this log product here that is sigma log fxi theta, so this is the sum. Now, if we are assuming x1, x2, xn IID random variables and we are making condition on the existence of the first moment here, then by the log of large numbers, this will converge.

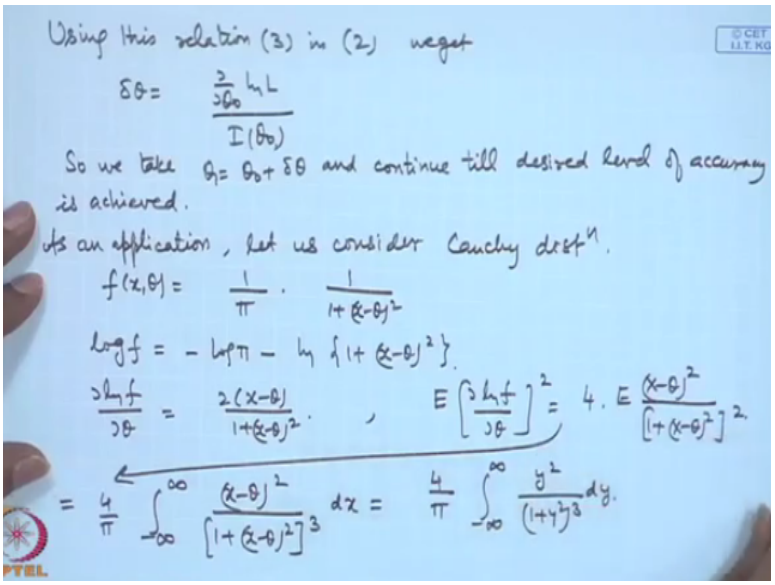
(Refer Slide Time: 28:20)



So, if it converges to expectations, so we can replace by that, so this term we replaced by expectation here and then we can determine; theta – theta0 = delta, because we have assumed, so that is –delta theta I theta0, this is the information, which I introduced a little earlier, this I theta0 here. So, now if we use this in this equation, so we are saying, using this in equation 1; using this let me call it, say 3.

Using this relation 3 in 2, so if you substitute there, we get delta theta = del/ del theta0 log L/ I theta0. So, basically if we start with an initial approximation theta0 and evaluate this, then delta theta is given by this, so now you consider theta1 = theta0 + delta theta and that will become the next approximation and then using that theta1, we can again calculate delta theta by substituting in this equation theta1 and continue till we achieve desired accuracy.

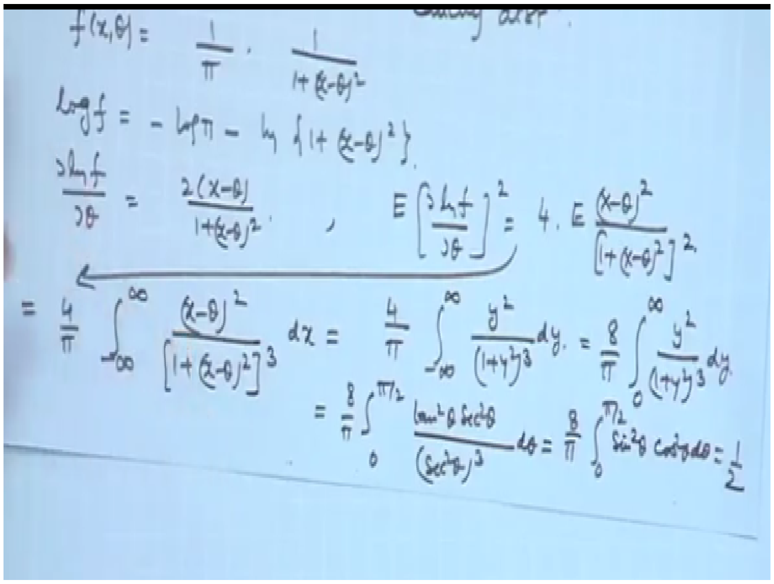
(Refer Slide Time: 29:25)



So, we take theta1 = theta0 + delta theta and continue till desired level of accuracy is achieved. So, as an example, let us consider; let us consider Cauchy distribution. In the Cauchy distribution, we just know saw, your fx theta is 1/pi, 1/ 1+ x- theta square, so we take log of f that is - log of pi - log of 1 + x - theta square, so del log f/ del theta is simply = twice x – theta/ 1+ x - theta square.

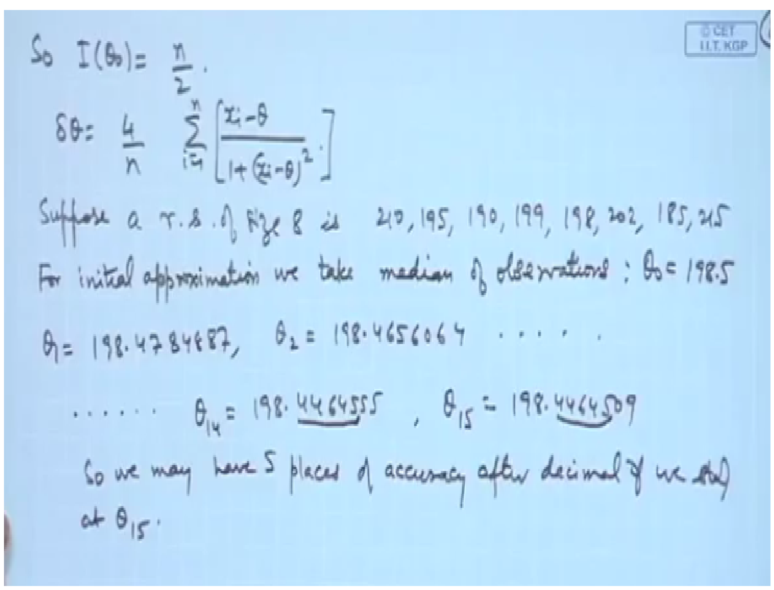
Let us consider say, expectation of del log f/ del theta whole square that is = 4 times expectation x - theta square/ 1+ x- theta square whole square. Now, for Cauchy distribution, this term can be evaluated, so this term is equal to 4/pi integral - infinity to infinity x- theta square divided by 1+x- theta square, now this will become cube. Because in the Cauchy distribution, we have another 1+ x- theta square in the denominator coming here.

(Refer Slide Time: 31:36)



So, if I substitute x- theta = y, I get 4/pi - infinity to infinity y square/ 1 + y square cube dy, so this can be easily evaluated, it is equal to 8/pi 0 to infinity y square/ 1 + y square cube dy and if we make a simple transformation likewise equal to 10 theta, then this becomes integral from 8/pi 0 to pi/2 tan square theta, sec square theta / sec square theta cube d theta that is 8/pi 0 to pi/2 sin square theta cos square theta d theta.

(Refer Slide Time: 32:27)

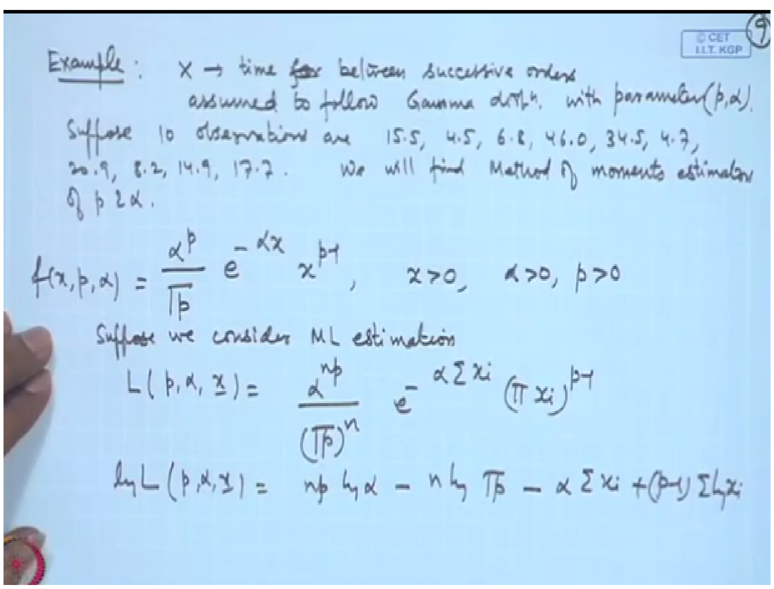


So, this can be evaluated and it turns out to be simply half, so this I theta 0; so I theta0 is simply equal to n/2, so delta theta, which we wrote as del/ del theta0 log L/ I theta0, that will be simply equal to 4/n sigma i=1 to n xi- theta/ 1+ xi- theta square, so we have obtained the formula, which can be used for the method of the scoring that means, if I consider theta0 as an initial approximation, then in substituting on the right hand side, we get the value of delta theta here.

So, that theta1 will become theta0 + delta theta, so as an application let us consider 1 problem. Suppose, a random sample of size say, 8 is 210, 195, 190, 199, 198, 202, 185 and 215, for initial approximation, we take say median of observations, so that is a theta0 = 198.5. Now, you can carry out the calculations, so theta 1 will become = 198.4784887, theta 2 = 198.4656064 and so on. If we continue like this, we get theta14 = 198.4464555, theta15 = 198.4464509.

So, we can see here up to 5 decimal places, the value is same, so we may stop here. So, we may have 5 places of accuracy after decimal, if we stop at theta 15, so this method of scoring is quite useful to obtain the solutions of the likelihood equation, if the likelihood equation cannot be solved in a closed from that means the solution cannot be obtained in an analytical form and now, I will just consider a couple of examples for application of method of moments, maximum likelihood estimator etc.

(Refer Slide Time: 35:49)



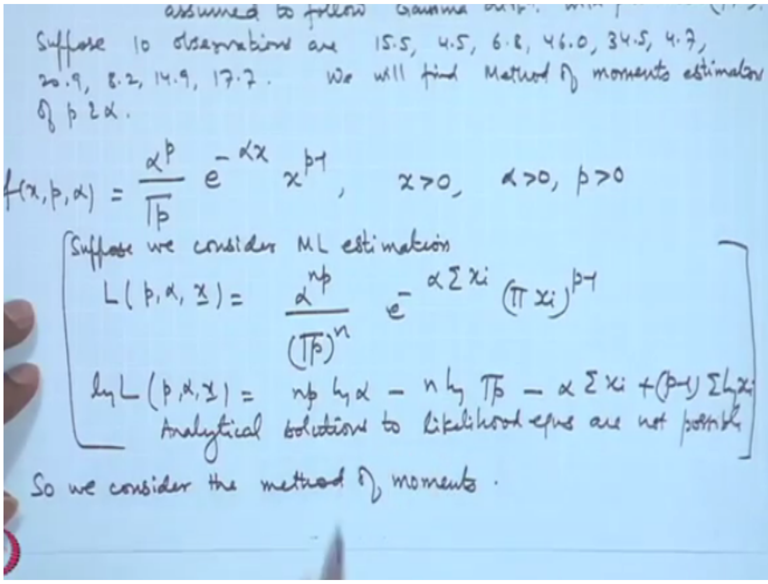
So, let us consider, say here that x is the time for; a time between successive orders, okay so it is given, it is said to follow; assumed to follow gamma distribution with parameters, say p and alpha. Suppose, 10 observations are taken to be say, 15.5, 4.5, 6.8, 46.0, 34.5, 4.7, 20.9, 8.2, 14.9, 17.7, we want the; we will find the method of moments estimators p and alpha that means we are assuming here the form of the distribution as alpha to the power p/ gamma p e to the power – alpha x, x to the power p-1 that is a form of the density function here.

And here it is assumed that both alpha and p are unknown, so the problem is of estimating both the parameters in this case. Sometimes, in a gamma distribution the parameter p is known and then we estimate only alpha, in that case maximum likelihood estimator can be easily derived but if both the parameters are unknown, then for maximum likelihood estimator becomes quite complicated.

In fact, the likelihood equation become quite complicated, as you can see here, suppose we consider; suppose we consider ML estimation okay. If we consider the ML estimation, then here the likelihood function will become L p alpha x that is = alpha to the power np/ gamma p to the power n, e to the power -alpha sigma xi product xi to the power p-1. So, if I take log here that is equal to np log alpha - n log gamma p – alpha sigma xi + p-1 log of product xi, which we can write as sigma log of xi.

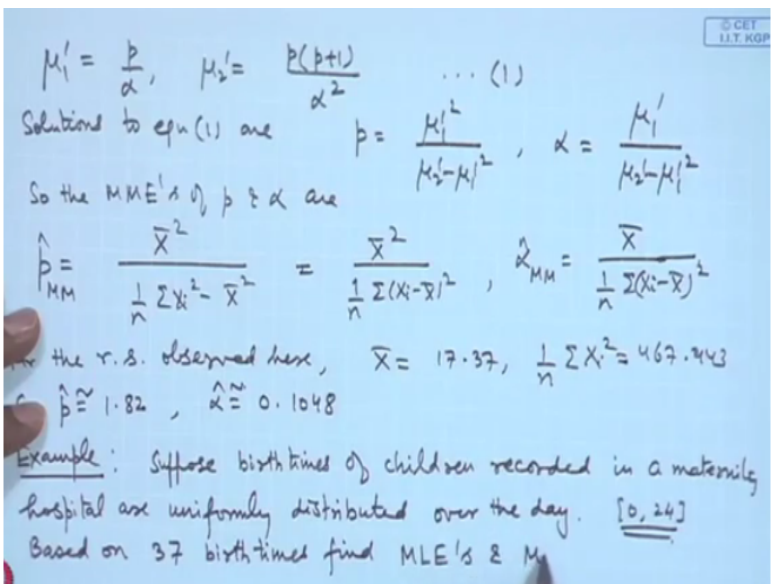
Now, easily you can see that if I want to differentiate with respect to alpha, easily we can do it but if we want to differentiate with respect to p, then there is a problem because p is occurring inside the gamma function here and therefore the solution of the likelihood equation will become complicated and we have to apply some numerical methods such as scoring method etc, to get the solutions.

(Refer Slide Time: 39:23)



So, you can see here that analytical solutions to likelihood equation are not possible, so we consider the method of moments.

(Refer Slide Time: 40:02)

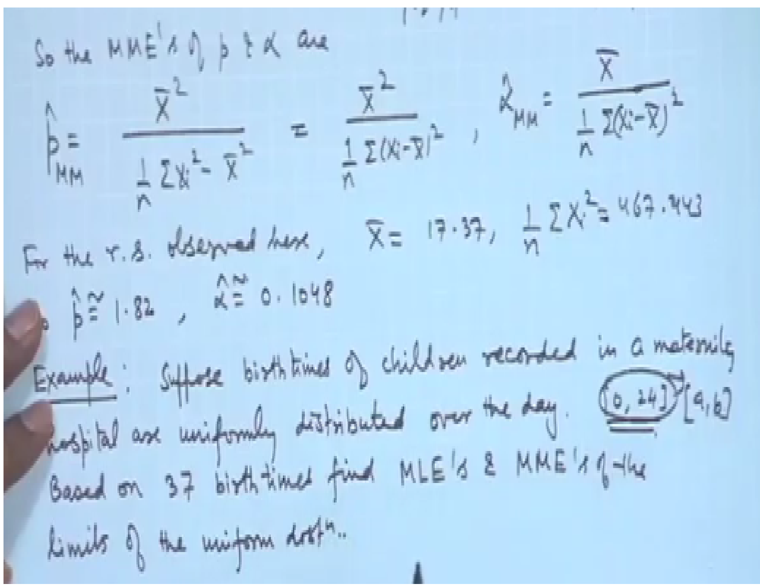


So, if we consider the method of moments here, we look at the first 2 moments about the origins, so Mu 1prime here is p/ alpha and Mu 2 prime = p*p+1/alpha square. So, now you consider the solution of this, let me call it equations 1, solutions to equation 1 are that is p = Mu 1 prime square/ Mu 2 prime - Mu 1 prime square and alpha = Mu 1 prime/ Mu 2 prime – Mu 1 prime square. If we remember, in the method of moments, we estimate Mu 1 prime/ alpha 1that is the first sample moment that is x bar.

And we estimate Mu 2 prime/ alpha 2 that is 1/n sigma xi square that is the second sample moment. So, if we substitute that; so the method of moments estimators of p and alpha are p head =; let me call it mm, x bar square/ 1/n sigma xi square – x bar square and which we can of course, write as x bar square/1/n sigma xi – x bar square and alpha head mm is then equal to x bar/ 1/n sigma xi – x bar square.

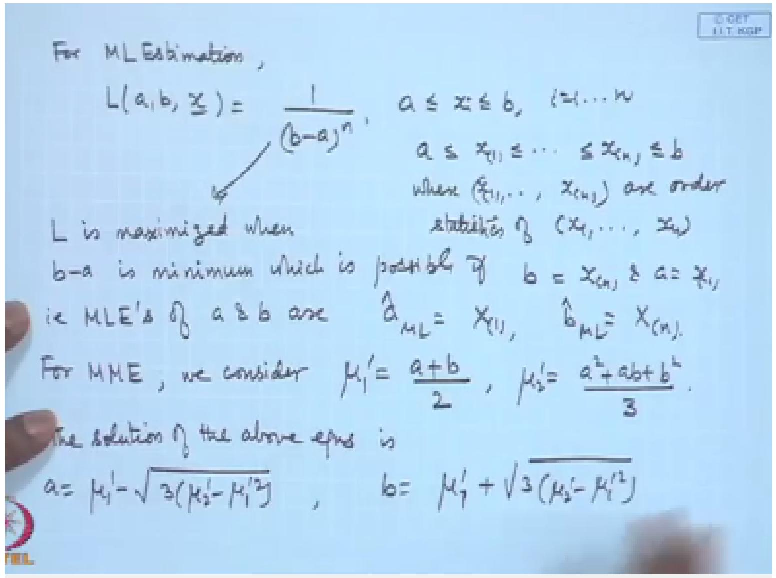
For the random sample observed, we can see that x bar = 17.37 and 1/n sigma xi square = 467.443, so p head will be equal to 1.82 approximately and alpha head = 0 .1048 approximately, so these are the method of moments estimators in this particular problem. Let us consider one application, where I can calculate both the method of moments estimators and the maximum likelihood estimators.

(Refer Slide Time: 44:03)



Suppose, birth times of children recorded in a maternity hospital are uniformly distributed over the da. So, if we say over the day we can consider say, 0 hours to 24 hours, we can consider like this and so based on 37 birth timing find maximum likelihood estimates and method of moments estimates of the limits of the uniform distribution. Since we are recording over the day is between 0 to 24 but it is some interval say, a to b here.

(Refer Slide Time: 44:33)

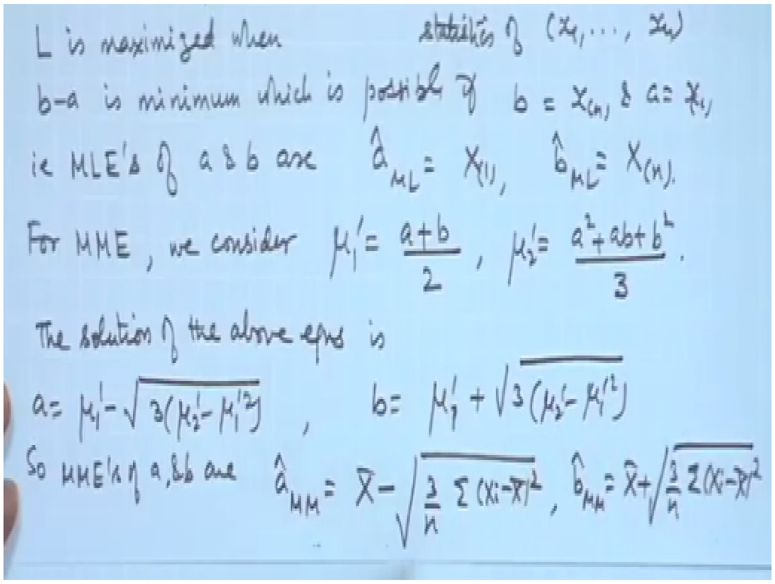


Now, we want to find out the realistic a and b here, which will be estimated from the data. Let me write down here the method of moments estimates and the maximum likelihood estimates here. So, for maximum likelihood estimation; the likelihood function is that is equal to 1/ b-a to the power n, where a is < or = xi, < or = b, for i=1 to n, each observation lies between a to b. Now, if you look at this term, if you want to maximise this it will be equivalent to minimising the value of b-a.

Now, minimising of the b-a can be done, if we can find the minimum value of b and the maximum value of a. Since, all the observations are between a to b, this restriction is realistically reducing to a < or = x1, < or = xn, < or = b, where x1, x2, xn they are denoting the order statistics of x1, x2, xn. So, if we consider the L is maximised, when b-a a is minimum, which is possible if b is chosen to be xn and and a is chosen to be x1 that is the maximum likelihood estimates of a and b are a head ML = say, x1 and b head ML = xn.

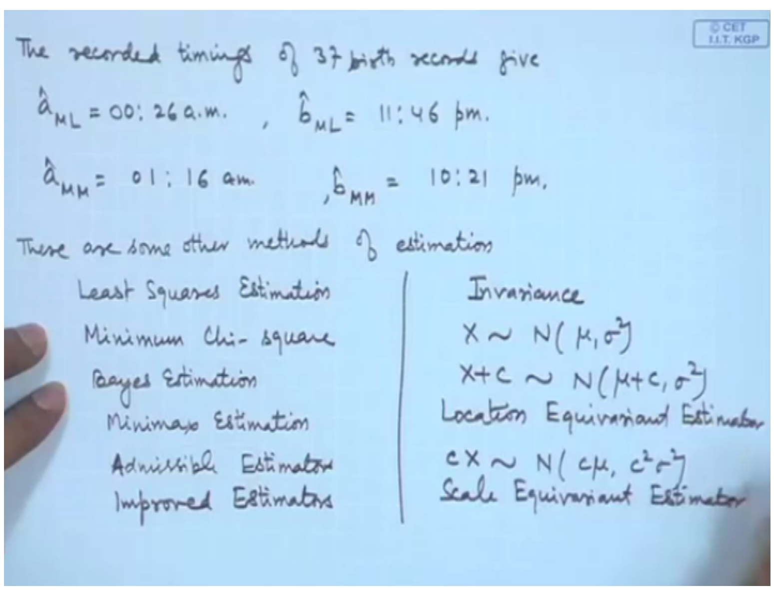
Now, let us consider the method of moments estimators in this particular problem. For method of moments, since here the parameters are a and b that is the 2 parameters are there, so we take the first 2 moments; Mu 1 prime for uniform distribution on the interval a to b, that is a+b/2 and Mu 2 prime will become equal to a square + ab+ b square/3, so we consider the solution of this. The solution of the above questions a = Mu 1 prime – square root 3 Mu 2 prime – Mu 1 prime is square, b = Mu 1 prime + square root 3 Mu 2 prime – Mu 1 prime square.

(Refer Slide Time: 47:41)



So, the method of moments estimators of a and b, they will be = a head, let me call it mm that is x bar – square root 3/n sigma xi – x bar square and b head mm = x bar + square root 3 / n sigma xi-x bar square. If we apply on the data that is available let me briefly mention the data, so the data in the terms of the timing.

(Refer Slide Time: 48:30)



The recorded timings of 37 birth records give, based on that we consider a head ML that is equal to; that is 00:26 hours that means night 12 o'clock 26 minutes and b head ML = 11:46 pm and a head mm turns out to be 0.1:16 am and b head mm = 10:21 pm, you can observe that there is some difference in the values, they are not the same. Now, the question comes that which one should be used, as I already mentioned for example, the mean squared error criteria, if we consider the mean squared errors of the estimates here, the maximum likelihood estimators would be preferred over the method of moments estimators here.

And in that case, we will prefer these as the realistic estimates of the limits of a and b in this particular problem. There are some other methods of estimation, so for example least square estimation, so I will discuss in detail the method of least squares estimation in the next module (()) (50:37), then there is a method of minimum chi square, then there are other methods, which have been developed using the concept of decision theory.

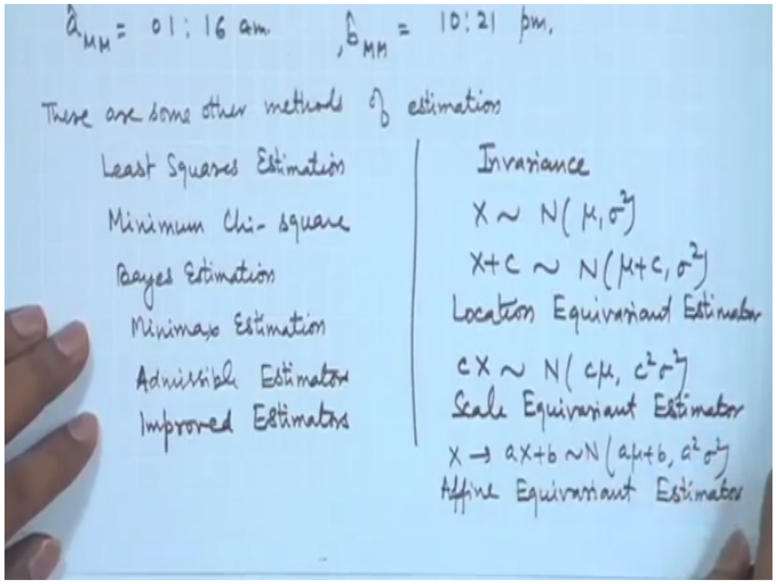
That means, we consider Bayes estimation, we have Minimax estimation and then there are some special things in the base and minimax estimation etc, so that means we put some conditions and then under those conditions, we do the base estimation, we have a base rules, we have empirical base rules, we have a limit of base rules, we have generalised base rules, we have extended base rules.

And similarly, in the Minimax T be the concept of light gamma minimax TL, minimax TN and so on. We have the concept of admissible estimators; admissible estimators and the consequently when we consider admissible estimators and then we have in admissible estimators, so therefore we consider improved estimators, one of the prominent concepts here that we are not discussed here but which is extremely useful in the decision theory that is the concept of invariance.

So, for example there are many statistical problems, which exhibit natural invariance say, I consider normal Mu square distribution, if I consider say x being shifted by; say c, then we are having observations say x+c, so now x+c will follow normal Mu + c sigma square that means the same shift is absorbed in the mean of the distribution are one of the parameters. So, if I say x follows normal Mu sigma square, then x+c will follow normal Mu + c sigma square.

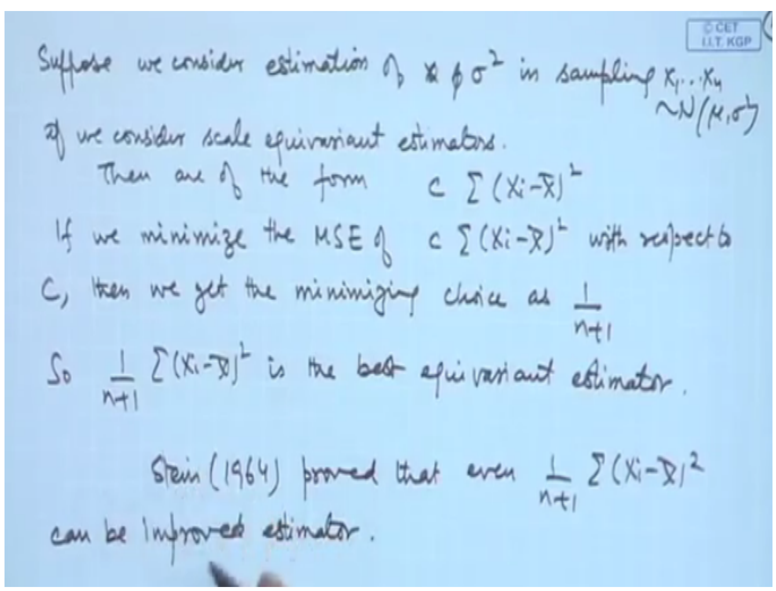
Along with this, if we impose some condition on the estimator also, that means the estimator for Mu should also shift by the same constant, then it will be called location equivariant estimators, so this is actually translation or location equivariance; location equivariant estimator and then we consider best location equivariant estimator. Similarly, we can consider say, cx then that will follow normal c Mu c square sigma square, so this is called the scale invariance.

(Refer Slide Time: 53:40)



And we consider the scale equivariant estimators are best scale equivariant estimator etc, we can consider x going to ax + b, then ax+b follows normal a Mu + b, a square sigma square, this is called affine invariance and we consider affine equivariant estimators. In many of estimation problems it has been observed that if we impose the condition of invariance, then we are able to get better estimators than the usual maximum likelihood estimators are the UMVUEs.

(Refer Slide Time: 54:25)



For example, in the case of estimation of sigma square in the normal distribution, suppose I consider; suppose we consider estimation of sigma square in sampling from; in sampling that means we are considering x1, x2, xn from normal Mu sigma square. If we consider a scale equivariant estimators, then they are of the form c times sigma xi – x bar whole square. If we minimise the mean squared error of c time sigma xi – x bar whole square with respect to c, then we get the minimising choice as 1/n+1.

So, 1/n+1 sigma xi – x bar whole square is the best equivariant estimator here, so problems of this nature and abound in practice and in fact, then if you consider some other group then even this can be improved and in 1964; 1964 Charles Stein proved that even 1/n+1 sigma xi- x bar whole square can be improved and he proposed an improved estimators, so are there are various methods of estimation, which are extremely useful in providing improved estimators.

So, those who are interested in this can refer to the books by Lehmann, J.X. Ferguson and many other texts and also the lectures on statistical inference NPTEL, so in the next part of this parametric methods, I will be starting the confidence intervals.