Pattern Recognition

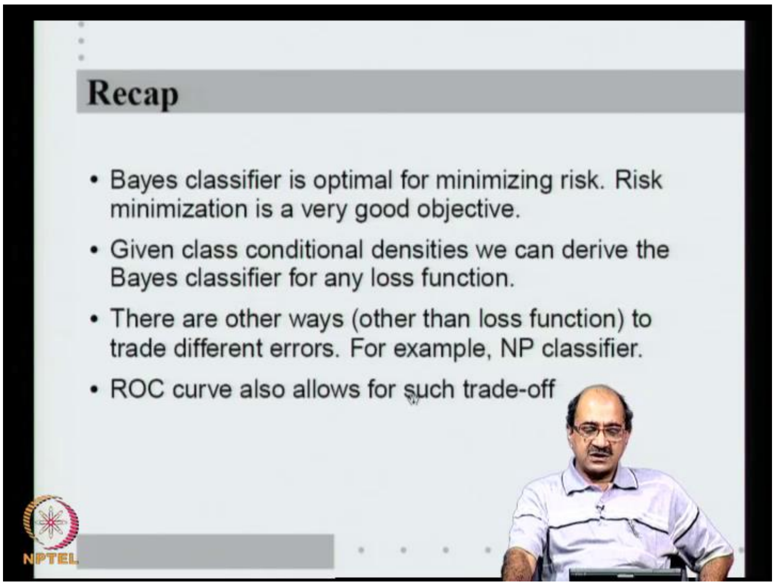
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 5
Implementing Bayes Classifier; Estimation of Class Conditional Densities

Let us get on with the next lecture, welcome to this lecture, just to recap what we have been doing so far, we were looking at the statistical way of looking a classifiers and we spent almost 2 lectures, discussing Bayes classifier and risk minimization. As we will see through the course, risk minimization is a, is the one generic technique, that is used again and again in for pattern classification and regression or functional learning.

So, Bayes classifier, as you have seen is optimal for minimizing risk, so and risk minimization is a good objective, is seen how we can get Bayes classifier for various special cases.

(Refer Slide Time: 01:08)



Recap

- Bayes classifier is optimal for minimizing risk. Risk minimization is a very good objective.
- Given class conditional densities we can derive the Bayes classifier for any loss function.
- There are other ways (other than loss function) to trade different errors. For example, NP classifier.
- ROC curve also allows for such trade-off

So, given all the class conditional densities, we can derive the Bayes classifier for any given loss function, we have derived it for different class conditional densities. And also last class, we saw a special example where the loss function is special in the sense, the actions of the classifier are not just class labels. But classifiers also allowed to reject a class pattern, that the classifier has k plus 1 options, whether there are only k classes.

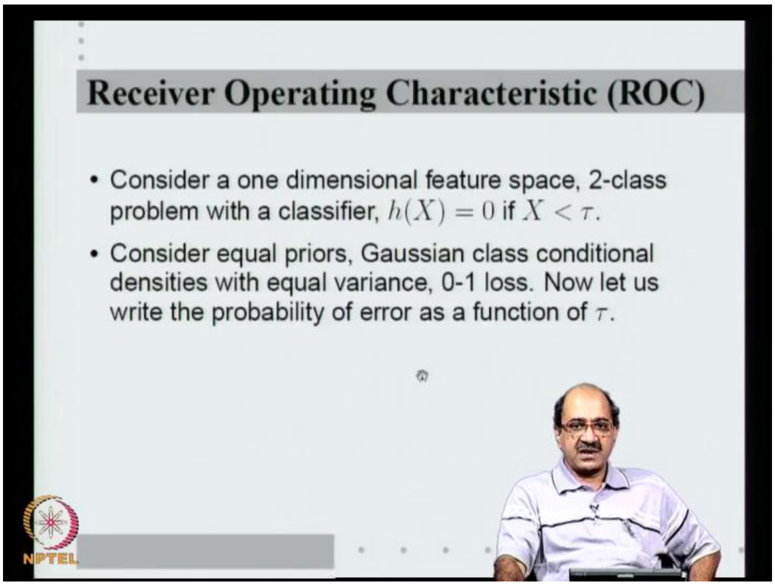
That examples should convince you that given any loss function, one can find a minimum risk classifier using the the Bayes classifier technique.

As I also said in the last class, risk minimization is only one of the many possible objectives, there are ways other than a loss function risk minimization, to think of classifiers. Essentially, one way of looking at loss functions is that, it assigns different amounts of loss to different kinds of errors say, classifier can make. So, when you take the risk, which is some expectation of loss so that, is a kind of weighted loss, the final risk through the loss function values tells you, how to trade one kind of error versus another.

So, given a loss function, which defines our acceptable trade-off risk, this minimization is one objective but there are ways other than through a loss function to trade off different kinds of errors. And one such example, we considered last class is the Neyman-Pearson classifier, where instead of saying you know, this error is so many times more costlier than that error and hence, minimize the total weighted error rate, we saying that, one kind of error should have probability below some alpha and then minimize the other kind of error.

So, there are different ways, in which I may want to trade one kind of error with another and Neyman-Pearson classifier is one good example of this trade-off. Another thing that we briefly considered last class is the, so called receiver operating characteristic curve, which is another way of explicitly affecting such a trade-off, so since we we went through ROC very fast let us, go over that again.

(Refer Slide Time: 03:23)



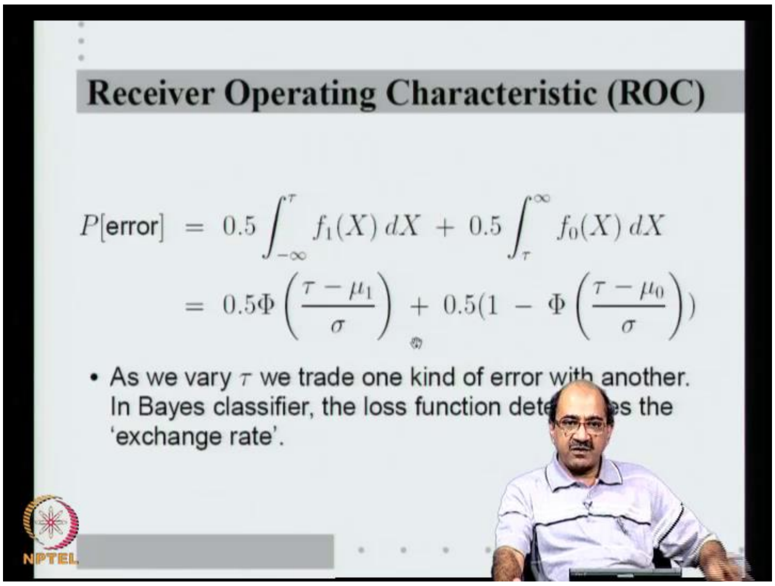
Receiver Operating Characteristic (ROC)

- Consider a one dimensional feature space, 2-class problem with a classifier, $h(X) = 0$ if $X < \tau$.
- Consider equal priors, Gaussian class conditional densities with equal variance, 0-1 loss. Now let us write the probability of error as a function of τ.

A receiver operating characteristic curve is, as I said another way to visualize trade-offs so as an example let us say, we have we have one dimensional feature space, 2 class problem, with the classifier being h(X) is class 0, if x less than tau. So, there is a single threshold and if X is less than tau, the threshold then I put in class 0 otherwise, I put in class 1.

We will consider equal priors and Gaussian class conditional densities with equal variance so we know, tau is the midway between the two means and since h(X) equals to 0 when x less than tau, we are assuming that, the mu 0 mean, for a class 0 is less than mu 1, the mean for class 1. Now, given the single threshold classifier, we can easily write the expression for probability of error, which we done in the general case, a couple of classes ago.

(Refer Slide Time: 04:12)



Receiver Operating Characteristic (ROC)

$$P[\text{error}] = 0.5\int_{-\infty}^{\tau} f_1(X)\,dX + 0.5\int_{\tau}^{\infty} f_0(X)\,dX$$

$$= 0.5\Phi\left(\frac{\tau-\mu_1}{\sigma}\right) + 0.5(1 - \Phi\left(\frac{\tau-\mu_0}{\sigma}\right))$$

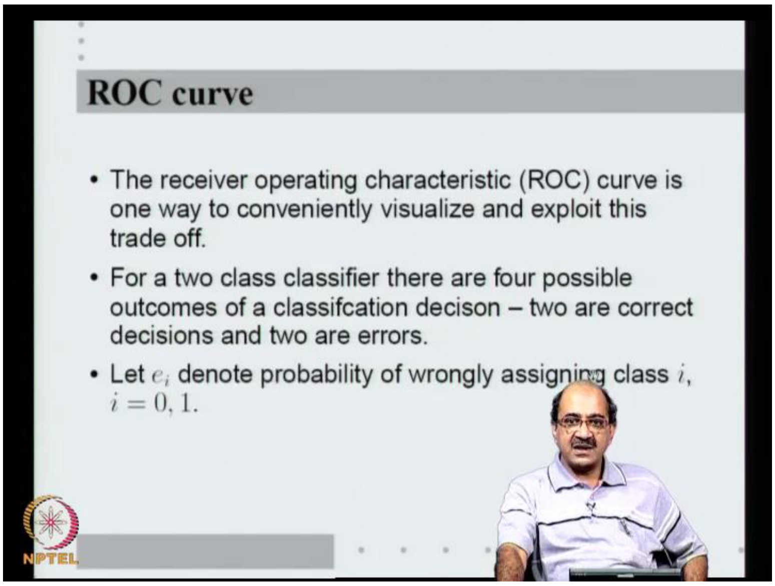
- As we vary τ we trade one kind of error with another. In Bayes classifier, the loss function determines the 'exchange rate'.

So, what is the probability of error, pairs are equal so both prior probabilities is 0.5 so what are the two kinds of error, if a pattern of class 1 will come with densities f 1 comes below tau, I would have said is class 0 say, this is one kind of error. So, this is the probability, that a feature vector of class 1 will have value less than tau that is why, minus infinity to tau f 1(X) dX. And similarly, a feature vector of class 0 comes with a value more than tau so that is why, tau to infinity f 0(X).

Since both f 1 and f 0 are Gaussian, f 1 with mean mu 1 and f 0 with mean mu 0 and both variance being sigma, this integral is nothing but phi of tau minus mu 1 by sigma. This integral nothing but 1 minus phi of tau minus mu 0 by sigma, where phi is the standard cumulative, the distribution function of standard Gaussian. When f 1 and f 0 are Gaussian, is easier to represent this integral in terms of standard Gaussian distribution function.

So, what we can see from this expression is, as I vary tau, probability of one kind of errors may increase and probability of other kind of error will decrease so essentially varying tau allows us to trade one kind of error with another. The Bayes classifier because there is one loss associated with one kind of error and another loss associated with another kind of error, fixes tau based on this weighted sum of loses.So, Bayes classifier is one way, in which I can fix the tau, and as we said tau allows you to trade one kind of error with another in that sense, we can say, the loss function defines the exchange rate between the two kinds of errors that is, one way of trading-off.

(Refer Slide Time: 06:04)

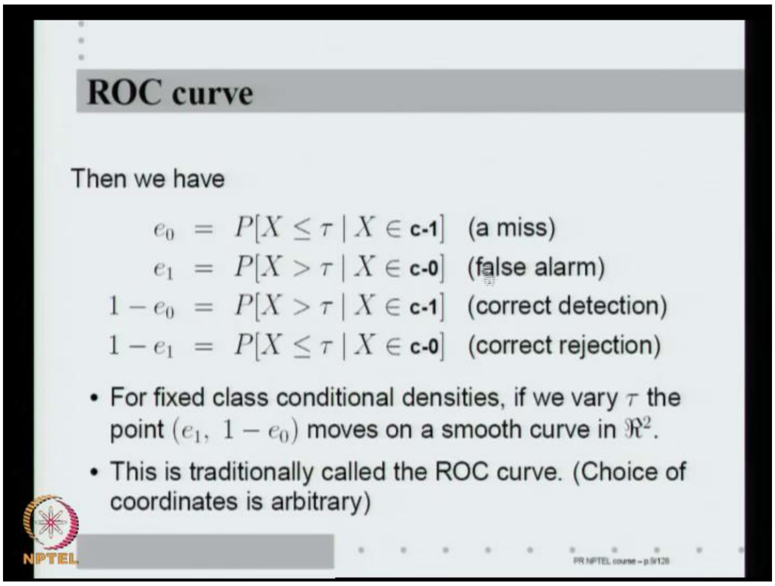


ROC curve

- The receiver operating characteristic (ROC) curve is one way to conveniently visualize and exploit this trade off.
- For a two class classifier there are four possible outcomes of a classifcation decison – two are correct decisions and two are errors.
- Let e_i denote probability of wrongly assigning class i, $i = 0, 1$.

But, actually look at in more general terms as follows, the so called receiver operating characteristic, I will I will shortly come to you as to where, this name comes from, is another way to conveniently visualize this trade-off.

(Refer Slide Time: 06:47)



ROC curve

Then we have

$$\begin{aligned} e_0 &= P[X \leq \tau \mid X \in \textbf{c-1}] && \text{(a miss)} \\ e_1 &= P[X > \tau \mid X \in \textbf{c-0}] && \text{(false alarm)} \\ 1 - e_0 &= P[X > \tau \mid X \in \textbf{c-1}] && \text{(correct detection)} \\ 1 - e_1 &= P[X \leq \tau \mid X \in \textbf{c-0}] && \text{(correct rejection)} \end{aligned}$$

- For fixed class conditional densities, if we vary τ the point $(e_1,\ 1 - e_0)$ moves on a smooth curve in $\Re^2$.
- This is traditionally called the ROC curve. (Choice of coordinates is arbitrary)

PR NPTEL course – p.9/120

When we have 2 class classifier, there are 4 possible outcomes of classification decision I can call, either 0 or 1 and the true class can be either 0 or 1, so there are 4 possibilities.

Two of them are correct decisions and two of them are wrong decisions let us say, e subscript i denotes the probability of wrongly assigning class i. What does that mean, e 0 is that, I actually call 0, class 0 but the feature vector actually belongs to class 1 so e 0 is the probability of wrongly assigning class 0 and similarly, e 1.

Now, we can write e 0 that is, I said, I say class 0 but it is actually class 1, what is the probability of that, it's probability X less than tau that is, when I will say class 0, given that X belongs to c 1. Similarly, e 1 when will I say 1, if X is greater than tau so it's probability X greater than tau, given X belongs to c 0, 1 minus e 0 and 1 minus e 1 can be written as complements of these probabilities.

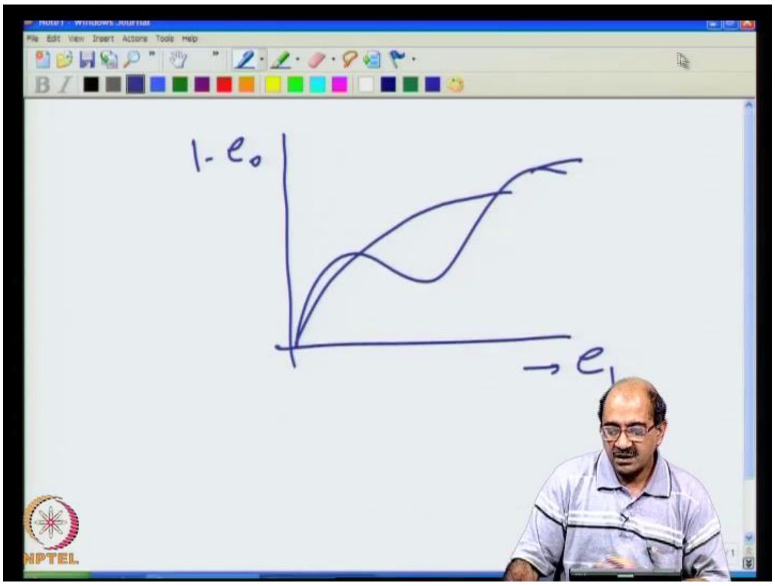
The the entire terminology comes from, as I mentioned earlier, much of this bayes decision theory was developed during second world war to make right decisions based on radar signal. The idea is looking at the radar signal, I have to call out, whether there is an enemy aircraft or not so calling 0 let us say, is that there is there is no threat, calling 1 means, there is an enemy aircraft and there is a threat. So, if I call 0 that is, I I put in class 0 whereas, it actually comes from class 1 is called a miss say, missed detection.

Similarly, if actually there is no enemy aircraft, but I call out a threat then that is a false alarm, when there is a enemy aircraft and actually, you call enemy aircraft there is a correct detection. If there is no enemy aircraft and I say there is no enemy aircraft that is a correct rejection so those are the four names. And the receiver operating characteristic name also comes, because all this decision theory is is you know, embedded into the radar receiver right.

And choosing tau is like choosing an operating point, for the receiver that is why, this is called receiver operating characteristics. Coming back, given these numbers e 0, e 1, 1 minus e 0, 1 minus e 1, if I, for any fixed class conditional densities, as we vary tau, these numbers keep changing. So, if I choose the point e 1 comma 1 minus e 0, for different tau's, I have different values of e 1 and e 0 and hence, different values of e 1 and 1 minus e 0.

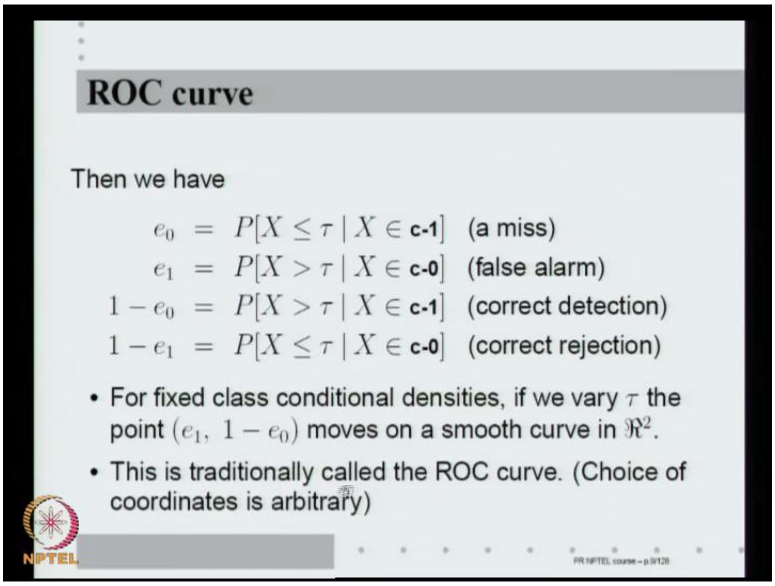
So, if I look at e 1, 1 minus e 0 space, which is R 2 and for each tau I note down, which is the point then for fixed class conditional densities, as we vary tau, the point e 1, 1 minus e 0 moves along a smooth curve in R 2. See, e 1 is a false alarm rate, 1 minus e 0 is correct detection so essentially plotting the false alarm rate on the x axis, correct detection rate on the y axis, and for different tau's, you will have different points.

(Refer Slide Time: 09:31)



So, the the curve will look something like this, so on the x axis, you have e 1, on the y axis, you have 1 minus e 0, this is the false alarm rate, this is the correct detection rate. For different taus, you get different values and it actually moves along a smooth curve of course, the curve does not always have to be like this, the curve can can have many other characteristics such curves are called receiver operating characteristic curves.

(Refer Slide Time: 10:09)



ROC curve

Then we have

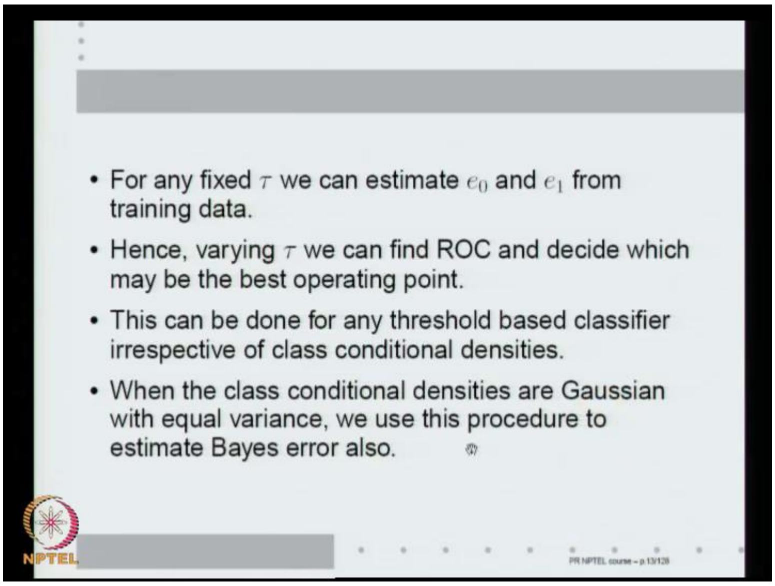
$$\begin{aligned} e_0 &= P[X \leq \tau \,|\, X \in \textbf{c-1}] \quad \text{(a miss)} \\ e_1 &= P[X > \tau \,|\, X \in \textbf{c-0}] \quad \text{(false alarm)} \\ 1 - e_0 &= P[X > \tau \,|\, X \in \textbf{c-1}] \quad \text{(correct detection)} \\ 1 - e_1 &= P[X \leq \tau \,|\, X \in \textbf{c-0}] \quad \text{(correct rejection)} \end{aligned}$$

- For fixed class conditional densities, if we vary τ the point $(e_1,\ 1 - e_0)$ moves on a smooth curve in $\Re^2$.
- This is traditionally called the ROC curve. (Choice of coordinates is arbitrary)

PR NPTEL course – p.9/128

Now, the the choice of coordinates is arbitrary but this curve when you, for various tau, you put the point e 1, 1 minus e 0 is called a receiver operating characteristic curve.

(Refer Slide Time: 10:30)



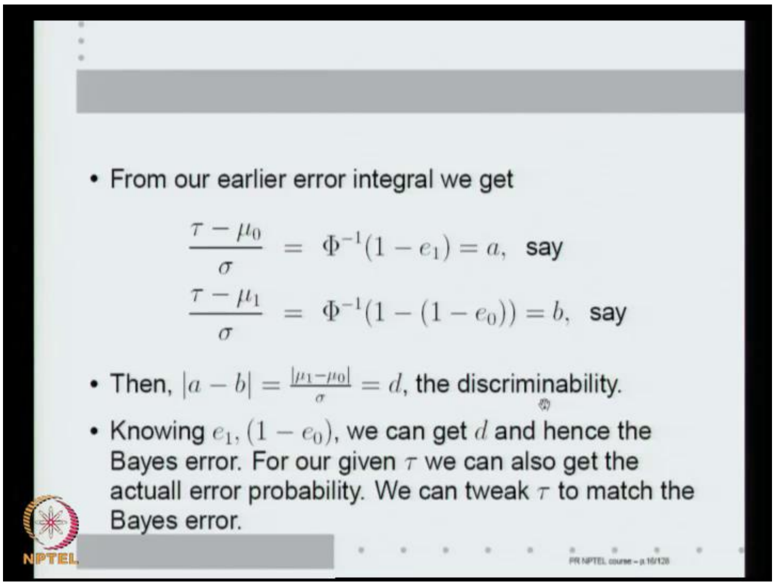
- For any fixed τ we can estimate e_0 and e_1 from training data.
- Hence, varying τ we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.
- When the class conditional densities are Gaussian with equal variance, we use this procedure to estimate Bayes error also.

PR NPTEL course – p.13/120

For any fixed tau, we can estimate e 0 and e 1 from the training data, I can fix a tau, I can calculate from the training data, how many are correctly classified, how many are wrongly classified. I can detect all the two kinds of errors and hence, I can get the fraction of errors and that is, my estimated probabilities. So, I can estimate e 0 and e 1 from the training data right then I can decide, as a varied tau, I will get different values of e 0 and e 1.

And I can decide, which tau is best for me right even, if I do not want to, even if I do not know the class conditional densities, as long as the classifier is a threshold based classifier. Simply by estimating e 0 and e 1, and plotting them for various values of the threshold, I can get the whole curve and then decide on which point on the curve, I want to be. This can be done for any threshold based classifier, irrespective of the class conditional densities. When the class conditional densities happen to be Gaussian with equal variance, this procedure is particularly helpful as follows.

(Refer Slide Time: 11:34)



- From our earlier error integral we get

$$\frac{\tau - \mu_0}{\sigma} = \Phi^{-1}(1 - e_1) = a, \text{ say}$$
$$\frac{\tau - \mu_1}{\sigma} = \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}$$

- Then, $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$, the discriminability.
- Knowing $e_1, (1 - e_0)$, we can get d and hence the Bayes error. For our given τ we can also get the actuall error probability. We can tweak τ to match the Bayes error.

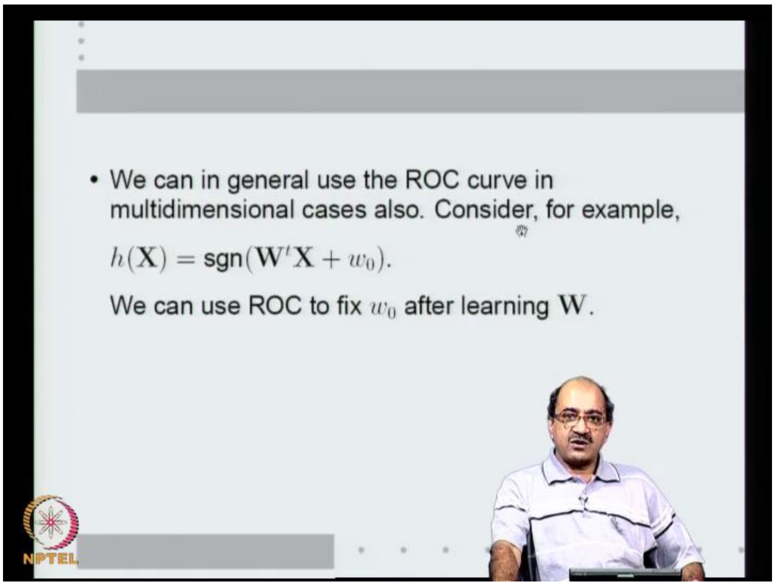
From our earlier error integrals, we know phi of tau minus mu 0 by sigma is 1 minus e 1 similarly, phi of tau minus mu 1 by sigma is e 0. So, I can write, tau minus mu 0 by sigma is phi inverse 1 minus e 1 and the other one, phi inverse 1 minus 1 minus e 0. Now, if I know the numbers e 1 and 1 minus e 0, which are the coordinates of the ROC then I can calculate phi 1 phi inverse of 1 minus e 1 and phi inverse of 1 minus 1 minus e 0.

Let us say, those numbers are a and b and the interesting thing is no matter, what tau I have used, if I can correctly estimate this a and b then the absolute difference between a minus b is mod of mu 1 minus mu 0 by sigma, which is the discriminability right. So, whether or not I know the class conditional densities exactly, whether or not I know mu 1 and mu 0, for I just take some tau, estimate e 0 and e 1.

Hence, calculate these numbers a and b using the standard Gaussian distribution function then the difference between a and b is the discriminability mu 1 minus mu 0 by sigma, which gives me very nice method of tweaking. In case of Gaussian class conditional densities, I can start with some tau, I can get my e 1 and e 0, once I have e 1 and e 0, I have e 1 and 1 minus e 0 and hence, I can calculate the discriminability d.

As we have as we have derived last class, d completely specifies the Bayes error so you know for this problem, what is the optimal Bayes error then I can ask, is the tau I am currently using achieves this error rate. If it is not, I can keep changing tau, till I achieve the Bayes error rate, no matter whatever I chose to the extent. I can estimate a and b, correctly I can estimate discriminability correctly, and hence I can estimate Bayes error rate correctly. Once I know Bayes error rate, I can keep tweaking tau, till I achieve the Bayes error rate, this is one way I can use ROC in in case, the class conditional densities are Gaussian.

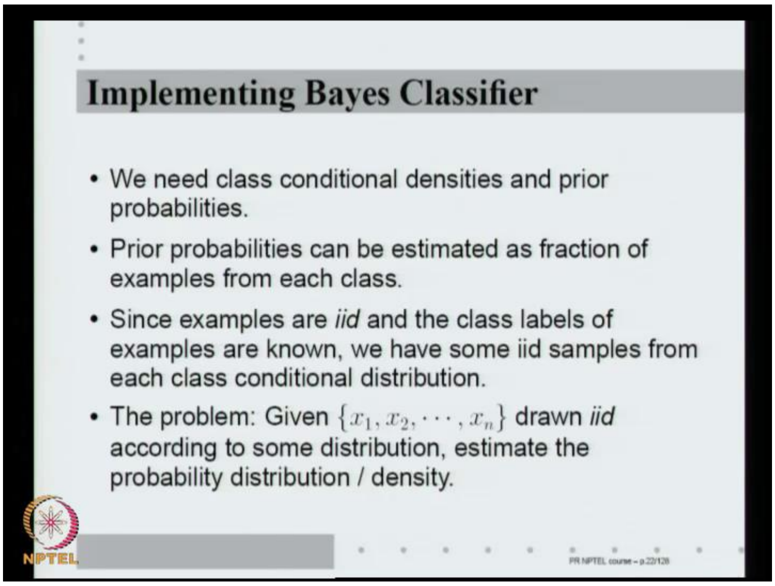
(Refer Slide Time: 13:39)



- We can in general use the ROC curve in multidimensional cases also. Consider, for example, $h(\mathbf{X}) = \text{sgn}(\mathbf{W}^t\mathbf{X} + w_0)$. We can use ROC to fix w_0 after learning $\mathbf{W}$.

Of course, you can use ROC in many other cases or suppose, I have a linear discriminant function h X is a sign of W transpose X plus w naught. If I have somehow, estimated W so I know the right direction in, onto which to project X then I can lend w 0 by using an ROC. Here, it does not matter, what class conditional this W transpose X has, just by plotting the ROC, I will be able to fix a threshold w 0. So, this this is another way, apart from Neyman-Pearson classifier to trade-off one kind of error with another right. So, that completes our general discussion of classifiers, in 2 class clear there are 2 errors loss function is one standard way, to trade these errors and risk minimization hence, is a very good objective. There are also methods such as, Neyman-Pearson classifier using the ROC curve whereby, one kind of error can be traded with another kind of error.

(Refer Slide Time: 14:35)



Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.
- Prior probabilities can be estimated as fraction of examples from each class.
- Since examples are *iid* and the class labels of examples are known, we have some iid samples from each class conditional distribution.
- The problem: Given $\{x_1, x_2, \cdots, x_n\}$ drawn *iid* according to some distribution, estimate the probability distribution / density.

PR NPTEL course – p.22/128

Now, let us move back to asking all this is fine, if I know the class conditional densities so if I want actually to implement Bayes classifier or implement Neyman-Pearson classifier, we need the class conditional densities in prior probabilities. So, let us stick to Bayes classifier so how do we implement Bayes classifier in practice, how do I get class conditional densities in prior probabilities, that is the next question.

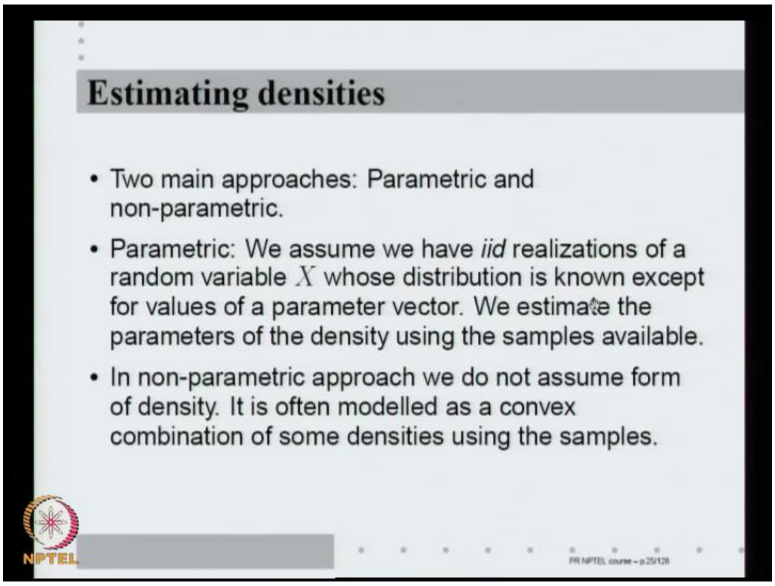
Now, prior probabilities may not be so difficult, I may know prior probabilities, I may want to assume prior probabilities of 2 classes to be same or I can simply estimate prior probabilities as the fraction of examples from each class. I have got some n examples, if n 1 of them are from class one class and n 2 of them are from the other class then n 1 by n and n 2 by n are good estimations for prior probabilities.

Then, how do I get class conditional densities, examples are i i d and class labels of examples are known, I can take the full example set and then separate them in 2 examples of class 0 and examples of class 1 right. Let us stick to to 2 classes of course, this will also work for many many more classes but so give me an example say, give me training data set where, some of the patterns will be class 0, some of the patterns will be class 1. So, I separate them out, so ultimately what I have is iid samples of class 0, iid samples of class 1.

What does that mean, if class 0, as density function f 0 I have some samples, which are drawn from a density function f 0 in a independent manner and then given to me. So, I have x 1, x 2, x n all of them are drawn from a particular density function. So, the problem now turns out to be given some x 1, x 2, x n, which are drawn in an i i d manner, according to some distribution, let us say, the class conditional density f 0, estimate the density correctly.

So, now, I do not have to look at the 2 classes together, if I can estimate the class conditional density of one class, I can estimate for the other class. So, my problem simply is, given a density function and I have I I know there is a density function and then I have got n samples drawn independently from the density function, how do I estimate the density function.

(Refer Slide Time: 16:50)



Estimating densities

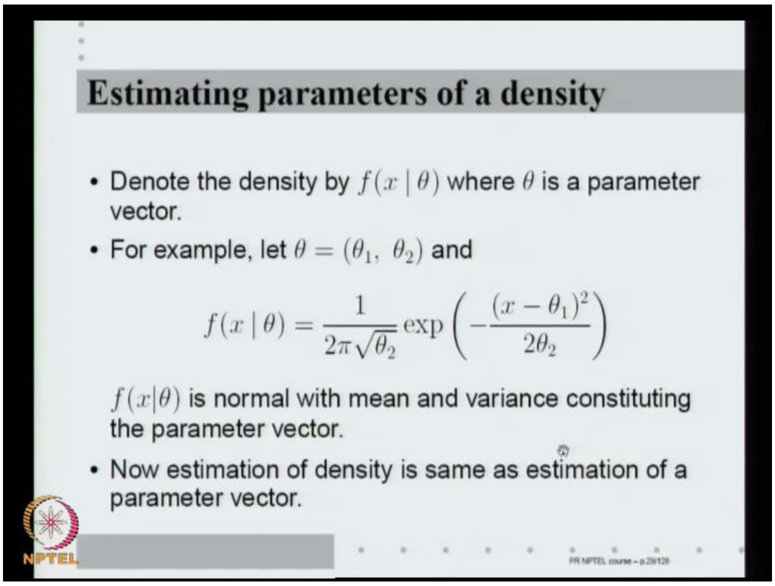
- Two main approaches: Parametric and non-parametric.
- Parametric: We assume we have *iid* realizations of a random variable X whose distribution is known except for values of a parameter vector. We estimate the parameters of the density using the samples available.
- In non-parametric approach we do not assume form of density. It is often modelled as a convex combination of some densities using the samples.

PR NPTEL course – p.25/128

So, the 2 main approaches for estimating density function from iid samples like this, these are called parametric and non-parametric approaches. What is the parametric approach, I assume that I know the density function except for some parameters that is, I know that, the class conditional density, from which I got the samples is normal. But, I do not know the mean and variance right, I may know that the class conditional density is exponential but I do not know the lambda parameter and so on.

So, in the parametric approach, we assume that the data given to us or iid realizations have a random variable X, whose distribution is known except for values of some parameters. Then we need to estimate the parameters from the density of the parameters, of the density from the samples available, this is the parametric method. In the non-parametric method, we do not assume any form for the class conditional density right, without any form for class conditional density, we want to estimate the density. Very often, it is estimated as some convex combination of densities using the sample data have, we will look at both the approaches but first we will look at the parametric approach.

(Refer Slide Time: 18:05)



Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.
- For example, let $\theta = (\theta_1,\ \theta_2)$ and

$$f(x \mid \theta) = \frac{1}{2\pi\sqrt{\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

$f(x|\theta)$ is normal with mean and variance constituting the parameter vector.
- Now estimation of density is same as estimation of a parameter vector.

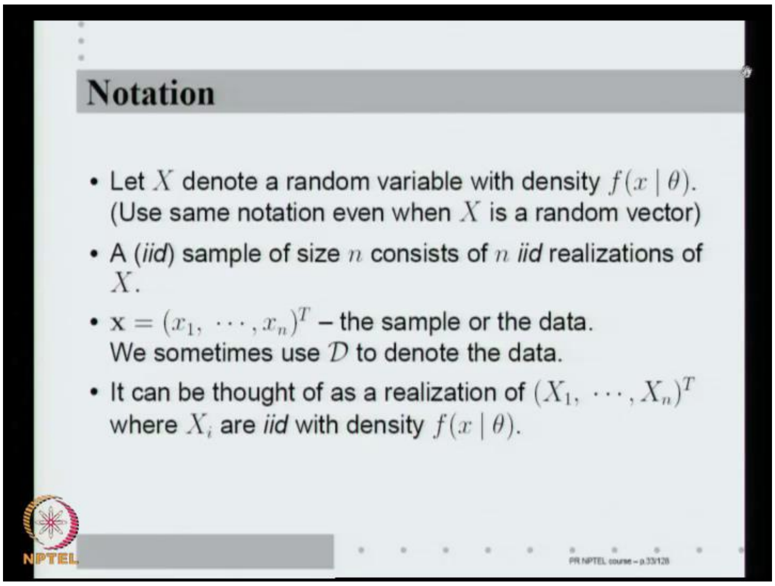
PR NPTEL course – p.29/128

In the parametric approach, let us say, we are getting data from the density f, we will write that, density as f x given theta where, theta is the parameter vector right. So, given theta does not has any mean theta as a random variable at this point but simply that, we will write f x given theta to denote that, the theta is the unknown parameter vector. So, for example, my theta could be a vector of 2 parameters theta 1 and theta 2, and the density f x given theta that is, specified in terms of theta, is 1 by 1 by I am sorry about the type, this should be root 2 pi, it is not 2 pi but it is root 2 pi.

1 by root 2 pi, root theta 2 exponential minus, x minus theta 1 whole squared by 2 theta 2 here, this is the normal density with theta 1, as the mean and theta 2, as the variance. So, f x given theta is normal with mean and variance, constituting the two parameters so this is what, we mean by specifying the parameter vector. That is the density is known except for the parameter vector, this means that the density is normal but I do not know the mean and variance. Those are given by the unknown parameters theta 1 and theta 2, once

again I am sorry this should be root 2 pi now, estimation of density is same as estimation of the parameter vector.

(Refer Slide Time: 19:44)



Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)
- A (*iid*) sample of size n consists of n *iid* realizations of X.
- $\mathbf{x} = (x_1, \cdots, x_n)^T$ – the sample or the data. We sometimes use $\mathcal{D}$ to denote the data.
- It can be thought of as a realization of $(X_1, \cdots, X_n)^T$ where X_i are *iid* with density $f(x \mid \theta)$.

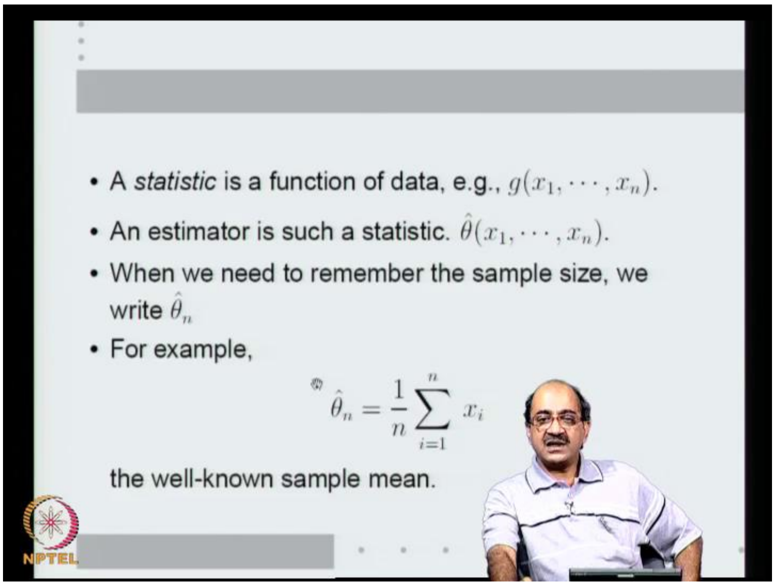
So, let us first get some notation in place to discuss about estimation, let us say, X is some random variable, which has density of x given theta that is, I know the density of x except for some parameters theta. From now on, we will not make any distinction between vector and scalar quantities so whether x is one dimensional, x is d dimensional whether it is a feature vector, whether it is a single feature, we use this same x, we do not use any boldface.

Things will become clear from context so X is a random variable, which could be a vector with density f x given theta. An iid sample of size n, consists of n iid realizations say, random variable X, you get n values, n iid values of the random variable X. So, we denote this as x 1, x 2, x n once again each of these x i's themselves may be vectors, if X is a random vector.

The entire set of data x 1 to x n, we denote by either a boldface x or a script D, this is the sample data, this is the data I have from the density, we sometimes denote it by this script D or sometimes denote it by the boldface x. When we want to think of it as a vector, we always think of it as a column vectors of x 1 to x n of course, with will be a vector, only if x i's are scalars.

If xi's then the if the random variable x itself is a vector, which is often the case because we have feature vectors then each of these x i's themselves are vectors. We can think of the data as a as one realization of the sort of random variables x 1 to x n where, each x i has density of x given theta and x i are i i d right. This sample can always so be thought of, as a realization of the the the the set of random variables x 1 to x n where, x i are iid and each of them have the same density x given theta.

(Refer Slide Time: 21:57)



- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.
- An estimator is such a statistic. $\hat{\theta}(x_1, \cdots, x_n)$.
- When we need to remember the sample size, we write $\hat{\theta}_n$
- For example,

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

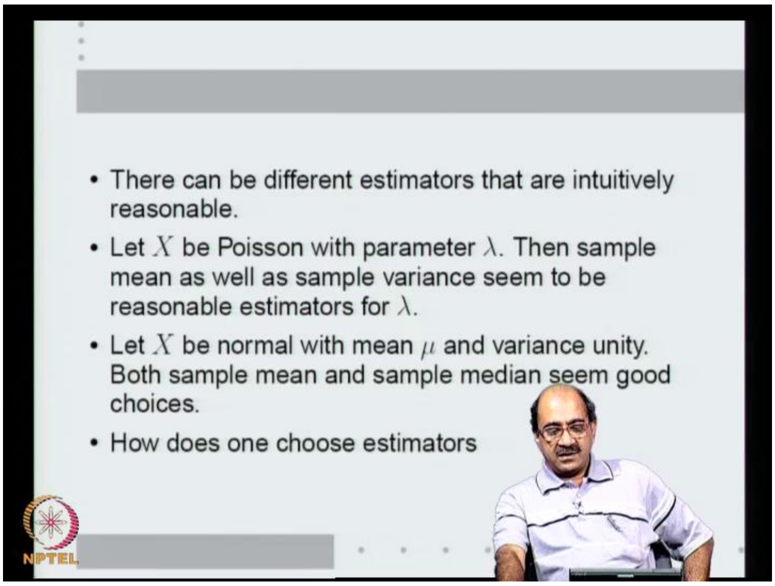
the well-known sample mean.

Now, a statistics given data is any function of data so if I am given data samples x 1 to x n, any function the data samples are x 1 to x n is called a statistics. So, essentially when I want to estimate a parameter, it is a statistic, given the data I am say, I am giving you what the parameter value is. So, the estimate essentially is a function, that maps the samples to parameter values, we generally denote estimation, such an estimate by putting a hat on the quantity.

So, theta hat is an estimate of theta and theta hat is always a function of data so we should write it as, theta hat of x 1 comma x n, when the data is cleared from context, we will simply write it as theta hat. This theta hat is obtained from n samples that is, the only thing that is really important to us so sometimes we write theta hat n to denote that theta hat is an estimate of theta, obtained from a sample of size n.

So, whenever is important to remember the sample size, we put that as a subscript of the estimate and once again, estimate is a statistic that is, a function that maps data to parameter values. So, here is an example of an estimate so an estimate obtained to n samples could be theta hat and could be defined as 1 by n, i is equal to 1 to n, x i. This is of course is the well known sample and as all of you know, this is a good estimate for the actual mean of the random variable so all estimates or functions of data like this.

(Refer Slide Time: 23:29)

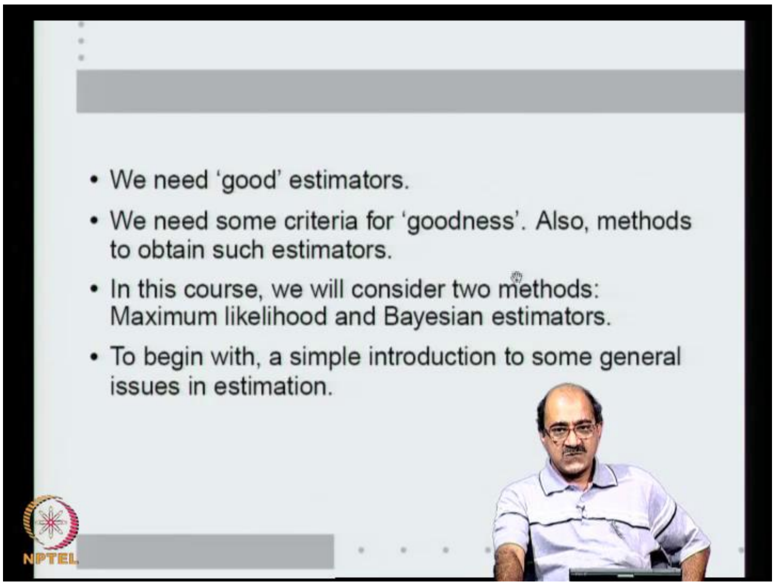


- There can be different estimators that are intuitively reasonable.
- Let X be Poisson with parameter λ. Then sample mean as well as sample variance seem to be reasonable estimators for λ.
- Let X be normal with mean μ and variance unity. Both sample mean and sample median seem good choices.
- How does one choose estimators

How could there can be different estimates, that are intuitively reasonable here are some examples suppose, X is a Poisson random variable with parameter lambda. Given the sample, the sample mean as well as sample variance seem to be reasonable estimators, for lambda for the actual Poisson random variable, both the mean and variance are lambda. So, if you give me a sample, I can take the sample mean as estimate of a lambda or I can take the sample variance as the estimate of a lambda right so both are equally reasonable.

If X is normal let us say, with some mean mu and variance unity, as you know because the normal density is symmetric, both mean and median are the normal densities mu. So, should I take the data mean or should I take the data median, both of them seem to be good choices for estimating mu. These are just some example, there are many many such questions, one can ask, so one would like some criteria to choose estimators. What should be a good way to choose estimators so let us look for some figures, I have made it for estimators.

(Refer Slide Time: 24:45)



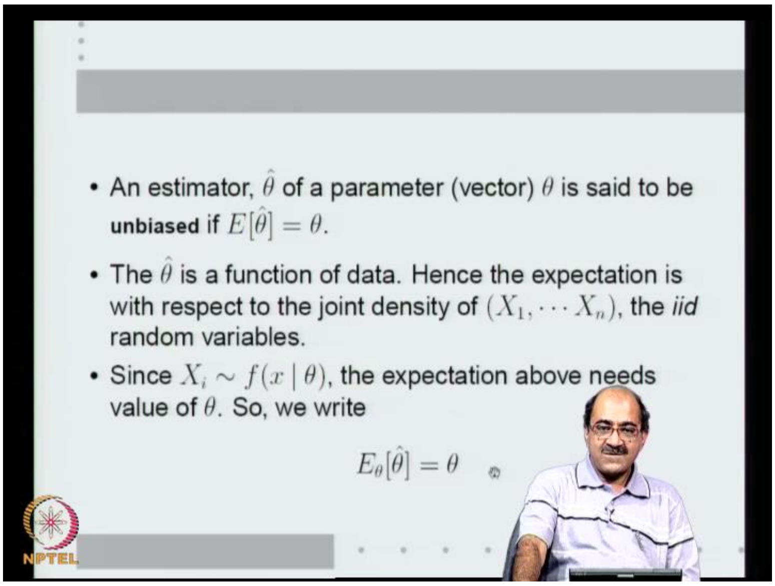
- We need 'good' estimators.
- We need some criteria for 'goodness'. Also, methods to obtain such estimators.
- In this course, we will consider two methods: Maximum likelihood and Bayesian estimators.
- To begin with, a simple introduction to some general issues in estimation.

Ultimately, we need good estimators right, what does good mean, good is based on a criteria so to decide what is good, I need some criterion right. Some criterion for goodness and of course, you know say, a criterion for goodness should somehow, allow me to obtain such estimates for various kinds of densities right otherwise, the criterion is useless.

So, let us ask for some simple criterion but anyway before you go to criterion, the methods that we use in this course, that only 2 methods that we discuss, one is called a maximum likelihood estimators other is called the Bayesian estimators. There are many other methods of obtaining estimators but these are the only two things that, we will consider in this course.

So, before we get into our methods, we will look at some general issues in the estimation so we will first discuss, what kind of properties do we want from our estimators. So that, we can decide, what are good estimators then we will ask what kind of methods will give us good estimators right, so to start with, we will just discuss general issues in estimation.

(Refer Slide Time: 25:57)



- An estimator, $\hat{\theta}$ of a parameter (vector) θ is said to be **unbiased** if $E[\hat{\theta}] = \theta$.
- The $\hat{\theta}$ is a function of data. Hence the expectation is with respect to the joint density of $(X_1, \cdots X_n)$, the *iid* random variables.
- Since $X_i \sim f(x \mid \theta)$, the expectation above needs value of θ. So, we write

$$E_\theta[\hat{\theta}] = \theta$$

An estimator theta hat of a parameter theta is said to be unbiased, if expectation of theta hat is equal to theta right, this this seems to be a nice thing to ask for. So, when I am estimating theta hat right, theta hat from data I obviously, make errors so the expectation of theta hat is equal to theta means, sometime I make errors on one set, sometime I make errors on the other set so that, all errors will cancel out right. So, at least in an expected sense, theta hat is same as the actual parameter I want to estimate because we have be little careful in understanding what this expectation means.

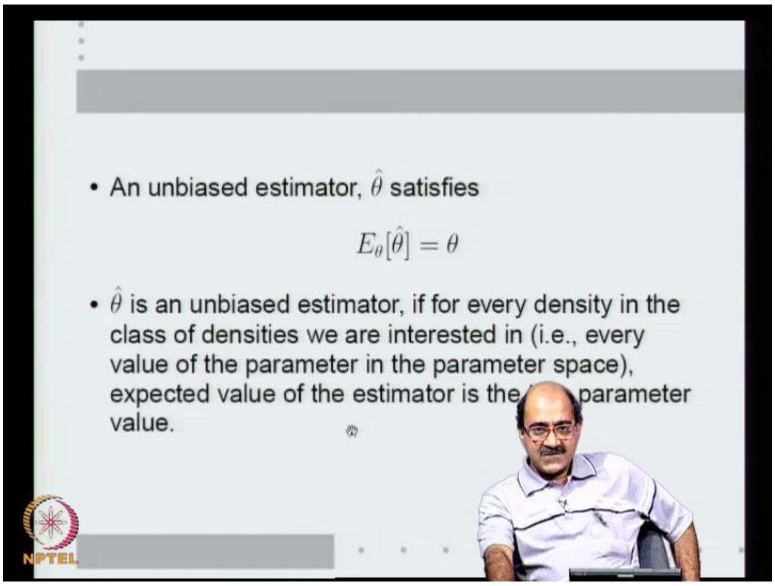
When we are saying expectation of theta hat, that is because theta hat is random that is, theta hat is a random variable, why is theta hat a random variable because theta hat is the function of data right. X 1 to X n theta hat is the function of X 1 to X n, these are iid random variables so theta hat is random because theta hat is a function of X 1 to X n. Because, theta hat is a function of X 1 to X n, when we say expectation theta hat, we mean the expectation with respect to the joint density of X 1 to X n right.

Because, the joint density of X 1 to X n is nothing but unfold product of the density of X i, which is f x given theta. We we are assuming that, the density of each X is f x given theta and X’s are independent so the joint density is simply a product of the marginals. So, this expectation here refers to expectation with respect to the joint density of X 1 to X n, which is same as the unfold product of the density model, we are using.

But then here is the catch, if X i is distributed as f x given theta then to do that expectation, we need the value of theta right. Because, theta hat is a function of X 1 to X n, I have to do that expectation with respect to joint density of X 1 to X n by the joint density of X 1 to X n, n was theta. So, what we mean by expected expectation of theta hat is equal to theta is the following, if I take any parameter value theta at the 2 parameter value and then take expectations of the random variable theta hat then that expectation should be equal to that particular parameter assumed.

So, to denote this, under this expectation we will put a subscript and say E theta, E theta is expectation with respect to joint density of X 1 to X n where, we assume the unknown parameter has actually value, this theta. Now, this equation makes sense in the following way, for any given parameter value, in the parameter space if I assume, that is the right parameter and take expectation, I should get back that parameter value.

(Refer Slide Time: 28:42)



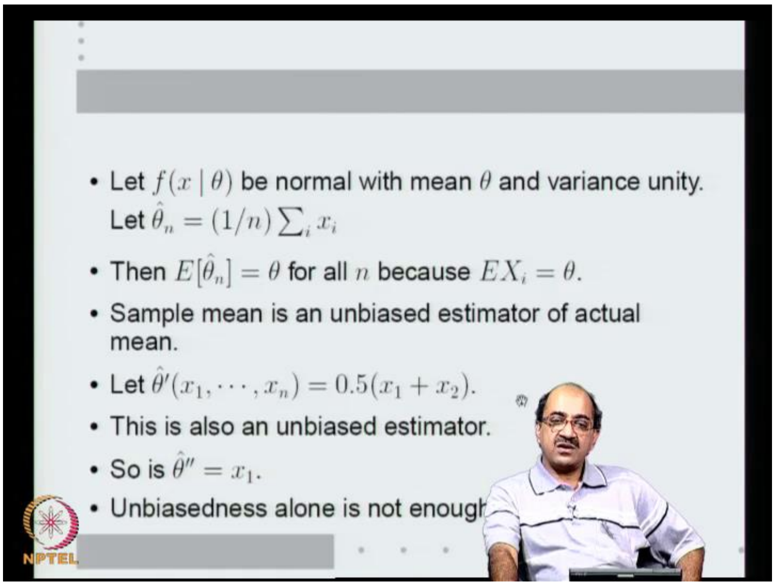
- An unbiased estimator, $\hat{\theta}$ satisfies

$$E_\theta[\hat{\theta}] = \theta$$

- $\hat{\theta}$ is an unbiased estimator, if for every density in the class of densities we are interested in (i.e., every value of the parameter in the parameter space), expected value of the estimator is the parameter value.

So, an unbiased estimator theta hat satisfies expectation of theta hat is equal to theta where, that expectation at the subscript theta, as I explained just now, what that subscript means. So, once we understand it, to give the notation, simple we will remove that subscript, we will just talk about expectation theta hat knowing this. So, a theta hat is an unbiased estimator, if for every density in the class of densities, we are interested in that, is a every value of the parameter in the parameter space, the expected value of the estimator is the true parameter value.

(Refer Slide Time: 29:15)



- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n)\sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator.
- So is $\hat{\theta}'' = x_1$.
- Unbiasedness alone is not enough

Here, are some examples suppose, f x given theta is normal with mean theta, the variance really does not matter but anyway I assume, variance as unity. And let us say, I define the estimator theta hat n that is, a estimator obtained from n samples as 1 by n summation x i. So, what is expectation theta hat, is 1 by n summation x i so if I assume that x i as distribution f x given theta with parameter theta, the expectation x i is theta so 1 by n summation expectation x i is equal to theta.

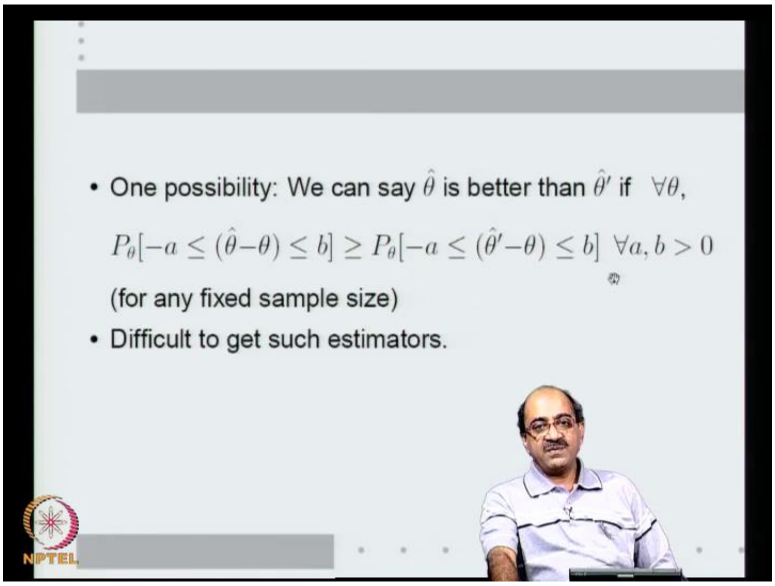
So, which means, expectation theta hat is equal to theta, for every theta and all n because expected value of x i is equal to theta right. So, the sample mean estimator, this as you can see, this is the sample mean, this is the mean of the data. So, the sample mean estimator is such that expected value of the estimator is the true value of the parameter since this seems good. But, before we can say this seems good, we we just now defined this as unbiased right, if this is satisfied that is called unbiased estimator.

So, I know that the sample mean is unbiased estimator but but this is the property that many other estimators have, sample mean is nothing special about this right. See, take the example let us say, theta hat of X 1 to X n is only X 1 plus X 2 by 2, even though I have n samples, let us say, I throw away all but n minus 2 samples and take my estimator as just the average of the first two samples.

What is the expected value of theta hat, it is 0.5 into expectation of X plus expectation of X 2, which is also equal to theta right. So, this is also an unbiased estimator suppose, I I take another estimator theta at double prime hat where, I take the estimate to be the first value I get. This is also unbiased right, it is like saying I want to calculate the probability of hats for a coin, I toss it a few times, I can take the number of hats by the number of tosses, for any number of tosses right, it is it is always unbiased.

So, basically what this means is, if I look at this theta hat and this theta theta prime hat, this theta double prime hat, all the three estimates are unbiased. So, saying an estimate is unbiased in not enough right, it does not really tell me whether estimate is good or not. So, we can go to the other extreme and say so what is that we want to say, obviously we will at our gut feeling that, this a better estimator than theta prime or theta double prime. But, on what basis, can I say this estimator 1 by n summation X i is better than this estimator or this estimator right.

(Refer Slide Time: 31:48)



- One possibility: We can say $\hat{\theta}$ is better than $\hat{\theta}'$ if $\forall\theta$,

  $P_\theta[-a \leq (\hat{\theta}-\theta) \leq b] \geq P_\theta[-a \leq (\hat{\theta}'-\theta) \leq b]$ $\forall a, b > 0$

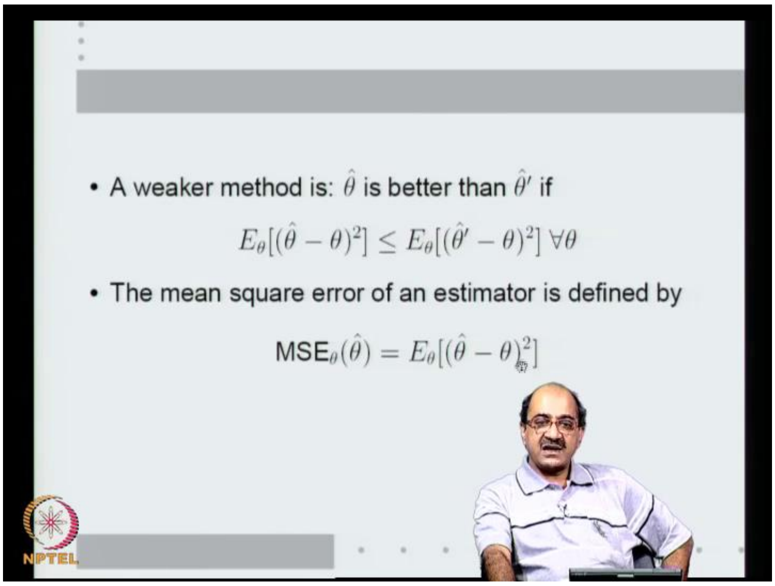
  (for any fixed sample size)
- Difficult to get such estimators.

So, one way we can ask now is, we will say that, theta hat is better than theta hat prime, if the probability that theta hat differs from theta by some quantity that is, minus a less than theta hat minus theta, less than b. This probability is greater than or equal to the probability over the same a and b, for theta hat prime, what does that mean, theta prime is closer to theta than theta hat prime.

The probability of theta hat being closer to theta is higher than the probability of theta hat prime being closer to theta because this has to hold good for all a b. I hope, all of you noticed that, I put a subscript theta on P, which means to calculate this probability, what this probably is respect to what, the random variable theta hat, theta hats distribution needs the two parameter.

So, what I am saying is ,I for whatever theta I assume in the distribution that is, the theta I am going to put here. This is a very strong requirement, no matter what my sample size is, no matter what is the accuracy level I want, no matter what values to a and b I give, theta hat is always more accurate than theta hat prime. So, if you can get this, this will be very good and then I can always say theta hat is better than theta hat prime. But, it is very difficult for any estimator to establish that level of superiority over any other. This means, for all sample sizes, for all accuracies, one estimator is uniformly better than the other estimator right that, we may or may not be able to establish.

(Refer Slide Time: 33:44)



- A weaker method is: $\hat{\theta}$ is better than $\hat{\theta}'$ if

$$E_\theta[(\hat{\theta}-\theta)^2] \leq E_\theta[(\hat{\theta}'-\theta)^2] \ \forall\theta$$

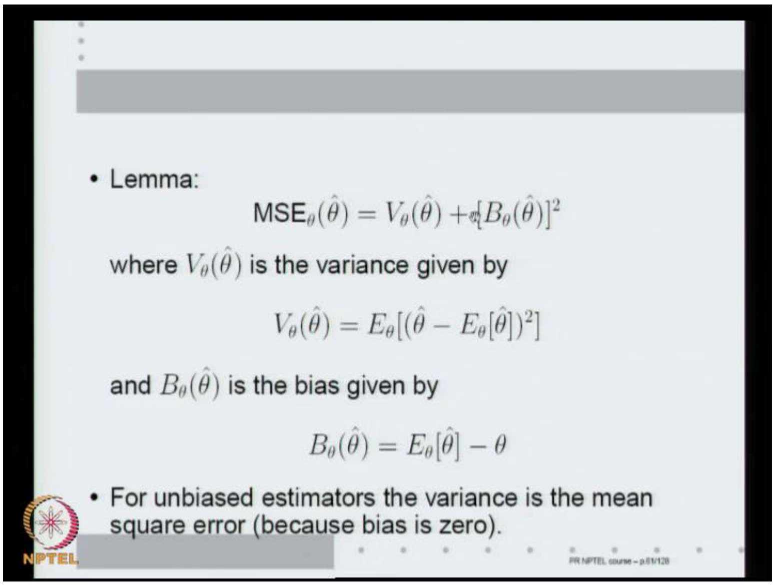
- The mean square error of an estimator is defined by

$$\text{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta}-\theta)^2]$$

So, instead of asking that, the error should be better at all levels of accuracy, we will simply say, the expectation of this square of the error that is, this is the mean square error that theta hat has. Theta hat minus theta is the error, theta hat minus theta whole square is the square of the error, if I take the expectation becomes mean square error of theta hat. So, you will say theta hat is better that theta hat prime, if the mean square error of theta hat is less than the mean square error of theta hat prime right.

This seems a reasonable thing to do because on the average, the error in theta hat is smaller than on the average, the error in theta hat prime and hence, I am willing to settle for theta hat rather than, theta hat prime. So, this is defined as the mean square error of a estimator, MSE of theta hat is defined as expected value of theta hat minus theta whole square. As you can see, all the expectations have theta at the subscript to say, this theta that I use here is whatever, the same theta I assume for taking the expectations that that is, when this is actually the error in theta hat. So, expected value of theta hat minus theta whole square is known as the mean square error of theta hat.

(Refer Slide Time: 34:59)



- Lemma:

$$\mathrm{MSE}_\theta(\hat{\theta}) = V_\theta(\hat{\theta}) + [B_\theta(\hat{\theta})]^2$$

where $V_\theta(\hat{\theta})$ is the variance given by

$$V_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$$

and $B_\theta(\hat{\theta})$ is the bias given by

$$B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$$

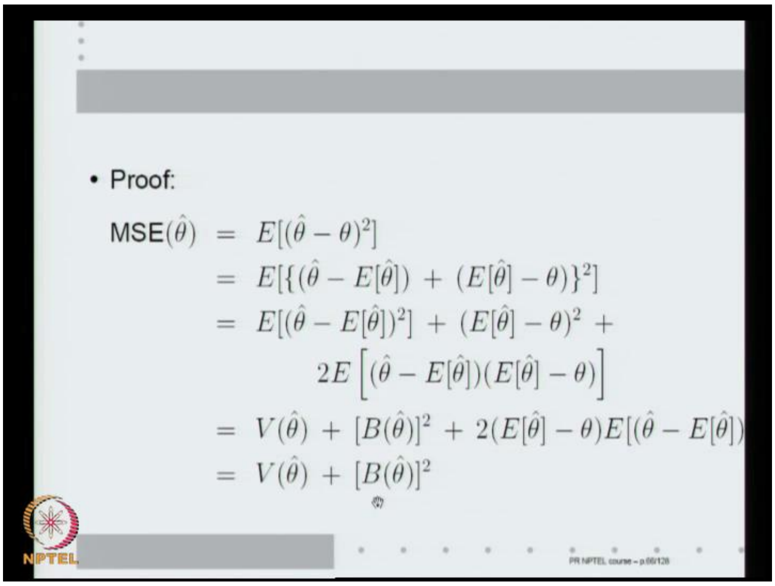
- For unbiased estimators the variance is the mean square error (because bias is zero).

Here is a very interesting result, for any estimator, the mean square of the estimator is given by sum of two quantities V theta, theta hat and V of theta hat and V of theta hat square where, V V of theta hat is the variance of theta. Because, theta hat is the random variable, what is it is variance, variance of any random variable is expectation of X minus expectation of whole square. So, variance of theta hat is expectation of theta hat minus expectation of theta hat whole square, the whole square is inside the expectation.

So, this is the variance of theta hat, this is the variance of the random variable theta hat so we call it the variance or the estimator theta hat. The bias B theta hat is called the bias of the estimator, bias of the estimator is simply expected value of theta hat minus theta right. Earlier, we we defined theta hat to be unbiased, if expectation of theta hat is equal to theta right. So, the difference between the expectation theta hat and theta is called the bias of the estimator.

So, essentially an estimator is unbiased, if it's bias is 0 and mean square error of any estimator is variance of the estimator plus square of it's bias, this lemma is not very difficult to prove so let us prove this. Before we go there, we will let us remember that, if an estimator is unbiased so that, the bias is estimated 0 then the variance of the estimator is equal to it's mean square error. So, for unbiased estimators, the mean square error is simply the variance.

(Refer Slide Time: 36:28)



- Proof:

$$
\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[\{(\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta)\}^2] \\
&= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta)^2 + \\
&\qquad 2E\left[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)\right] \\
&= V(\hat{\theta}) + [B(\hat{\theta})]^2 + 2(E[\hat{\theta}] - \theta)E[(\hat{\theta} - E[\hat{\theta}]) \\
&= V(\hat{\theta}) + [B(\hat{\theta})]^2
\end{aligned}
$$

PR NPTEL course – p.66/128

How do you prove this, what is the mean square error, expectation of theta hat minus theta whole square, I have stopped putting theta as subscript sometimes, I will put sometimes, I would not put. But, all of us understand, what this expectation means so I can rewrite this by adding and subtracting expectation of theta hat. So, I wrote theta hat minus theta whole square as, theta hat minus expectation theta hat plus expectation, theta as minus theta whole square.

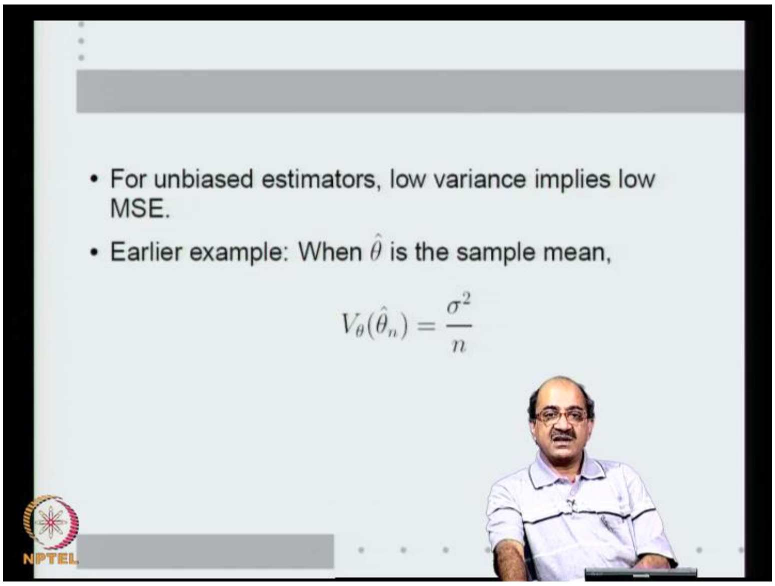
Now, I can group the first two terms and second two terms, and think of this as a plus b whole square and expand. If I expand it, I I get expectation of a square, which is theta hat minus expectation theta at whole square, the expectation of b square that is, expectation of expectation theta hat minus theta whole square. Now, theta is a constant, expectation of any random variable is a constant so expectation theta hat minus theta is a constant so expectation of that is also a constant.

I do not have to take a expectation so the second term simply becomes, expectation theta hat minus theta whole square, the third term is the 2 a b from the square that is, expectation of two times expectation of theta hat minus expectation theta hat into expectation theta hat minus theta. Now, concentrate on this 2 a b term, this the second factor here, expectation theta hat minus theta is a constant.

As we already seen, expectation theta is the constant, theta is a constant, so this constant can come out of this expectation. If it comes out, this expectation what is left is, expectation of theta hat minus expectation theta hat. Push this expectation inside, I get expectation theta hat minus expectation theta hat because expectations of expectations is itself.

And expectation theta hat minus expectation theta hat is 0 that gives us, MSE is this term expectation theta hat minus expectation of theta hat minus expectation of theta hat whole square that, we already defined at the variance of theta hat. The second term by definition is, bias square of theta hat, this is the third term where, I pulled out the constant out of the expectation. Now, this is 0 giving us mean square of theta hat is variance of theta plus bias of theta square.

(Refer Slide Time: 38:42)

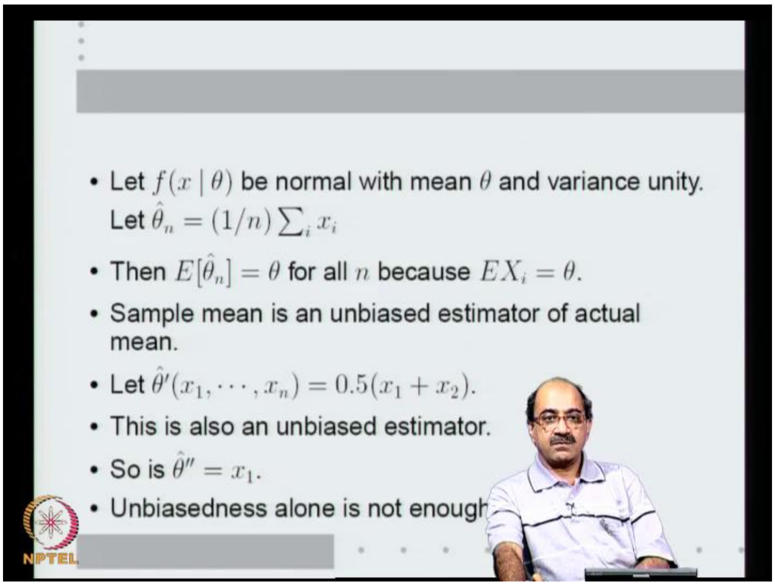


- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}$ is the sample mean,

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}$$

So, for unbiased estimators, low variance implies low mean square error so among all unbiased estimators I can choose the one, which has lower variance. So, if I go back to my earlier estimates, my when theta hat is the sample mean estimator.

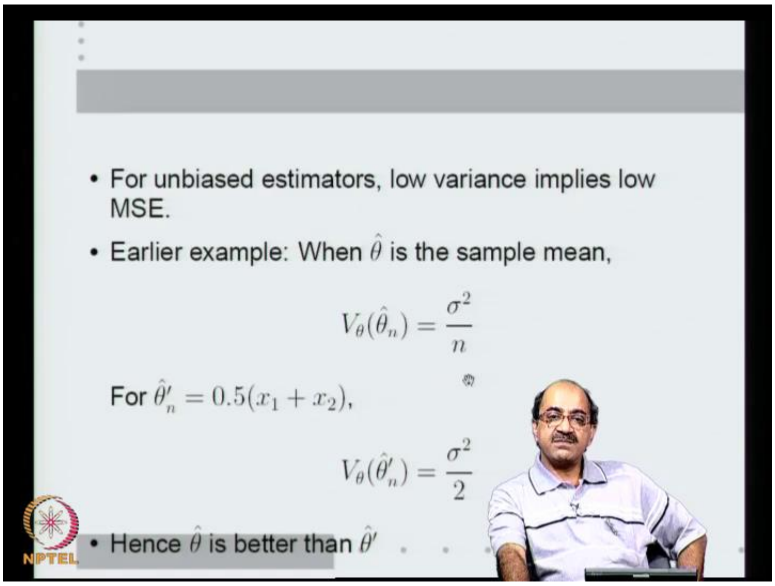
(Refer Slide Time: 39:07)



- Let $f(x \mid \theta)$ be normal with mean θ and variance unity. Let $\hat{\theta}_n = (1/n) \sum_i x_i$
- Then $E[\hat{\theta}_n] = \theta$ for all n because $EX_i = \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator.
- So is $\hat{\theta}'' = x_1$.
- Unbiasedness alone is not enough

When theta hat is the sample mean because X i's are independent random variables, variance of a sum of independent random variable is a sum of variances. So, this will be, inside the summation, this variance of this sum is, sum of variances that will be n sigma square because they are multiplying with 1 by n, it becomes 1 by n whole square into n sigma square. So, this becomes sigma square by n right whereas, this becomes sigma square by 2 and this becomes, sigma square right, so that is basically what I am getting.

(Refer Slide Time: 39:43)



- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}$ is the sample mean,

$$V_\theta(\hat{\theta}_n) = \frac{\sigma^2}{n}$$

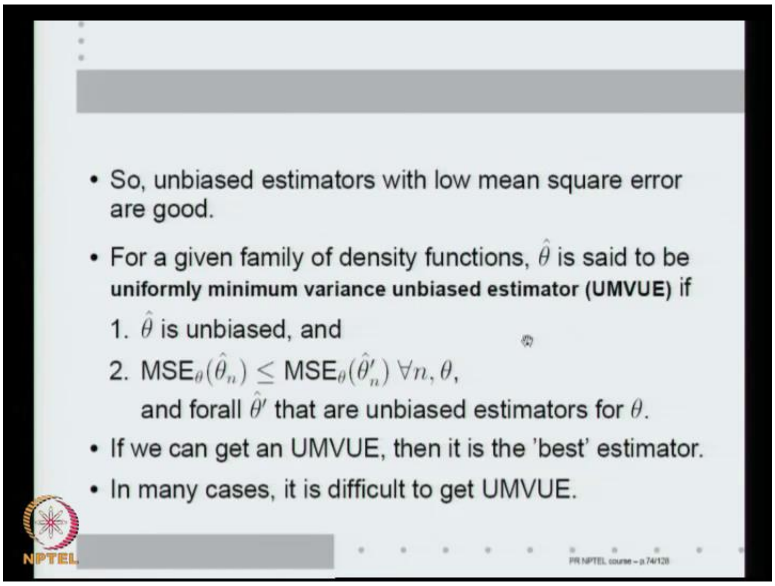
For $\hat{\theta}'_n = 0.5(x_1 + x_2)$,

$$V_\theta(\hat{\theta}'_n) = \frac{\sigma^2}{2}$$

- Hence $\hat{\theta}$ is better than $\hat{\theta}'$

So, in my earlier example for the sample mean, the variance is sigma square by n whereas, for my other estimators, when I take this, it is sigma square by 2 right. So, because the variance is smaller here than here, this is a better estimate because both are unbiased, variance is the mean square error. So, the mean square error of this estimator is smaller than mean square error of this estimator, so I can say that, this estimated theta hat n is better than theta hat prime n. So, mean square error is a good way to compare estimators and for unbiased estimators, mean square error is simply the variance.

(Refer Slide Time: 40:21)



- So, unbiased estimators with low mean square error are good.
- For a given family of density functions, $\hat{\theta}$ is said to be **uniformly minimum variance unbiased estimator (UMVUE)** if
  1. $\hat{\theta}$ is unbiased, and
  2. MSE$_\theta(\hat{\theta}_n) \leq$ MSE$_\theta(\hat{\theta}'_n)$ $\forall n, \theta$, and forall $\hat{\theta}'$ that are unbiased estimators for θ.
- If we can get an UMVUE, then it is the 'best' estimator.
- In many cases, it is difficult to get UMVUE.

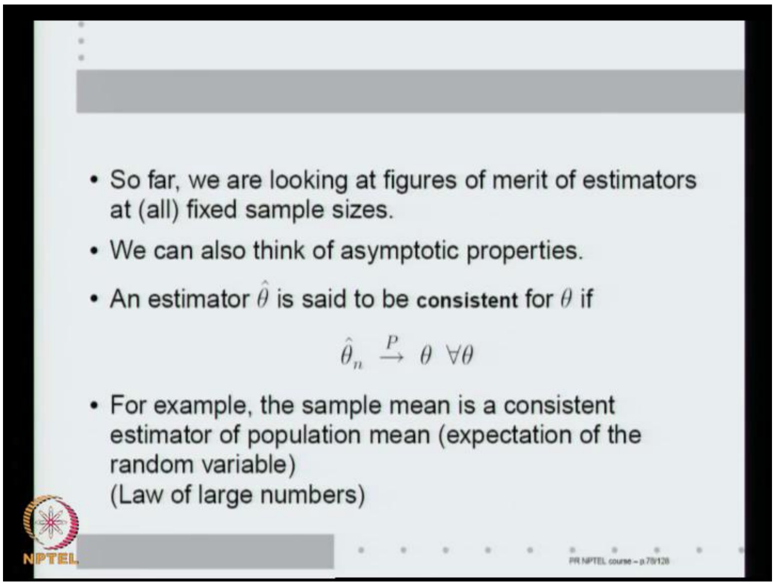
PR NPTEL course – p.74/128

So, unbiased estimators with low mean square error are good estimators, for a given family of density functions theta hat, a specific estimate at theta hat is said to be uniformly minimum variance unbiased estimator, often written as UMVUE, uniformly minimum variance and unbiased. If firstly, theta hat is unbiased and mean square error of theta hat, which is same as variance because unbiased, is less than the mean square error of any other theta hat prime where, theta hat prime is unbiased estimator.

So, theta hat is UMVUE that is, uniformly minimum variance unbiased, if variance of theta hat is smaller than variance of any other unbiased estimator theta hat prime. Now, this has to hold for every single n that is, what is meant by uniformly minimum variance. We are not saying that, at some n it is minimum, for every single n, the variance of theta hat n is less than the variance of theta hat prime n.

So, for every n, for all theta, if the variance at theta hat is less than theta hat prime and theta hat is unbiased then that theta hat is called the uniformly minimum variance unbiased estimator. So, as a matter of fact, if I can get UMVUE nothing like that but in many cases, it is difficult to get UMVUE and also, there may not be many standard procedures for getting UMVUE.

(Refer Slide Time: 41:50)



- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.
- We can also think of asymptotic properties.
- An estimator $\hat{\theta}$ is said to be **consistent** for θ if

$$\hat{\theta}_n \xrightarrow{P} \theta \ \forall\theta$$

- For example, the sample mean is a consistent estimator of population mean (expectation of the random variable)
  (Law of large numbers)

PR NPTEL course – p.70/128

So, let us look at some other figures of merit so everything that we looked so far, we are saying, for every single n something has to hold, unbiased (( )) is the expectation of theta hat, n is equal to theta, for all n. We have talked of UMVUE once again, minimum variance at all fixed sample sizes now, if I if you let go of that and only ask for, as the samples size goes to infinity, is the estimated good right, that is also a good way of looking at it, we can think of asymptotic properties of estimators right.

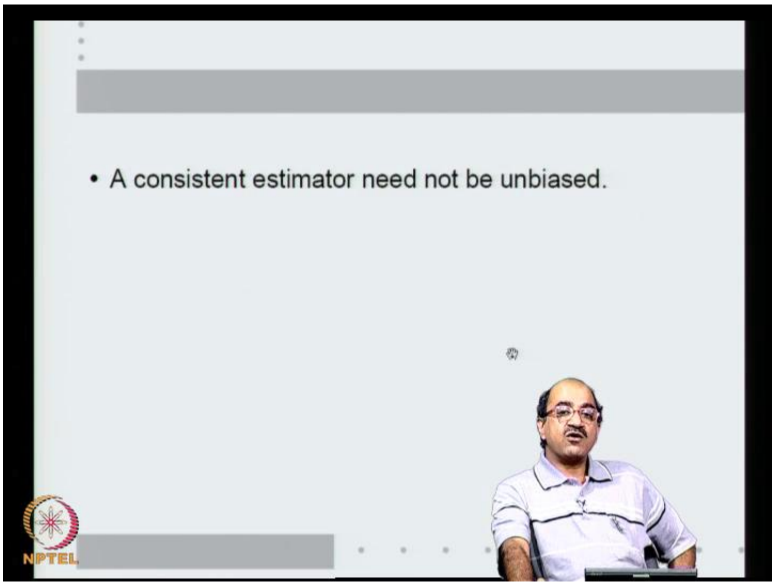
So, one asymptotic property is to say, at the sample size goes to infinity thus, the estimator converts to the true value thus, theta hat n converts to theta. Since theta hat n is a sequence of random variables, as I vary n, it becomes a sequence of random variables so when I say convergence, I have to say convergence in what sense, is convergence in distribution, convergence almost truly. So, we will take convergence in probability because there is a convenient mode of convergence for our purposes.

So, we will say, an estimate of theta hat is said to be consistent for theta, if theta hat n converges in probability to theta, as n tends to infinity, this is an asymptotic property.

But, essentially what it what this means is because of the convergence in probability, if n is sufficiently large, the probability that theta hat n and theta differ by say, some epsilon, can be made less than delta that is what, this convergence in probability means.

So, a consistent estimator is good because for large sample size the estimator will be close to the true value of the parameter. So, an estimator theta hat is said to be consistent, if theta hat n converges in probability to theta. We know, the sample been estimator converges to (()), sample estimator 1 by n summation X i now, by law of large numbers, sample mean converges to the population mean in probability, with law of large numbers. So, a sample mean estimator in addition to, being unbiased in addition to, being minimum variance is also a consistent estimator. Our interest in the consistent estimators come from the fact, before we go there, a consistent estimator does not have to be unbiased.

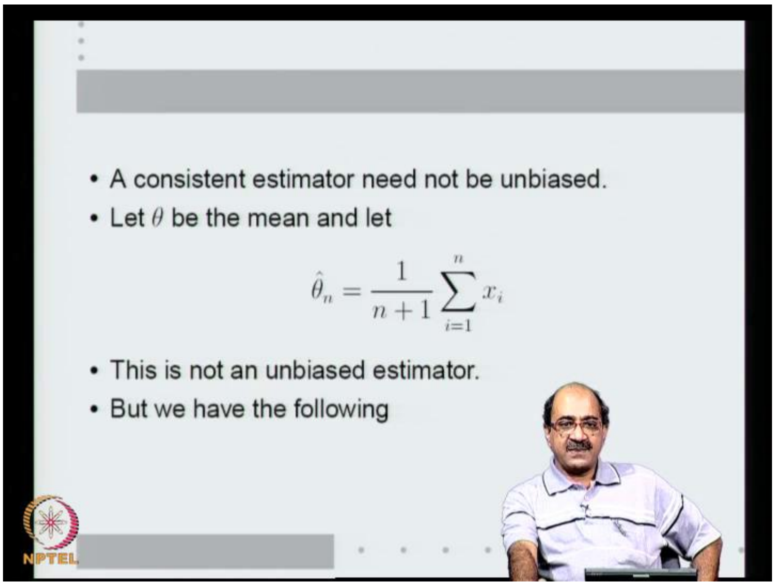
(Refer Slide Time: 44:03)



- A consistent estimator need not be unbiased.

Consistency is an asymptotic property so even if the estimator is biased, it may still be consistent right.

(Refer Slide Time: 44:04)



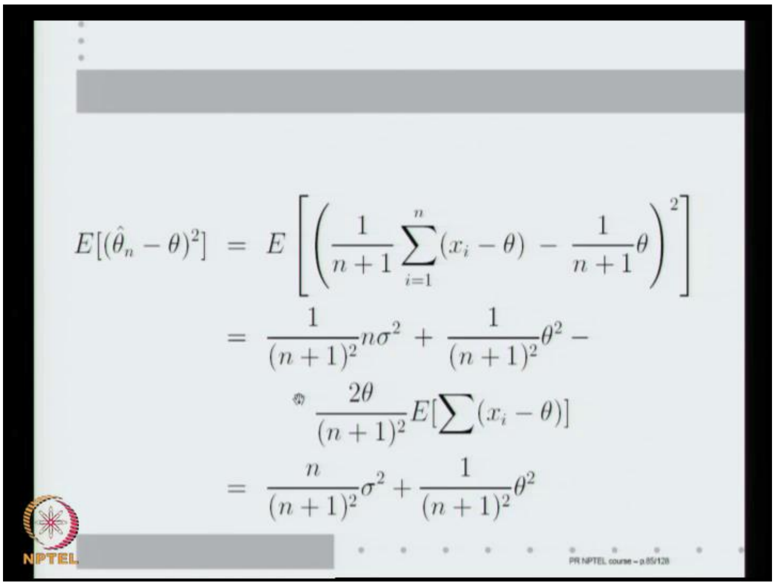
- A consistent estimator need not be unbiased.
- Let θ be the mean and let

$$\hat{\theta}_n = \frac{1}{n+1}\sum_{i=1}^{n} x_i$$

- This is not an unbiased estimator.
- But we have the following

For example, let us instead of taking sample mean, I will take 1 by n plus 1 summation X instead of, 1 by n summation X n, is obviously biased because expected value of theta hat n is not equal to theta but it is n by n plus 1 theta. So, there is bias but as you can see, as n goes to large then whether you divide by n or n plus 1 may may not make much difference. And one would expect that, theta hat n will converse to theta and we can prove that right, this is not an unbiased estimator.

(Refer Slide Time: 44:36)



$$\begin{aligned} E[(\hat{\theta}_n - \theta)^2] &= E\left[\left(\frac{1}{n+1}\sum_{i=1}^{n}(x_i - \theta) - \frac{1}{n+1}\theta\right)^2\right] \\ &= \frac{1}{(n+1)^2}n\sigma^2 + \frac{1}{(n+1)^2}\theta^2 - \\ &\quad \frac{2\theta}{(n+1)^2}E[\sum(x_i - \theta)] \\ &= \frac{n}{(n+1)^2}\sigma^2 + \frac{1}{(n+1)^2}\theta^2 \end{aligned}$$

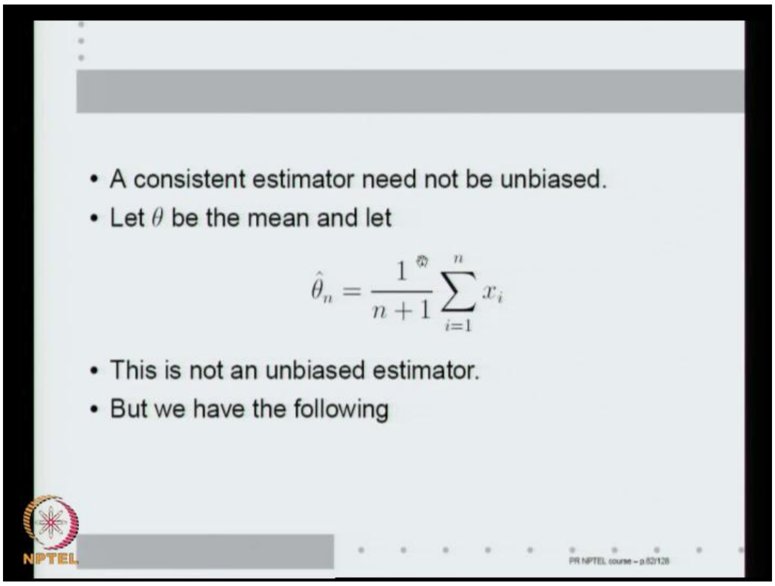
PR NPTEL course – p.85/128

But, we can prove the following, expectation of theta hat and minus theta whole square, this will because it is a this gives me the mean square error of the estimator. So, this is theta hat n, theta hat n is 1 by n plus 1 x i so I wrote theta as n by n plus 1 theta. If I push this summation side, I will get n by n plus 1 theta and the remaining, 1 by n plus 1 theta I wrote separately, square. Now, you can square this because x i's are independent, when I squared it, all the cross terms will cancel, when I take the expectation side.

So, that will ultimately give me 1 by n plus 1 whole square into only the squares of x i minus theta whole square will be there. So, there will be n sigma n such sigma comes, which are sigma square, this will give me 1 by n plus 1 theta square. And the 2 a b term, which is 2 theta by n plus 1 whole square into expectation of x i minus theta, which is 0 because if I put the expectation inside, expectation x i is equal to theta.

So, which just gives me 1 by n plus 1 whole square n sigma, n by n plus 1 whole square sigma square, 1 by n plus 1 whole square theta square, as n tends to infinity, this goes to 0. So, expected value of theta hat and minus theta whole square goes to 0, as n tends to infinity, which means theta hat n converges to theta n quadratic mean and hence, it will also converge in probability.

(Refer Slide Time: 46:00)



- A consistent estimator need not be unbiased.
- Let θ be the mean and let

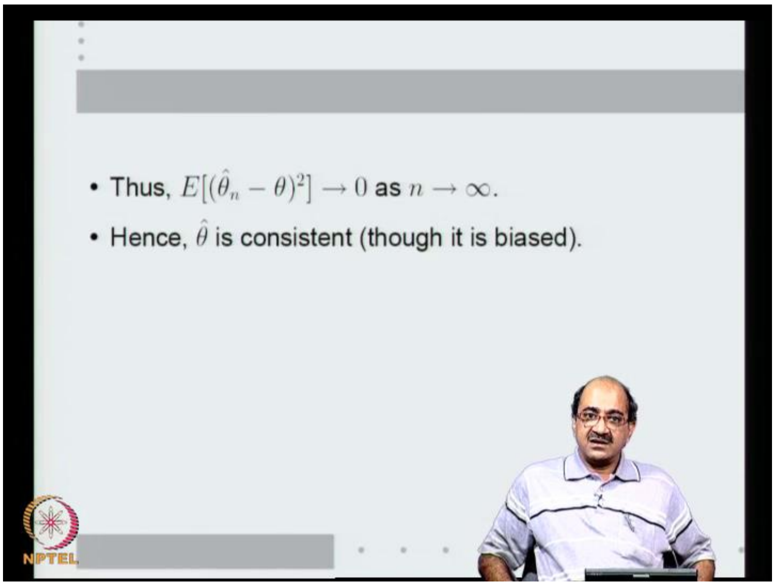
$$\hat{\theta}_n = \frac{1}{n+1}\sum_{i=1}^{n} x_i$$

- This is not an unbiased estimator.
- But we have the following

PR NPTEL course – p.82/128

So, this theta hat n what we saw earlier, even though it is a biased estimator, it is a consistent estimator right, because it as n tends to infinity converges to two value.

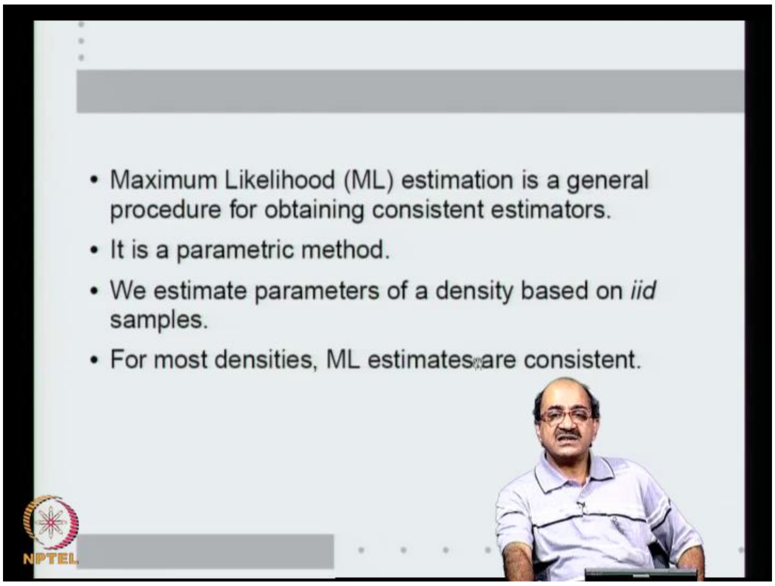
(Refer Slide Time: 46:09)



- Thus, $E[(\hat{\theta}_n - \theta)^2] \rightarrow 0$ as $n \rightarrow \infty$.
- Hence, $\hat{\theta}$ is consistent (though it is biased).

So, this goes to 0, as n tends to infinity hence, theta hat is the consistent estimator though it is biased.

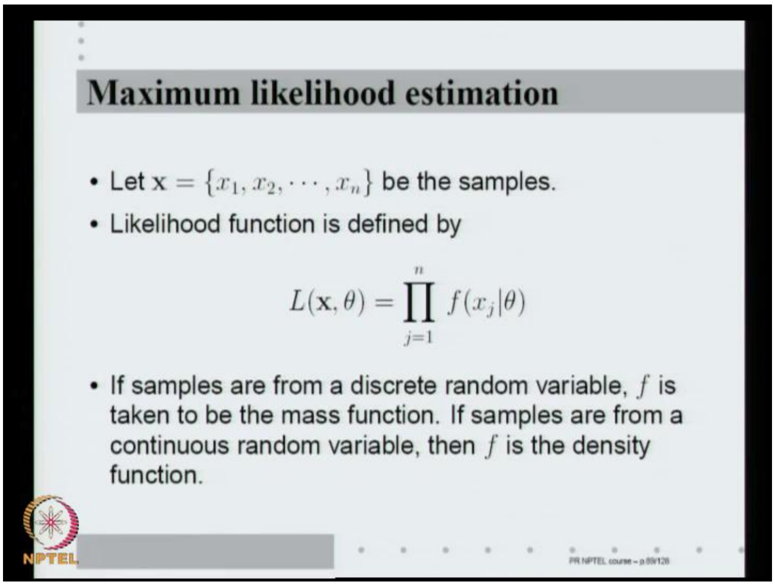
(Refer Slide Time: 46:28)



- Maximum Likelihood (ML) estimation is a general procedure for obtaining consistent estimators.
- It is a parametric method.
- We estimate parameters of a density based on *iid* samples.
- For most densities, ML estimates are consistent.

But, the good thing about consistency is that, we have a general procedure for obtaining consistent estimator, maximum likelihood estimation is a general procedure for obtaining consistent estimators. It is a parametric method, we estimate parameters of a density based on iid samples and the nice thing is, if the density satisfies some simple regularity conditions then the maximum likelihood estimates can be proved to be consistent.

(Refer Slide Time: 46:50)



Maximum likelihood estimation

- Let $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ be the samples.
- Likelihood function is defined by

$$L(\mathbf{x}, \theta) = \prod_{j=1}^{n} f(x_j|\theta)$$

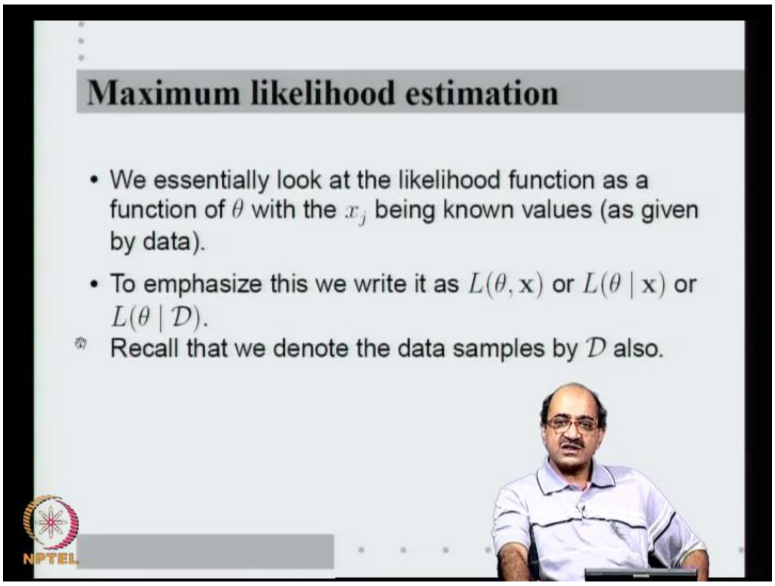
- If samples are from a discrete random variable, f is taken to be the mass function. If samples are from a continuous random variable, then f is the density function.

PR NPTEL course – p.89/128

This is how, we do maximum likelihood estimation once again, x 1, x 2, x n be the samples, the likelihood function, we defined a likelihood function, which is the function of x and the parameter vector theta. As L x theta is product, j grown 1 to n, f x j given theta so far, we have always been talking of f as density but it really does not matter. If the samples are from a discrete random variable, f is taken to be the mass function, if they are taken from a continuous random variable, f is the density function, in both cases we define this product as the likelihood.

Intuitively, if it is the density function, if it is the mass function, this product gives you the probability of obtaining the sample. This density of course, is not a probability but even then it is called the likelihood. Of course, the the reason for calculating likelihood is not about calculating about x j because x j's in any, when I am estimating, I have got specific sample and I am estimating theta.

(Refer Slide Time: 47:48)

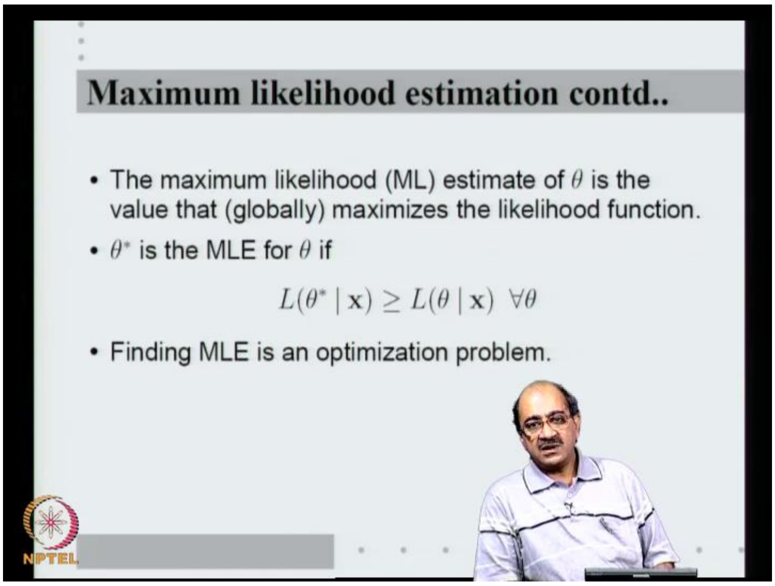


Maximum likelihood estimation

- We essentially look at the likelihood function as a function of θ with the x_j being known values (as given by data).
- To emphasize this we write it as $L(\theta, \mathbf{x})$ or $L(\theta \mid \mathbf{x})$ or $L(\theta \mid \mathcal{D})$.
- Recall that we denote the data samples by $\mathcal{D}$ also.

So, we often would like to think of likelihood, not often we always think of likelihood, as a function of theta, with the sample being known, the sample is data. So, to emphasize this, we often write theta as the first variable or more more often, we write it as L of theta given x or L of theta given D. Because, D is the notation for samples, we always write the likelihood function as L of theta given x, L of theta given D because likelihood function is viewed as a function of theta.

(Refer Slide Time: 48:18)



Maximum likelihood estimation contd..

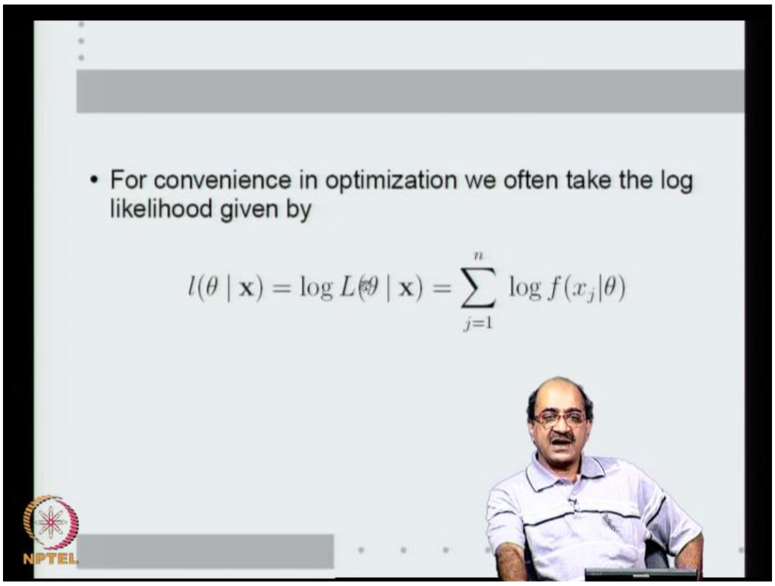
- The maximum likelihood (ML) estimate of θ is the value that (globally) maximizes the likelihood function.
- θ^* is the MLE for θ if

$$L(\theta^* \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}) \ \forall \theta$$

- Finding MLE is an optimization problem.

So, the maximum likelihood estimate is the value of theta that globally maximizes the likelihood, I said we will look at likelihood function, as a function of theta. So, theta star is MLE that is, maximum likelihood estimate for theta, if L of theta star given x is greater than equal to L of theta given x, for all theta. So, the value of theta that globally maximizes the function L theta given x is called the maximum likelihood estimator. So, finding MLE is essentially an optimizational problem, if I am given the function L theta given x, as a function of theta, how to find it is global maximum, this is an optimization problem.

(Refer Slide Time: 48:58)

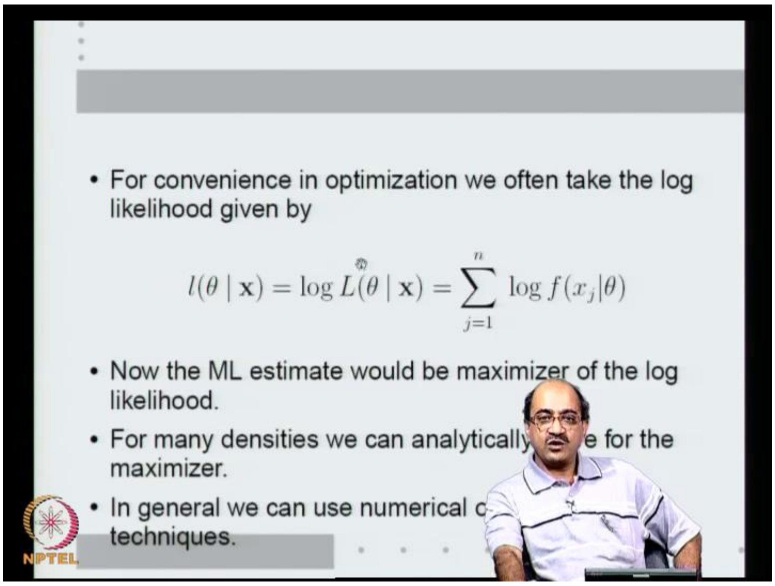


- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

Very very often for convenience in the optimization, we take what is called the log likelihood, take the log of the likelihood function, the reason is, likelihood function as you as you have seen as the product so if I take log, it becomes summation. So, log of L theta given x is summation of log of f x, x j given theta and we represent the log likelihood by little l given theta x, likelihood as capital L and log likelihood as small l.

(Refer Slide Time: 49:37)



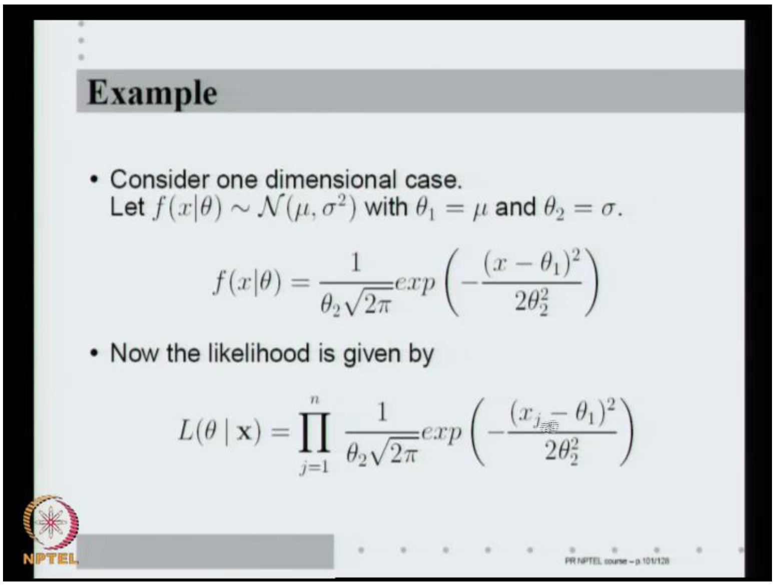
- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- Now the ML estimate would be maximizer of the log likelihood.
- For many densities we can analytically e for the maximizer.
- In general we can use numerical c techniques.

Now, the ML estimate would be maximize of the log likelihood because log is a monotone function, whatever maximizes, the likelihood will also maximise the log likelihood. So, for many densities, we can analytically calculate the maximize and if you cannot of course, you can always use a numerical technique, if you know the likelihood function to obtain the MLE estimate.

(Refer Slide Time: 49:54)



Example

- Consider one dimensional case.
Let $f(x|\theta) \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta_1 = \mu$ and $\theta_2 = \sigma$.

$$f(x|\theta) = \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

- Now the likelihood is given by

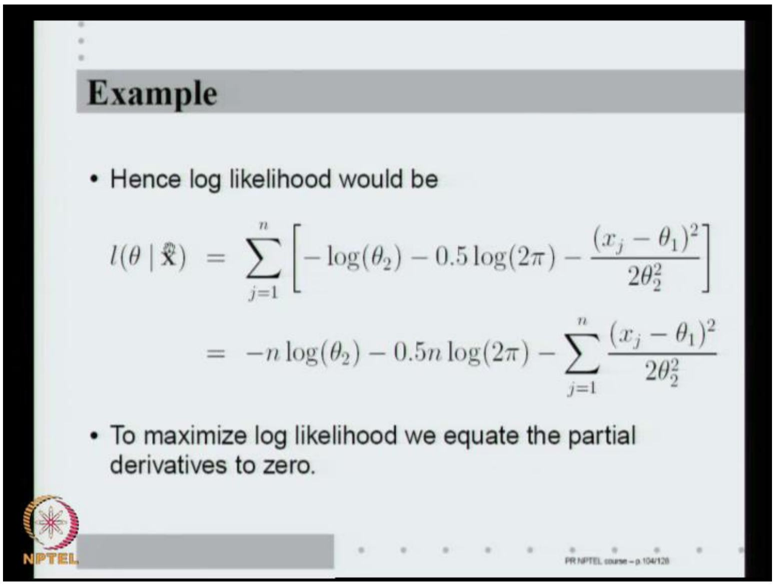
$$L(\theta \mid \mathbf{x}) = \prod_{j=1}^{n} \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x_j-\theta_1)^2}{2\theta_2^2}\right)$$

So, let us consider one example, let us say, one dimensional case let us say, x is normal with mean mu and variance sigma square, so there are two parameters theta 1 and theta

2, theta 1 is mu and theta 2 is sigma square. So, we take a theta to be sigma instead of, sigma square, we could have taken second parameter to be sigma square or sigma, we have taken to be sigma. Then the density becomes 1 by theta 2, root 2 pi exponential minus x minus theta one whole square by 2 theta 2, square.

So, what is my likelihood, L theta given x is product over j is equal to 1 to n, f of x j given theta, f of x j given theta is the same expression where, x is now replaced by x j. Now, what will be the log likelihood, you take log of this. So, log of with they will be sum and log of this, log of this will be log of the first term plus log of second term. Log of exponential will be only, what is inside the exponential, so that becomes the log likelihood. Sum j is equal to 1 to n, log of this will be minus log theta to minus half log two pi and then what is inside.

(Refer Slide Time: 51:07)



Example

- Hence log likelihood would be

$$l(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j - \theta_1)^2}{2\theta_2^2} \right]$$

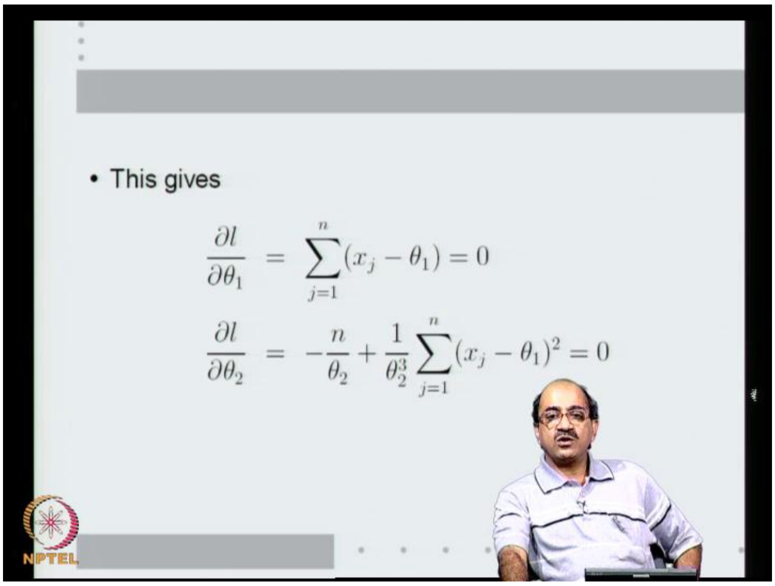
$$= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2^2}$$

- To maximize log likelihood we equate the partial derivatives to zero.

PR NPTEL course – p.104/128

So, minus log theta 2 minus half log 2 pi minus x j minus theta 1 whole square by 2 theta 2 square, so this is the log likelihood function. We can simplify it, push the summation inside so there n log theta 2 minus 0.5, 1 log 2 pi minus summation, j is equal to 1 to n x j minus theta 1 whole square by 2 theta 2, square. This is my log likelihood function, I am asking which values of theta 1 and theta 2 maximize the log likelihood right. So, how do I maximize, there are two which is a function of two two variables, theta 1 theta 2, we will find the partial derivatives and equate them to 0 right.

(Refer Slide Time: 51:43)



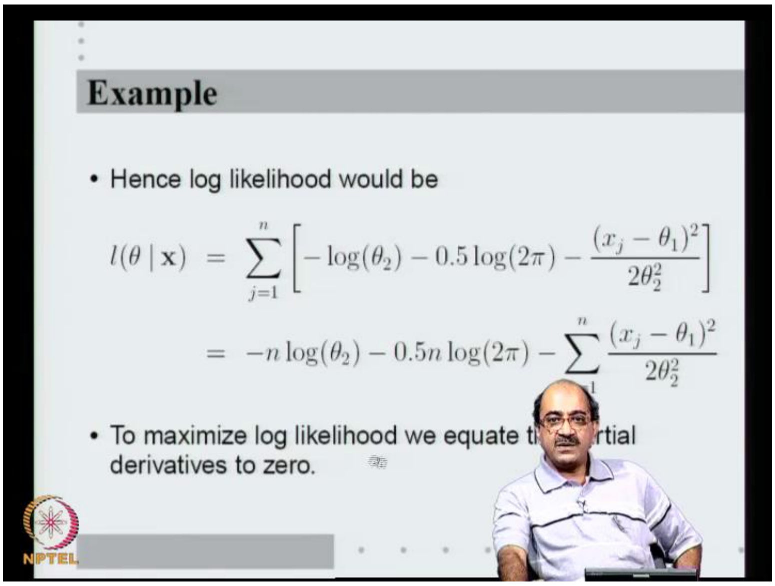
- This gives

$$\frac{\partial l}{\partial \theta_1} = \sum_{j=1}^{n}(x_j - \theta_1) = 0$$

$$\frac{\partial l}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j - \theta_1)^2 = 0$$

So, what you do, dou l by dou theta 1, dou l by dou theta 2 equate them to 0.

(Refer Slide Time: 51:49)



Example

- Hence log likelihood would be

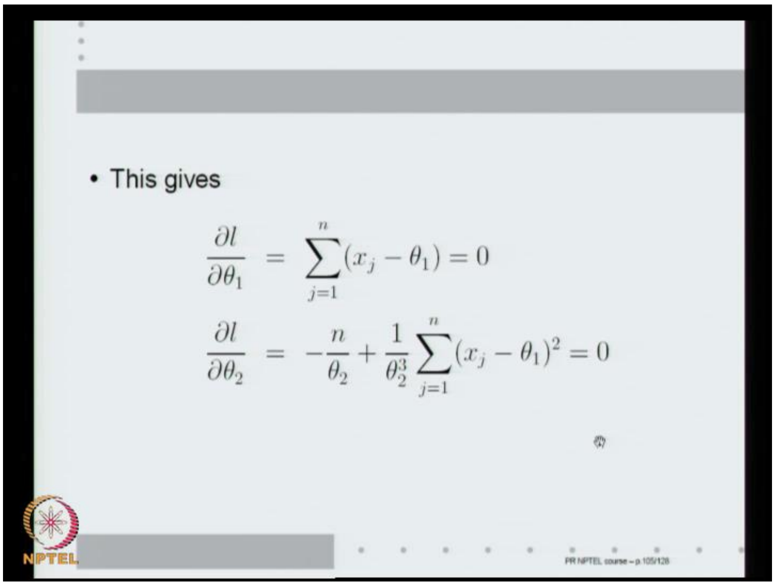
$$l(\theta \mid \mathbf{x}) = \sum_{j=1}^{n}\left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j - \theta_1)^2}{2\theta_2^2}\right]$$

$$= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n}\frac{(x_j - \theta_1)^2}{2\theta_2^2}$$

- To maximize log likelihood we equate the partial derivatives to zero.

What will be dou l by dou theta 1. If I differentiate this with respect to theta 1, derivative of this is 0, derivative of this is 0, derivative of this will be half, summation of this half, x j minus theta 1 by 2 theta 2 square right into minus 1, that minus will cancel.

(Refer Slide Time: 52:08)



- This gives

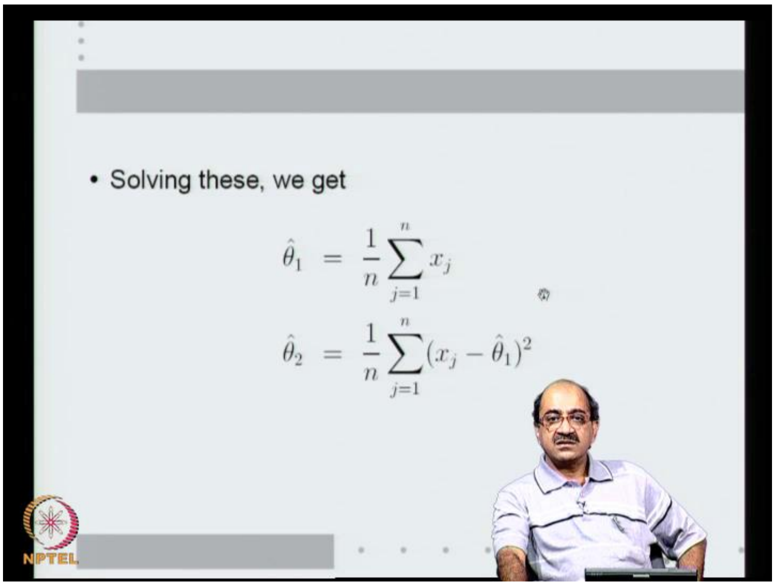
$$\frac{\partial l}{\partial \theta_1} = \sum_{j=1}^{n}(x_j - \theta_1) = 0$$

$$\frac{\partial l}{\partial \theta_2} = -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j - \theta_1)^2 = 0$$

PR NPTEL course – p.105/128

So, if I equate that to 0, this is what I will get, j is equal to 1 to n, x j minus theta 1 is equal to 0 right. That half will also go and 1 by 2 theta 2 will square will also go because equal equate it to 0. Now, if I crunch this, I get n times theta 1 is equal to summation x j, which means theta 1 is 1 by n summation x j.

(Refer Slide Time: 52:30)



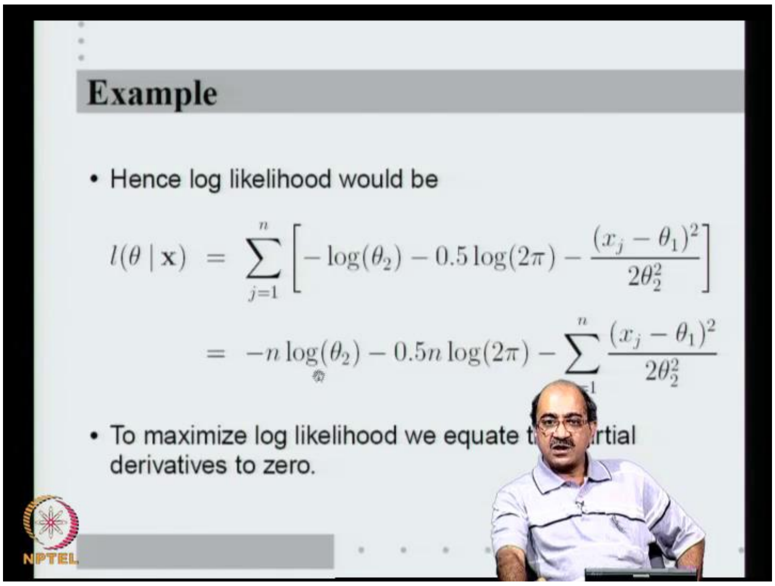
- Solving these, we get

$$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\theta}_1)^2$$

So, that is what, I get as my estimator of theta 1 similarly, partial derivative with respect to theta 2.

(Refer Slide Time: 52:41)



Example

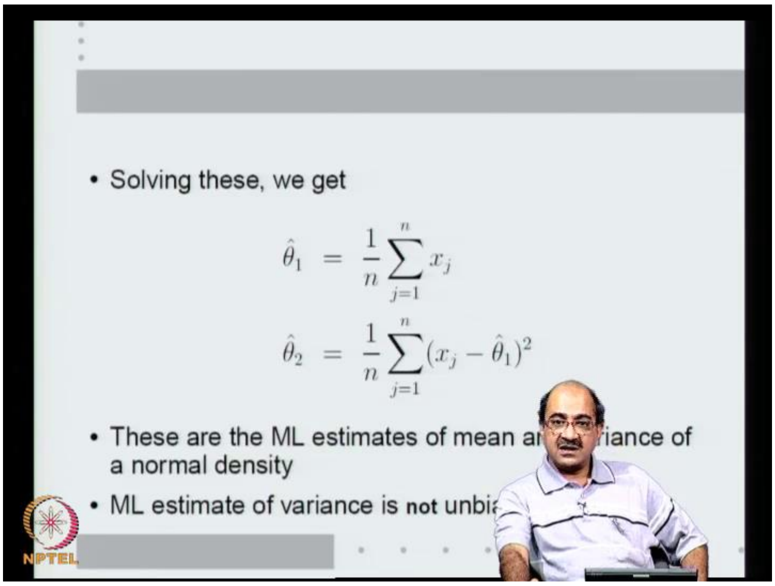
- Hence log likelihood would be

$$l(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j-\theta_1)^2}{2\theta_2^2} \right]$$
$$= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j-\theta_1)^2}{2\theta_2^2}$$

- To maximize log likelihood we equate t rtial derivatives to zero.

So, I have to differentiate, this is theta 2, this give me n by theta 2, this will give me x j minus theta 1 whole square that is a constant now we can define, 2 will also stay here. So, 1 by theta 2 square will give me minus 2 by theta 2 cube, 1, 2 will go away so minus n by theta 2, that minus has gone, 1 by theta 2 cube into this equal to 0 right. So, if I take this term on this side, multiply by theta 2, so I will get 1 by theta 2 square into something and then equate to 0 and solve.

(Refer Slide Time: 53:16)



- Solving these, we get

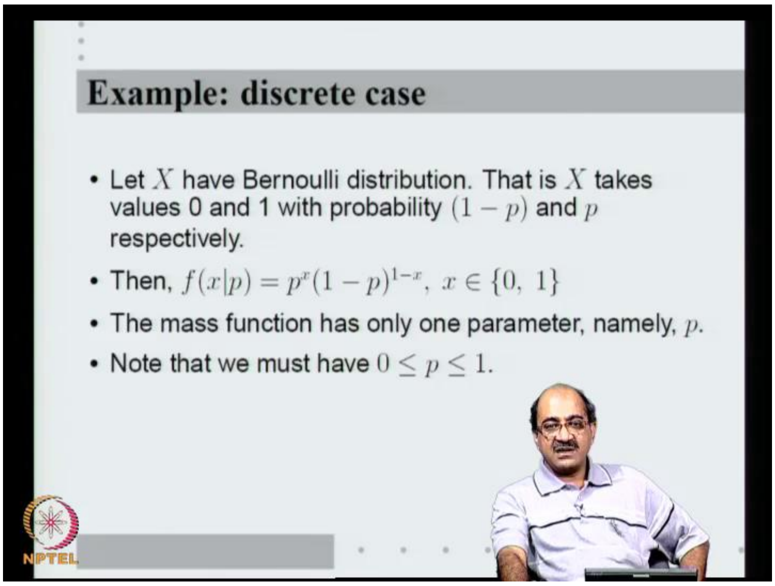
$$\hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$
$$\hat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2$$

- These are the ML estimates of mean ar iance of a normal density
- ML estimate of variance is **not** unbia

I get theta 2 hat is 1 by n x j minus theta 1 whole square so by solving del l by del theta 1 is equal to 0 and del l by del theta 2 is equal to, dou l by dou theta is equal to 0 and dou l by dou theta 2 equal to 0, we get the maximum likelihood estimators. The estimator for theta 1 is the sample mean estimator for theta 2 square, not theta 2 hat, the estimator for variance is the sample variance.

As some of you may know, 1 by n x j minus theta 1 whole square is not an unbiased estimator variance, if I want to get unbiased estimator, I have to get 1 by n minus 1. So, these are the ML estimator mean and variance of a normal density and ML estimator of variance is not unbiased. So, as we have already seen, consistent estimators need not have to be unbiased and ML estimation only guarantees consistency sometimes, we may land up with unbiased with biased estimators. But for one dimensional normal density, these are the estimators, a maximum likelihood estimators for mean and variance.

(Refer Slide Time: 54:28)

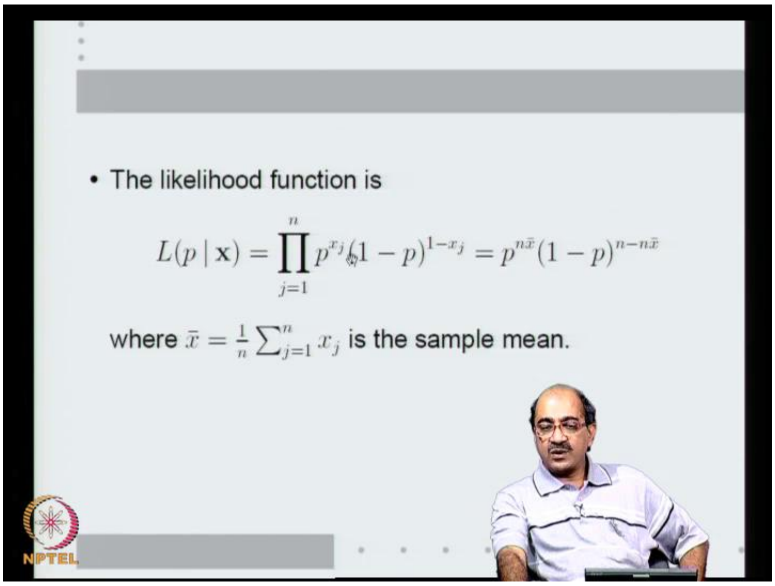


Example: discrete case

- Let X have Bernoulli distribution. That is X takes values 0 and 1 with probability $(1-p)$ and p respectively.
- Then, $f(x|p) = p^x(1-p)^{1-x}$, $x \in \{0,\ 1\}$
- The mass function has only one parameter, namely, p.
- Note that we must have $0 \leq p \leq 1$.

We can do this same thing for a discrete random variable also let us say, X is a Bernoulli distribution that is, x takes values 0 and 1, with probability p and 1 minus p. So, I can write the mass function of x parameter, as p has power x, and 1 minus p power 1 minus x, x takes only 0 and 1. So, f of 0, the mass function at value 0 will be 1 minus p, mass function value at 1 will be p so this is the Bernoulli density right, this the mass function Bernoulli takes values 0 and 1, takes value 0 with probability 1 minus p, 1 with probability p.

(Refer Slide Time: 55:26)



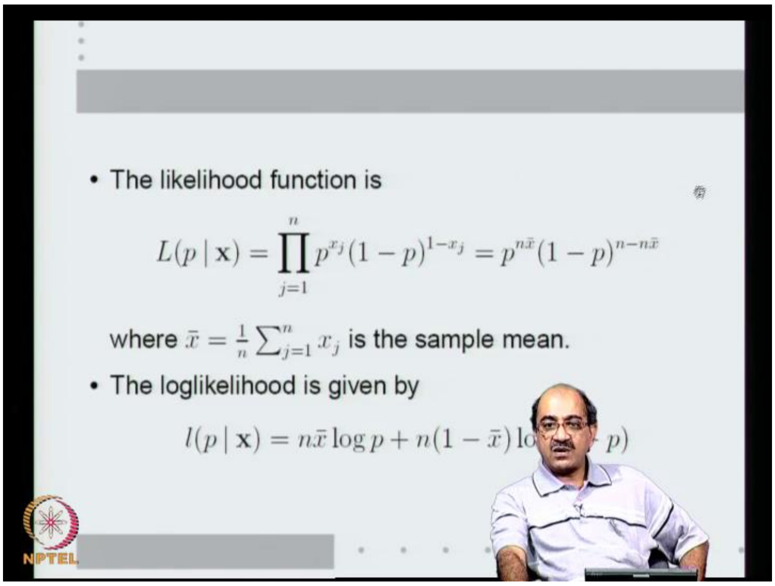
- The likelihood function is

$$L(p\,|\,\mathbf{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is the sample mean.

So, the mass function has only one parameter namely p of course, we must ensure 0 less than p less than 1. So, if I simply maximize this, over all p is not what I want, I have to maximize this over maximize the likelihood, over p between 0 and 1. But, as it turns out, if we just unconstraintly maximize, you will anyway get p between 0 and 1.

(Refer Slide Time: 54:26)



- The likelihood function is

$$L(p\,|\,\mathbf{x}) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

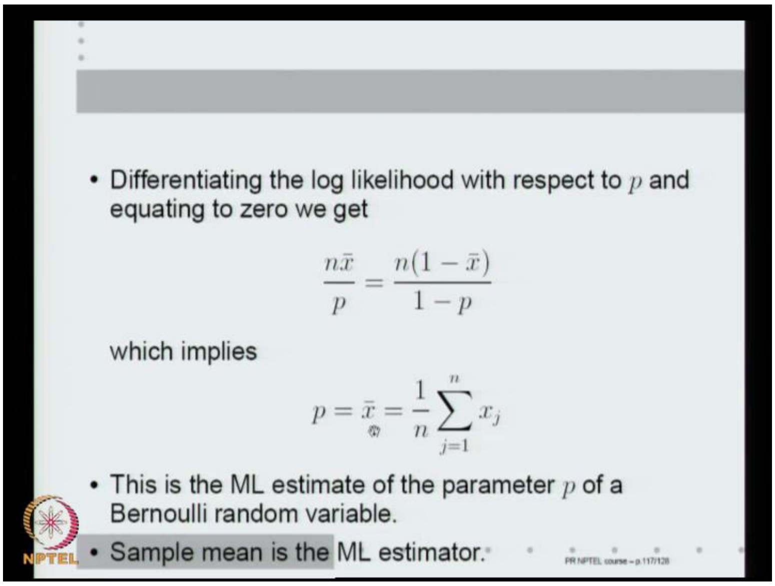
where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is the sample mean.

- The loglikelihood is given by

$$l(p\,|\,\mathbf{x}) = n\bar{x}\log p + n(1-\bar{x})\,\mathrm{lo} \quad p)$$

So, what is the likelihood function so j is equal to 1 to n so this is my thing so I put x j here. So, p x j, 1 minus p, 1 minus x j because the product, it is p times summation x j, 1 minus p times summation 1 minus x j. If I write x bar at the sample mean then summation x j is nothing but an x bar so this becomes p times n x bar, 1 minus p times, n minus n x bar. So, the log likelihood will be n x bar log p plus n into, 1 minus x bar, log 1 minus p, this is very simple thing to differentiate.

(Refer Slide Time: 56:03)



- Differentiating the log likelihood with respect to p and equating to zero we get

$$\frac{n\bar{x}}{p} = \frac{n(1-\bar{x})}{1-p}$$

which implies

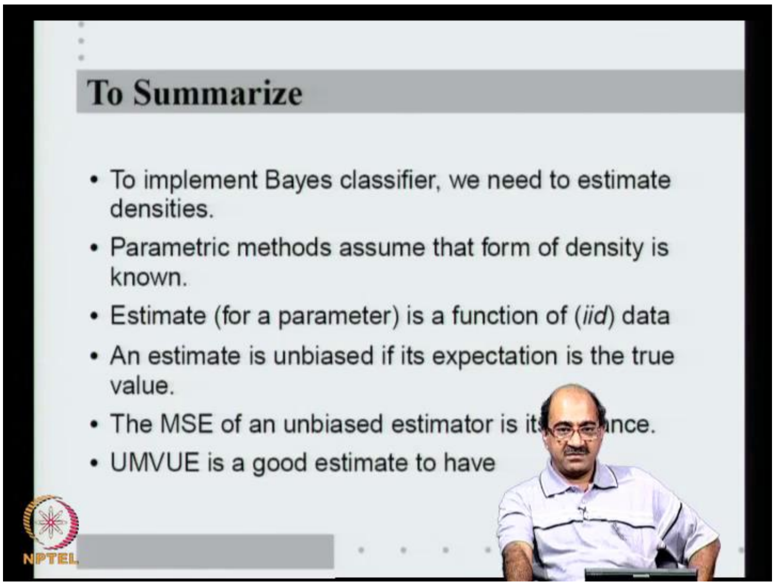
$$p = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- This is the ML estimate of the parameter p of a Bernoulli random variable.
- Sample mean is the ML estimator.

PR NPTEL course – p.117/128

If we differentiate, we get this and from there, you get p once again as the sample mean because x j's take only values 0 and 1, summation x j by n will be between 0 and 1. So, even though, I have maximized likelihood in a unconstrained manner, I am still getting p between 0 and 1, so that is all right. So, this is the maximum likelihood estimator for the parameter of a Bernoulli random variable.

(Refer Slide Time: 56:28)

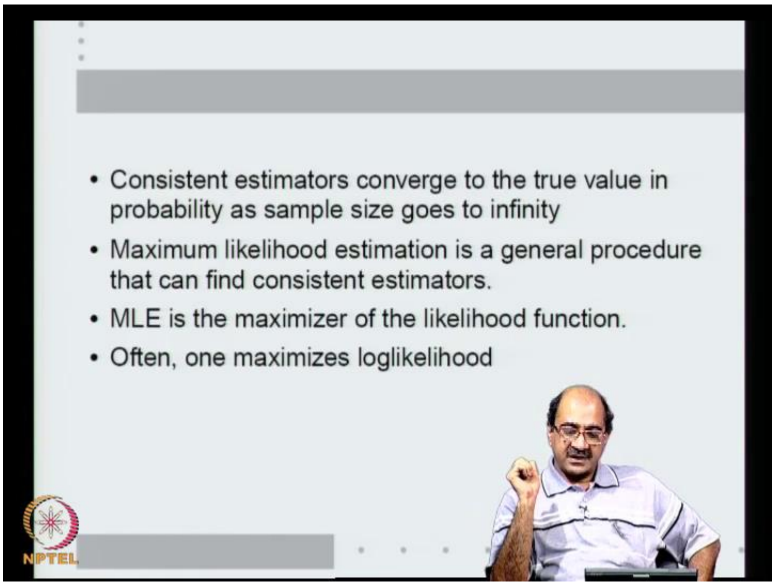


To Summarize

- To implement Bayes classifier, we need to estimate densities.
- Parametric methods assume that form of density is known.
- Estimate (for a parameter) is a function of (*iid*) data
- An estimate is unbiased if its expectation is the true value.
- The MSE of an unbiased estimator is it[illegible]nce.
- UMVUE is a good estimate to have

So, let us summarize today's lecture, to implement Bayes classifier, we need to estimate densities. Parametric methods assumed that, form of density is known and then obtain the parameters from the data. An estimate for a parameter is a function of the data, for all estimation we assume we have iid realizations of the random variable or iid data from a density and an estimate is a function of this data, is a statistic, is the the function of this data. So, estimate is the function of the iid data, an estimate is unbiased, if it's expectation is the true value. The mean square of an unbiased estimator is it is variance and uniformly, minimum variance unbiased estimators are very good to have, if you can have them.

(Refer Slide Time: 57:20)



- Consistent estimators converge to the true value in probability as sample size goes to infinity
- Maximum likelihood estimation is a general procedure that can find consistent estimators.
- MLE is the maximizer of the likelihood function.
- Often, one maximizes loglikelihood

A consistent estimator converges the true value in probability, as the sample size goes to infinity right, maximum likelihood estimation is a general procedure, that can find consistent estimators. MLE is a maximizer of the likelihood function, often one maximizes the log likelihood, just for convenience in maximization. And for many standard densities, one can obtain Bayes likelihood through simple analytical means.

Thank you.