Statistical Inference

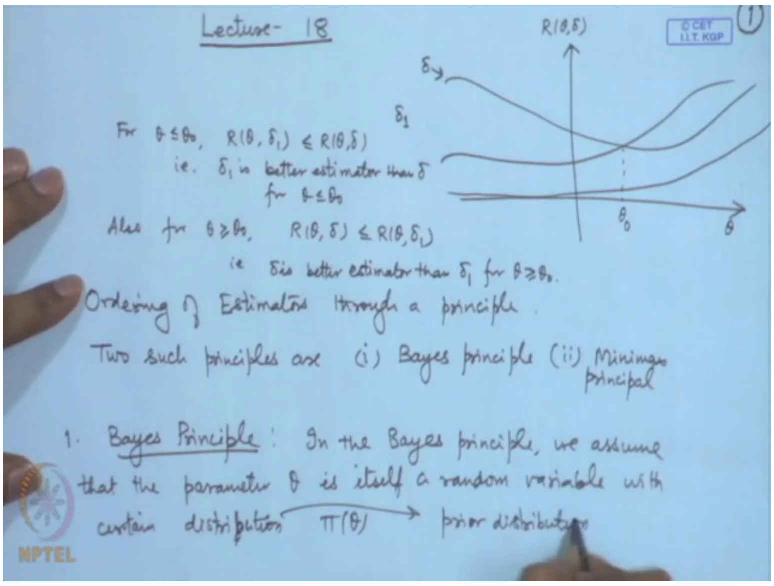
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture No. # 18

Bayes and Minimax Estimation – I

In the last lecture I introduced the concept of loss function and the risk function of an estimator. When we consider the risk function of an estimator, the criteria of choosing an estimator is based on risk optimulty, that is, the estimator, which is having smaller risk, is considered to be better.

(Refer Slide Time: 00:44)



For example, if I consider on the x-axis the parameter theta and on this side, I denote the risk functions of any given estimator delta, then R theta delta denotes the risk function, it may have certain shape.

Suppose, this is corresponding to the estimator delta, suppose corresponding to estimator, say delta 1, the risk function is like this. Suppose, this is the point, say theta naught, then we can say here, that for theta less than or equal to theta naught, risk of delta 1 is less than or equal to the risk of delta, that is, delta 1 is better estimator than delta for theta less than or equal to theta naught. Also, for theta greater than or equal to theta naught, we have R theta delta less than or equal to R theta delta 1. So, we will say, that delta is better estimator than delta 1 for theta greater than or equal to theta naught.

Now, the question arises, is there an estimator, which will have the risk function uniformly lowest? The answer to this question is no. As I explained in the last lecture through the example of a squared error loss function, that means, the class of all the estimators cannot be completely ordered in general situations. Of course, if we consider the situation where only one estimator is there, or if only the parameter space contains one point, then this is a trivial situation.

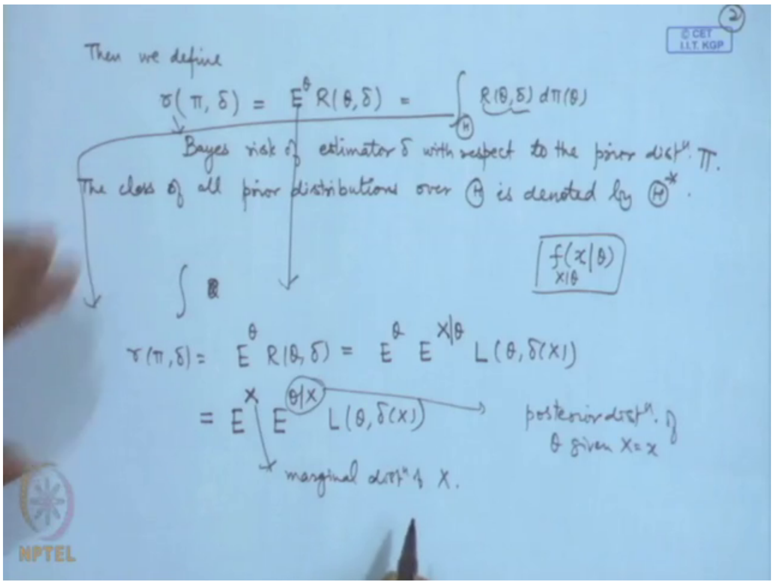
So, in the previous lecture I introduced, that one method of overcoming this problem is to restrict the class of available decision rules, like we have used the criteria of unbiasedness. So, we consider unbiased estimators and in the class of unbiased estimators, find out the best choice, which we call minimum variance unbiased estimator. Another criteria was, that we can introduce invariance concept in the problem, that means, consider equivariant estimators.

And in the class of equivariant estimators, if possible, find out the best equivariant estimator. This is one method of optimizing or finding out the optimal estimators. There is another method or another approach to look at this problem. We can introduce another ordering, we should find out a way how to order the class of decision rules or the estimators. So, we can say, that we introduce ordering of estimators through a principle. The main problem for the ordering is that R theta delta is a function and therefore, we are not able to find out the minimum. Somehow, if we can reduce this R theta delta for any given estimator to a single quantity, to a single number, then we can consider because the set of real numbers is ordered, we can find out the best choice.

So in that case there are two important principles. Two such principles are, one is called the Bayes principle and the 2nd one is called minimax principle. Let us consider the Bayes principle. In the Bayes principle we assume, that in the Bayes principle we assume, that the parameter theta is itself a random variable with certain distribution call, say pi theta. So, this distribution, that I am calling, it is called prior distribution. We can interpret it in this way, for example, we are considering estimation of the mean mu of a normal distribution.

Now from the past knowledge we have information, that the mean of a normal distribution itself has been following a normal distribution, say, with mean 1 and variance 2. In that case the prior distribution for mu is normal 1, 2. Similarly, suppose we are considering Poisson distribution with the parameter lambda and we may have a prior information on the lambda. By looking at the behavior over the past data, that lambda itself follows, say, exponential distribution with certain parameter, say a or 1 or omega, etcetera. Where that is a known quantity this is called a prior distribution.

(Refer Slide Time: 07:05)



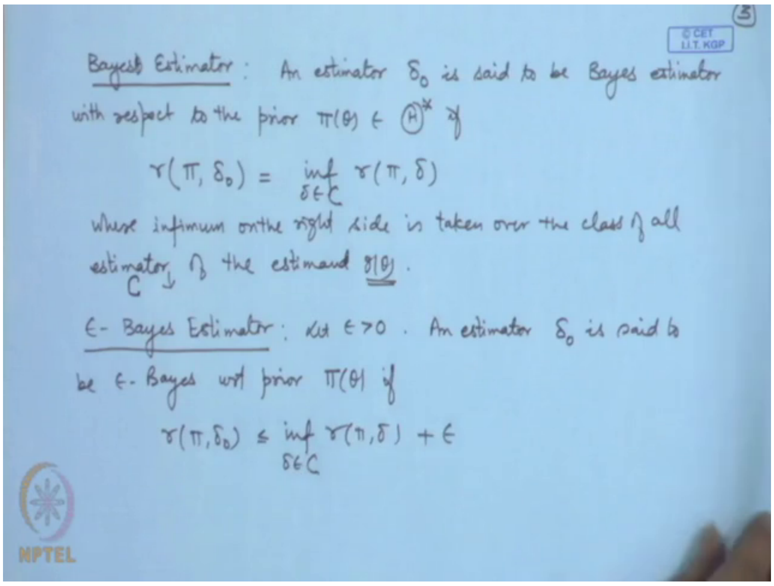
Then, what we consider here? Then, we define r pi delta as Expectation of R theta delta, where… Now, what do you mean by this expectation? This expectation is over because theta is considered as a random variable, now so suppose you are dealing with the continuous distribution, then it could be d pi theta. So, here, I have just used a general notation here as theta varies over theta d pi; theta means (( )) integral, so this could be integral or summation depending upon whether we are assuming a discrete or continuous distribution. This is called Bayes risk of estimator delta with respect to the prior distribution pi. So, the class of all prior distributions over theta is denoted by, say, theta star. So, we can talk about the Bayes risk provided this exists.

Now, we should also be able to talk about certain other quantities. For example, we may be able to talk about the, see if you expand this quantity, this could be R theta delta. Now, this R theta delta is nothing, but so this quantity for example, R pi delta, that is equal to expectation with respect to theta of R theta delta. And this R theta delta itself is an expectation with respect to x of L theta delta x. Now, the distribution of x involves theta, so we are treating it as a conditional distribution of x given theta now. Now, this is a new interpretation. Earlier, when we are considering the distribution of x, then the parameter of x, the parameter theta is considered to be a fixed, but unknown quantity.

But in the Bayesian principle since theta itself is now considered as a random variable, therefore the distribution of theta, distribution of x. But we usually use a notation f x theta. Now, we can actually use this notation in this particular fashion f x given theta to take care of the fact, that theta itself is a random variable. So, this is now considered as a conditional distribution of x given theta.

Now, as you all know, this can be written in a reverse way also, that is, firstly, we consider the conditional distribution of theta given x and then we take in integral with respect to our expectation, with respect to marginal distribution of x. So, this is called posterior. So, we should be able to talk about the posterior distribution of theta given x and this is called the marginal distribution of x.

(Refer Slide Time: 10:59)



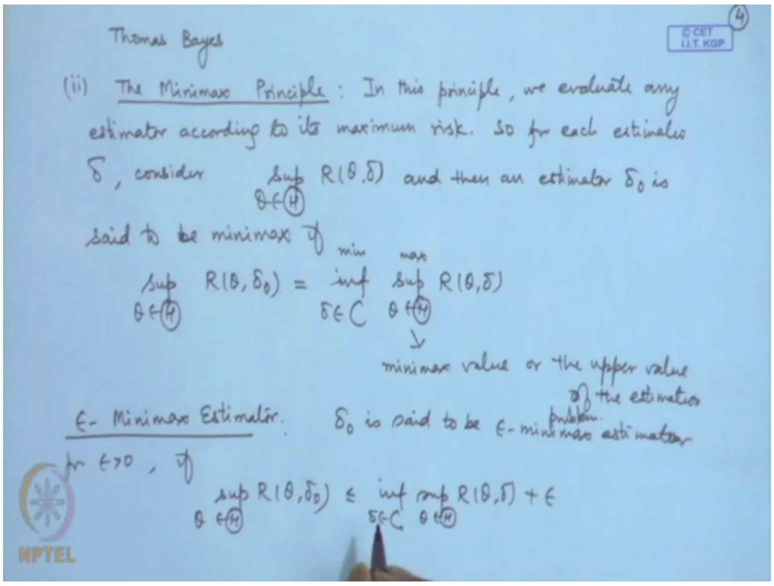
If we can talk about these quantities, then we can define a Bayes estimator as, so what is a Bayes estimator? An estimator, say delta naught is said to be Bayes estimator with respect to the prior pi theta belonging to theta star if the Bayes risk of delta naught is the minimum Bayes risk of all the estimators, where, where infimum on the right side is taken over the class of all estimators of the estimand; whatever be the parameter, that we are considering to be estimated.

Now, in this one we are assuming, that the infimum exists. Now, sometimes the infimum may not exist, in that case we may have to get ourselves happy with being close to the minimum. And we introduce the concept of epsilon Bayes estimator. Let epsilon be greater than 0, then an estimator delta naught is said to be epsilon Bayes with respect to prior pi, if r pi delta naught is less than or equal to infimum, r pi delta plus epsilon.

Now, this class of all the estimators, let us use some notation, so let us use the notation, say capital C. So, we can write here delta belonging to C and here also we can say delta belonging to C class of all the estimators.

Now, let us give a historical introduction to, that is, what is the history of Bayes estimators or Bayesian procedures.

(Refer Slide Time: 14:30)



Now, certainly, the concept of Bayes estimators or Bayesian procedures, follows the name of Thomas Bayes, the name of the statistician against whom we know the famous bayes theorem. Now, in the Bayes theorem what did we do? We had the prior probabilities of certain events, using that we were able to calculate the posterior probabilities of certain events. Now, in the case of estimation what we are doing? We are replacing the probabilities by a prior probability distribution and the posterior probabilities by a posterior probability distribution. Therefore, this name Bayes estimators or Bayesian procedures is given.

Now, when the theory of statistics in 1920s to 1940s was being developed, that time the Bayesian procedures were not very popular. In fact, the founders like Jerzy Neyman and R A Fisher, etcetera, they did not agree to the use of Bayesian procedures. The main reason was that they said how can you be sure of the prior information because the parameter to be estimated or on which you are finding out the inference is not known and therefore, it is not possible to pinpoint a prior distribution. Therefore, if the prior distribution is wrong, you get a different estimator or different procedure. So, in that case everybody will have its own estimator.

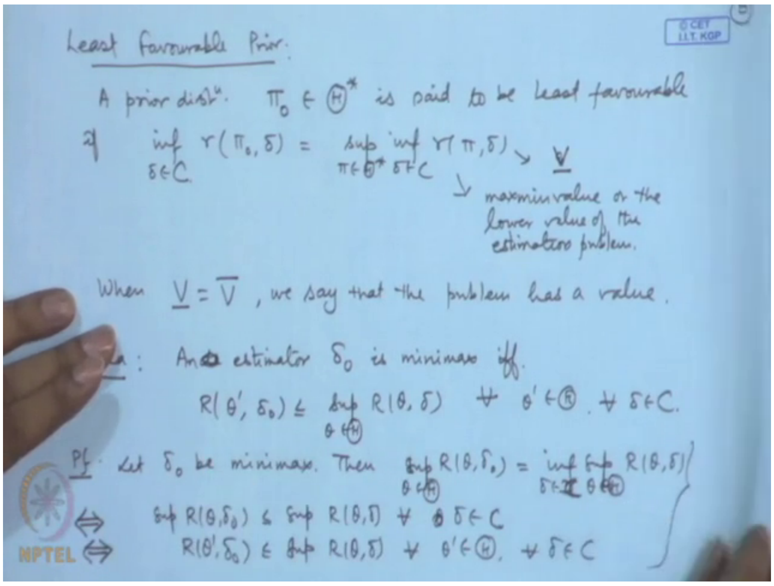
However, later on, in 1960s etcetera, by L J Savage and Bruno de Finetti, etcetera, they said, that if there is a prior information, it should be used. And then, another result was that complete class results, which were proved in the decision theory, which said, that essentially, any good rule or any admissible rule must be Bayes or limit of Bayes rule, etcetera. So, nowadays, the Bayesian procedures are much in use and especially, with the computational power, that has been developed, with the help of computational power one can actually derive the Bayesian procedures.

Now, let me also introduce the minimax principle in the Bayesian procedures or in the Bayesian principle. We reduced the risk function R theta delta to R pi delta, a single number and therefore, it was possible to find out the minimum. In the minimax principle we consider any or we can, you can say, that we evaluate any estimator by the maximum risk that may get. So, in this principle we evaluate any estimator according to its maximum risk. So, for each estimator delta, consider supremum of R theta delta for theta belonging to theta and then, an estimator delta naught is said to be minimax if R theta delta naught is equal to, infimum, the supremum of R theta delta naught, that is, the maximum risk of delta naught is actually the minimum among all the available estimators, where C is the class of all the estimators. This right hand side is called the minimax value or the upper value of the estimation problem.

Now, this minimaxity principle actually assumes, that the statistician or the decision maker is trying to optimize the worst, that can happen. That means, given any situation, what is the worst possibility and then we choose, that estimator for which that worst is the smallest or you can say, whatever worst could happen have the minimum of that. So, in a sense, it is a negative approach. Nevertheless, it is good in the sense, that we cannot do worse than the minimax value now. So, here we can use min and max for supremum and infimum etcetera and that is why the name minimax is there.

Once again, like in the case of Bayes estimation, one may not be able to find out the infimum. So, we can define epsilon minimax estimator delta naught is said to be epsilon minimax estimator for epsilon greater than 0 if supremum of R theta delta naught is less than or equal to infimum supremum R theta delta plus epsilon.

(Refer Slide Time: 21:10)

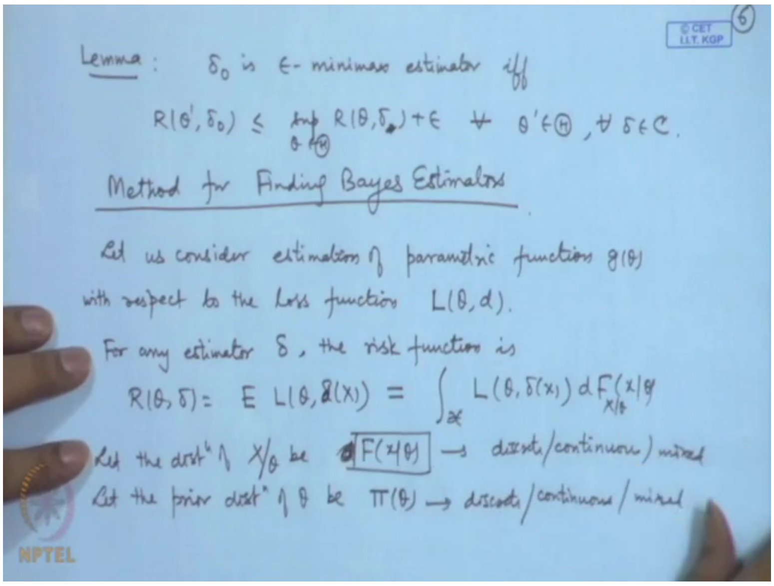


We also introduced least favorable prior, a prior distribution say pi naught is said to be least favorable if infimum of r pi naught delta is the maximum. This is called the maxmin value or the lower value of the estimation problem. We have some notations here, generally we can use V lower bar and for the upper value of the game, generally we use V upper bar as a notation. When V lower bar is equal to V upper bar, we say, that the problem has a value in the game theory. This is actually called the point of equilibrium or the problem has a saddle point, etcetera. There are certain equivalences to these definitions, I will point out one or two before giving the methods for determining the Bayes and minimax estimators.

Let me take up one or two such, let me call it a lemma here. A decision rule or an estimator delta naught is minimax, if and only if R theta prime delta naught less than or equal to supremum of R theta delta for all theta prime, and for all delta, let delta naught be minimax, then supremum of R theta delta naught is equal to infimum supremum R theta delta. Now, this implies, that supremum of R theta delta naught is less than or equal to supremum of R theta delta for all delta belonging to C. Now, this implies, that R theta prime delta naught is less than or equal to supremum of R theta delta for all theta prime belonging to theta and for all delta belonging to C.

Now, you can see here, that this implication implies this implication and this implication implies, that this implication because if I say supremum of R theta delta naught is less than or equal to supremum of R theta delta for all delta, then certainly it is equal to the infimum value. So, therefore, this is if and only if condition.

(Refer Slide Time: 25:33)



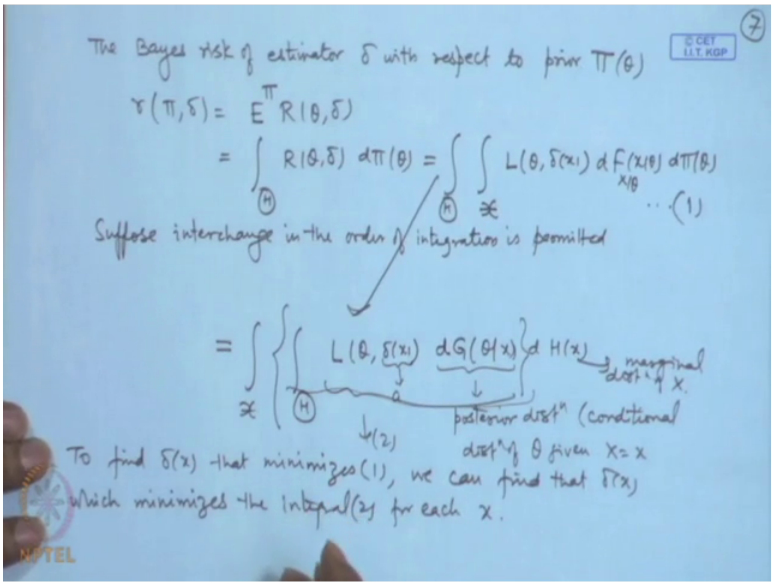
In a similar way one can consider, say, delta naught is epsilon minimax estimator if and only if R theta prime delta naught is less than or equal to supremum R theta delta naught delta plus epsilon for all theta prime and for all delta.

Now, let us see how to obtain a Bayes estimator? How to obtain minimax estimators? So, method for finding Bayes estimators, so according to the definition we must calculate r pi delta for every estimator delta and then see among all such values, that what is the one, which is minimizing, that is, we should be able to minimize r pi delta with respect to delta. So, let us look at this function. We have defined R pi delta as expectation of R theta delta, which is actually equal to expectation of theta expectation x given theta L theta delta x or…

So, if you consider this, actually this is becoming, let me use the general Lebesgue-Stieltjes integral notation, so this is becoming L theta delta x d F x x given theta into d pi theta. Therefore, it looks almost impossible how to minimize this with respect to delta. However, we try to write it in a different way. So, consider again, let us consider estimation of say parametric function, say g theta with respect to the loss function, say L theta d. So, for any estimator delta the risk function is R theta delta is equal to expectation of L theta delta x.

Now, let us consider here, that let the distribution of x be, so we may write d F x V f x theta. So, F x theta is the, it could be discrete or continuous, so we will consider the conditional distribution of x given theta and let the prior distribution of theta be, so we will give some notations, so pi theta. Once again this could be discrete or continuous or mixed and similarly, this could be discrete or continuous or mixed. In that case this is nothing, but L theta delta x d F x given theta, the integral is over x here.

(Refer Slide Time: 30:08)

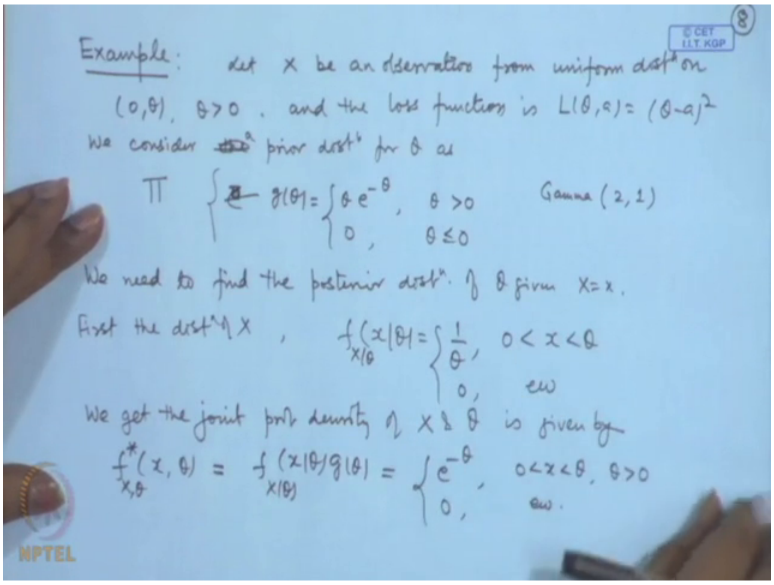


Now, the Bayes risk of estimator delta with respect to prior pi, that is, r pi delta, that is, expectation of R theta delta with respect to pi, that is equal to integral of R theta delta d pi theta over the parameter space, but this is nothing, but L theta delta x d F x given theta into d pi theta. As such the problem is to find out that value of delta for which this double integral or double summation is minimum.

What we do, suppose interchange in the order of integration is permitted, then we can write it as L theta delta x, now the notation will change here, we will consider it as d of say, G of theta given x into d of H x. Now, this is with respect to theta and this is with respect to x. What has happened? I have changed the order of integration here, so this is nothing, but denoting the posterior distribution, that is, conditional distribution of theta given x and this is denoting the marginal distribution of x.

Now, if I write like this as an iterated integral, then for each fix value of x if we look at this inside quantity, then this becomes a fix number. Therefore, we can consider it as a function of a and minimize with respect to a. Now, if for each x you are able to minimize this, then overall, so naturally this is a function of x, it will, the value, minimizing value will be dependent upon x, then that minimizing value will be called the Bayes estimator. So, to find delta x, that minimizes, let me give these numbers here 1 and this is I call 2, that minimizes. 1 we can find, that delta x, which minimizes the integral 2, only this integral for each x. So, this will give us the Bayes estimator. Let me explain through an example here.

(Refer Slide Time: 34:13)



Let us consider, say, let x be a, be an observation from uniform distribution on 0 to theta, theta is positive. So, our parameter space is zero to infinity and the loss function is, say theta minus a square, that is the squared error loss function. That means, our criteria is mean squared error.

Now, we consider the prior distribution, so a prior distribution for theta as, say pi, so we will give some notation here, let us call it, say g theta is equal to theta e to the power minus theta; for theta greater than 0 it is 0, for theta less than or equal to 0.

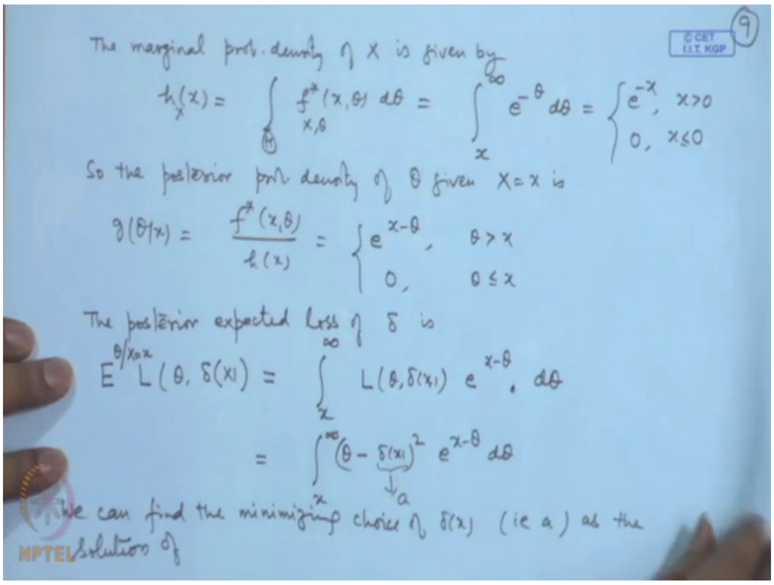
If you observe it carefully, it is actually gamma distribution, gamma distribution with parameter 2 and 1. So, we call it pi, the distribution is called pi. Now, we, in order to determine the Bayes estimator we need to calculate the conditional risk function with respect to the distribution, posterior distribution here. So, what we do now? We need to find the posterior distribution of theta given x. So, firstly, let us look at the distribution of x, the distribution of x, x follows uniform 0 theta, so you will write it as f x given theta is equal to 1 by theta 0 less than x less than theta, it is equal to 0 elsewhere.

Note here, that in the usual theory we have been considering it as a distribution of x because theta was considered to be fixed, but unknown quantity. But now, it is considered as a conditional distribution because theta is also a random variable. So, now, we are treating it as a conditional distribution of x given theta. Now, we have a conditional distribution and we have a distribution of theta, which we can consider as a marginal distribution of theta.

So, using this we can write the joint probability density of x and theta is given by, I will use the notation F star because F I am using for the marginal, so for the, for the conditional, so for the joint I will use f star. This is nothing, but f x given theta into g theta. So, that is equal to e to the power minus theta for 0 less than x less than theta, theta greater than 0, it is equal to 0 elsewhere.

In order to calculate the posterior distribution of theta given x, now I need the marginal distribution of x.

(Refer Slide Time: 38:24)

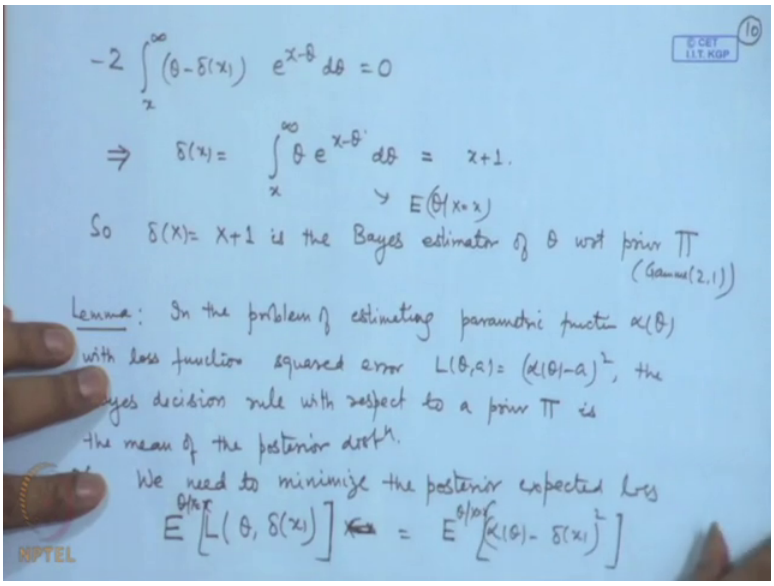


So, the marginal probability density of x is given by, let us use the notation, say h x, that is the integral of f star x theta with respect to theta. Now here, if you look at the joint distribution, the joint distribution is e to the power minus theta when theta is greater than x, at other points it is zero. So, this will become x to infinity e to the power minus theta d theta. So, this is naturally equal to e to the power minus x for x greater than 0 and it is 0 for x less than or equal to 0. That means, the marginal distribution of x is nothing, but the exponential distribution with parameter, scale parameter 1 and the location parameter 0. So, the posterior probability density of theta given x, that is, f star x theta divided by h x, that is equal to e to the power, e to the power x minus theta for theta greater than x, it is equal to 0 if theta is less than or equal to x. This is nothing, but exponential distribution with location parameter x.

So, now what is our aim? Our aim is to find out the value of delta, which will minimize this. This I call posterior expected loss of estimator delta with respect to prior pi. So, we calculate this, the posterior expected loss of delta is expectation of L theta delta x, this is considered with respect to the conditional distribution of theta given x, that is equal to L theta delta x. Since here x is fix, I can use this small x here e to the power x minus theta d theta from x to infinity, that is equal to integral theta minus delta x square e to the power x minus theta d theta x to infinity.

Now, we can substitute, say a here we can minimize. Now, if you observe this, this is nothing, but a convex function of a. Therefore, the minimizing choice is obtained if you differentiate with respect to n put equal to 0.

(Refer Slide Time: 42:35)



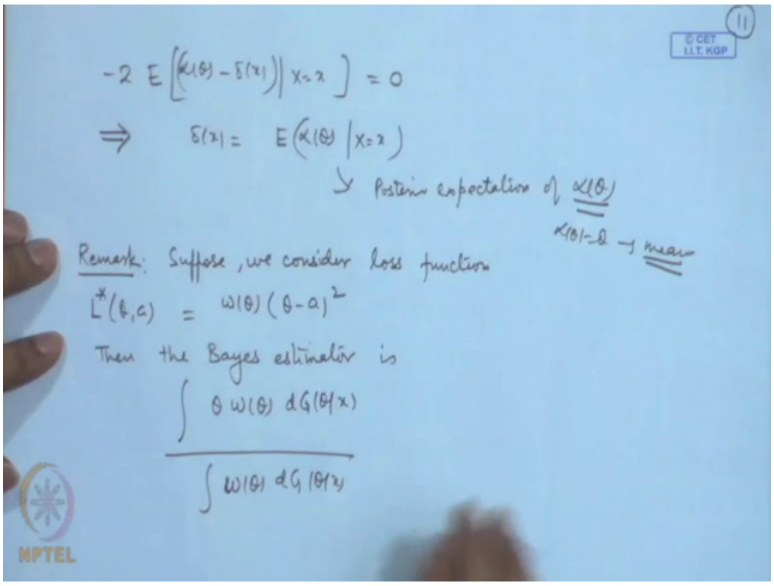
We can find the minimizing choice of delta x, that is, a as the solution of, so if you differentiate that, you get minus twice theta minus delta x e to the power x minus theta d, theta is equal to 0. This implies delta x is nothing, but integral of theta e to the power x minus theta d theta from x to infinity. So, this integral is nothing, but x plus 1. So, in fact, you can observe this as the mean of this distribution. Actually, it is expectation of theta given x, so delta x is equal to x plus 1 is the Bayes estimator of theta with respect to prior pi, that is taken as gamma 2, 1 distribution.

Now, one thing we noticed here, we had considered the squared error loss function. In the squared error loss function when we differentiate and put is equal to 0, the solution is turning out to be the mean of the posterior distribution here.

Now, in fact, this is a more general phenomena, which I will state it as a lemma here. In the problem of estimating parameter, parametric function, say alpha theta with loss function squared error, that is, L theta a is equal to alpha theta minus a whole square. The Bayes decision rule with respect to a prior pi is the mean of the posterior distribution.

Let us look at the proof of this. We can, give it in a general sense, we need to minimize the posterior expected loss, that is, expectation of L theta delta x given x equal x, so this is integral or expectation with respect to the conditional distribution of theta given x. Now, this is nothing, but expectation of alpha theta minus delta x whole square. This is with respect to the conditional distribution of theta given x. Now, in this one delta x is fixed, so we are calling it as a.

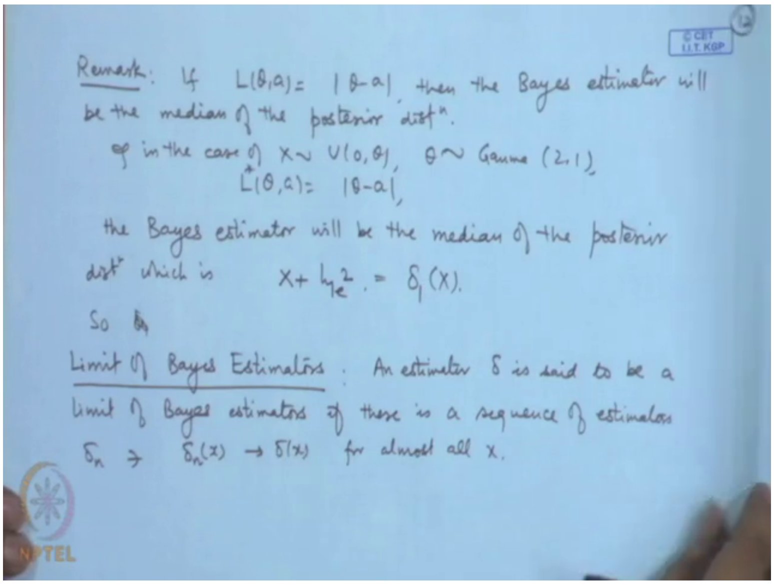
(Refer Slide Time: 46:25)



So, when you differentiate and put equal to 0, you get twice alpha theta minus delta x given x equal to x, with a minus sign, this is equal to 0, that is giving you delta x is equal to expectation of alpha theta given x equal to x, that is the posterior expectation of alpha theta. Now, in the previous problem alpha theta was equal to theta, that means, it was simply the mean.

In a similar way, we have general statements regarding weighted quadratic error loss function. For example, if we consider, say suppose, we consider loss function as, say omega theta into theta minus a square, let me call it L star, then the Bayes estimator is obtained as theta omega theta d G theta given x.

(Refer Slide Time: 48:18)

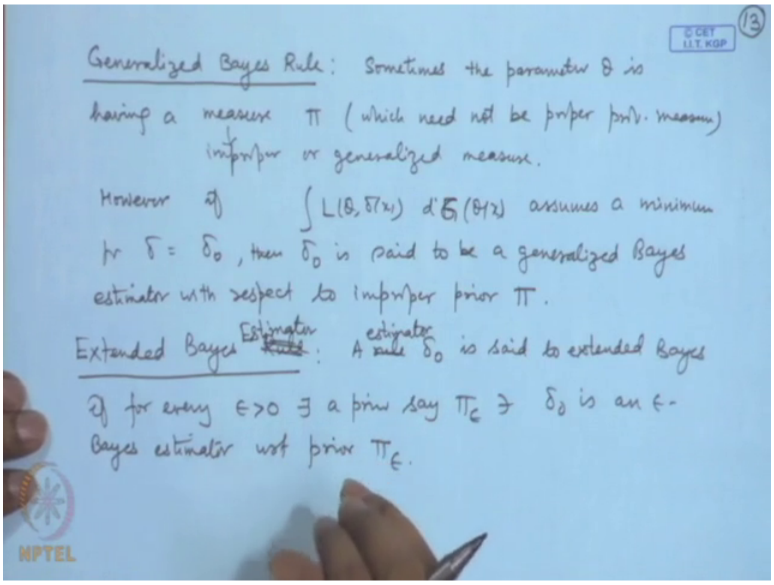


Similarly, if I have absolute error loss function, if L theta a is, say modulus of theta minus a, then the Bayes estimator will be the median of the posterior distribution. Just to give an example here, in the previous case when we have calculated the posterior distribution as exponential distribution with location parameter a x. Here, if we calculate the median, in this case the median is, for example, in the case of x following uniform 0 theta, theta following gamma distribution with parameter 2 and 1 and loss function, say modulus of theta minus a, the Bayes estimator will be the median of the posterior distribution, which is x plus log 2.

So, let me give another notation, let us call it delta x. So, delta 1 x is Bayes estimator with respect to the absolute error loss function, whereas with respect to squared error loss function, we got x plus 1. So, the change of the loss function certainly changes the form of the Bayes estimator.

We also talk about, what is known as, limit of Bayes estimators. An estimator delta is said to be a limit of Bayes estimators. If there is a sequence of estimators, say delta n such that delta n x converges to delta x for almost all x, that means, the probability, that this will converge is 1.

(Refer Slide Time: 51:51)



We defined also a generalized Bayes rule. Let us recollect the process of finding out the Bayes rules. We consider a prior distribution, which is treated as a probability distribution for the parameter theta. Using this we are able to calculate the joint distribution of x and theta, from there we calculate the marginal distribution of x, from there we derived the conditional probability distribution of theta given x.

In deriving the solution, that is, in order to find out the minimum value of the posterior expected loss, what we need is ultimately the posterior distribution, that mean, we should be able to talk about a posterior distribution. Now, this procedure of finding out the posterior distribution or you can say, the solution of minimizing the posterior expected loss has one advantage. It could be possible, that initially, whatever pi theta we are taking is not a proper probability distribution, that means, we may even be having infinite probability measure, that means, not a probability measure, it is simply a measure.

But can we talk about the joint distribution and then, we can calculate in the same way a marginal distribution, which may again not be a proper probability distribution. Ultimately, when we write down the conditional, that is, g theta given x, is it a probability distribution? If that is, so we can still talk about a Bayes rule or a Bayes estimator. Now, such a, such an estimator is called a generalized Bayes rule. So, sometimes the parameter theta is having a measure, so I am saying it is not a probability measure pi, which need not be proper probability measure.

So, we call it a improper measure or generalized measure. However, however, if L theta delta x, d F, d g theta given x assumes a minimum for delta is equal to delta naught, then delta naught is said to be a generalized Bayes estimator with respect to improper prior pi. We also define extended Bayes rule, A rule delta naught is said to be extended Bayes if for every epsilon greater than 0, there exists a prior, say pi epsilon, such that delta naught is an epsilon Bayes estimator with respect to prior pi epsilon. Let me modify this statement here, estimator, estimator.

Sometimes I am using the word rule in the framework of the general decision theory here. In the next class I will give examples of extended Bayes rules, generalized Bayes rules, limit of Bayes rules and how, in the usual estimation problems they are same or different from the say, maximum likelihood estimators or the best invariant estimators, etcetera. We will also consider the desirable properties of the Bayes estimators and we will connect it to the finding out minimax estimators.

So, in the next two lectures we will be discussing these various connections and the methods of finding out Bayes and minimax estimators.