Statistical Methods of Scientists and Engineers

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

Lecture - 09
Parametric Methods - I

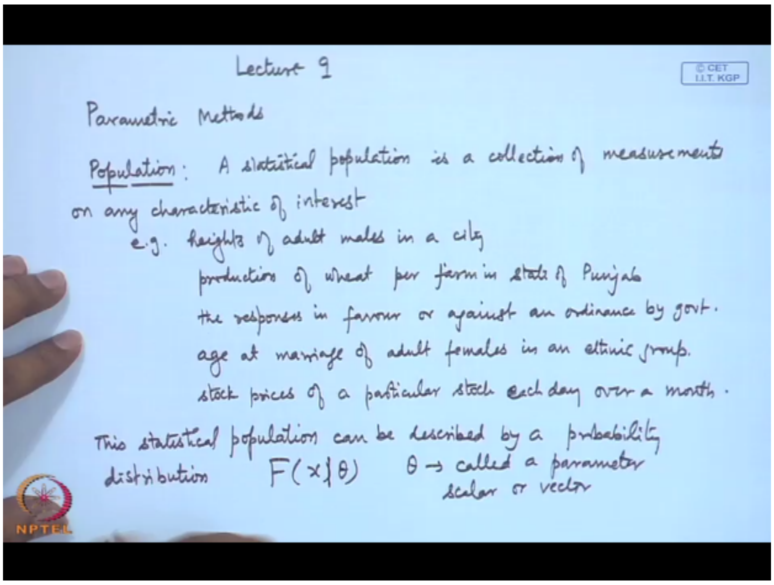
Today, we will start our second module of this course on Parametric methods. I will firstly introduce the problem of statistical inference. Broadly speaking, if we consider our day to day terminology of making scientific statements then they are like that in various areas. For example, in the area of agriculture, we talk about what is our per hectare production of wheat say in the state of Punjab. Is it more than per hectare production in the state of say Tamil Nadu.

We make a statement what is the average size of the holding for Indian farmers. In the field of atmospheric sciences, we make a statements like what could be the what is the expected rainfall during the next monsoon season in Indian peninsula. We talk about in atmospheric sciences, if there is a cyclone approaching what would be the average wind speed during the peak of the cyclone.

In the area of medicine, suppose there is a treatment which is discovered for treating a particular kind of disease then we will be interested in knowing the average effectiveness of the medicine that means out of a n number of people, how many people will be effectively cured from that disease by taking that particular line of treatment. Or if there is another previously known medicine, which is already available, then whether this new medicine will be more effective than the previous one or less effective or equally effective.

Whether it is more costly treatment than the previous one. Statements of this nature around in every area of human activity be it economics, be it social science, be it industry, be it trade, be it physics and so on. Now, statisticians treat this problem as a problem of statistical inference, we broadly classify it into 3 categories. We assume so for example we consider the concept of population.

(Refer Slide Time: 03:02)



So we firstly talk about a statistical population is a collection of measurements on any characteristic of interest. For example, it could be heights of adult males in a city, it could be production of say wheat per farm in state of say Punjab. It could be say for example the responses in favor or against an ordinance by government. It could be say age at marriage of adult females in an ethnic group.

In trade it could be say the stock prices of a particular stock each day over a month. So, all of these are examples of a statistical populations. As I mentioned little while ago, a typical problem of a statistical inference could be to estimate average heights of adult males in a city, it could be to estimate the average production of wheat, it could be for example the proportion of people who favor a particular ordinance by government.

It could be the estimation of the average age at marriage of adult females in an ethnic group, it could be stock prices that means estimating say average stock price of a particular stock. So to answer these questions we are dealing with the populations of the measurements on this. Now, there are 2 ways of looking at this, we may assume a distributional model for these measurements and of course there are methods to determine that what could be an appropriate distribution for that.

For example, heights of adult males in a city could follow a normal distribution. The responses in favor are against may follow a binomial distribution. Age at marriage may be following say a gamma distribution and so on. There are methods of fitting distributions right now we are not going to discuss that. But once we have a fitted a model, then we can say that, this statistical population can be described by a probability distribution.

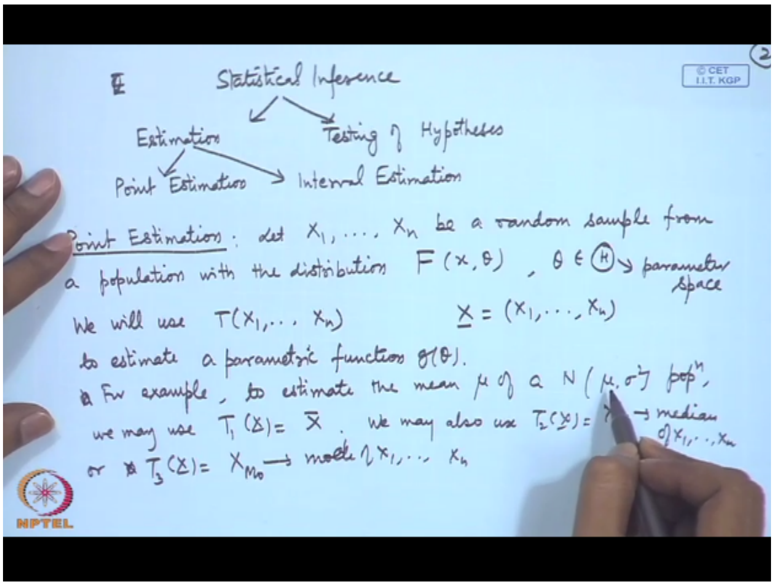
Say capital Fx. Now here we may have option. As I mentioned, if we assume a model like normal distribution, binomial distribution, exponential distribution etc. then this F is actually specified but the parameter of the distribution may be unknown and in that case we will say the distribution is Fx theta, that means it is correct traced by a parameter theta. This is called a parameter which could be of course scalar or vector.

For example, if we are writing say binomial n p distribution, then here the parameter could be p or it could be n p, if the total number of trials is fixed here then n could be known and then the parameter is p otherwise it is n and p. Suppose I consider exponential distribution with parameter lambda, then your parameter is, suppose I say normal mu sigma square, then my parameter is mu and sigma square. That means by parameter we refer to the characteristics of the population.

In whatever way they may be defined. For example, in normal distribution mu defines mean and sigma square denotes variance, whereas in a binomial distribution n denotes the total number of trials and p denotes the probability of success in each trial. In an exponential distribution, lambda is actually the reciprocal of the average, 1 / lambda will be the mean. The problem of parametric inference is to make certain statements about the unknown parameter of the population.

For example, I want to estimate the average height of adult males, so this brings us to the one area of statistical inference and that is called estimation.

(Refer Slide Time: 09:44)



So, broadly speaking let us write the problem of statistical inference, I divide into various parts so one is estimation, where I want to know the value of the parameter through some method. The other problem is that of making a doing a testing of hypothesis. For example, I would like to check whether the heights of adult males in a city is different from the height of adult males in another city or is it less or is it more.

If you want to do a testing, then the statement has to be in the form of a assertion and then an alternative to that will be given whether it is not true or so. So, broadly speaking we subdivide the problem of a statistical inference into estimation and testing of hypothesis. Later on I will also discuss the case when there is no theta here. That means the distributional model itself may not be known and that is called the problem of non parametric methods and are non parametric inference.

We will discuss that in another module of this particular course. Right now we are assuming that our population is partially specified in the sense that capital F is specified but the parameters may not be specified. So if the parameters are not specified and we want to make certain inference about that, then we are dealing with the parametric methods as I mentioned here. Now let us consider the problem of estimation.

In the problem of estimation, we will give on the basis of a random sample so because we do not have the full population with us so to make the statistical interference a sample from that population is taken and then the sample is used to draw any particular inference. It could be in the form of estimator. Now an estimator can be of 2 types, it can be a point estimator or it can be an interval.

For example, we may say the average height of adult males is say 6 feet, then we are giving a single value for the average height. Then this is called a point estimator. On the other hand, we may say that 95 % of the times the average height would lie between 5 feet 11 inches and 6 feet 1 inch. Then we are giving an interval of the values and at the same time we are associating a probabilistic statement along with that.

That means roughly how many times the statement was likely to be true. This is known as a confidence interval or interval estimation. So we subdivide the problem of estimation into the problem of point estimation and the problem of interval estimation. To begin with I consider the problem of point estimation. So let us consider let X1, X2, Xn be a random sample from a population with the distribution Fx theta.

As I mentioned theta could be scalar or vector and it lies in a space of values which is called parameter space. We will use T X1, X2, Xn. Let us use a abbreviation X vector denotes X1, X2, Xn to estimate a parametric function say g theta. Now the question comes that, what is the nature of g theta and what is the nature of T. Let us consider say binomial n p. we may be interested in estimating the probability of success. For example, if we consider responses in favor of an ordinance.

So we have taken a sample of size n and we record the responses which are in favor. So for example that number is capital X. So now the sample proportion is capital X/n, which can be used as an estimator for the probability of favorable response. So X is a collection of the responses which are in, so total number of responses are in terms of yes you can say 1 and if it is no, you can say 0 and if you add them you get capital X there.

So it is function of the observations and we are using it to estimate the proportion here. In a similar way, one may consider suppose I take the problem of normal mu sigma square and suppose I am considering the heights of adult males, I assume that they follow a normal mu sigma square distribution. Now the problem is to estimate mu. So I take a sample of adult males and I record their heights. So let me call them X1, X2, Xn.

I can use the sample mean X bar for estimating mu. Now the question arises that what is the methodology by which I can assign the estimators, what are the criteria for that. So let us look at this problem. So for example, to estimate the mean mu of a normal mu sigma square population, we may use, let me call it T 1 X as X bar. We may also use say T2X, which is nothing but the X median. That is the median of observations.

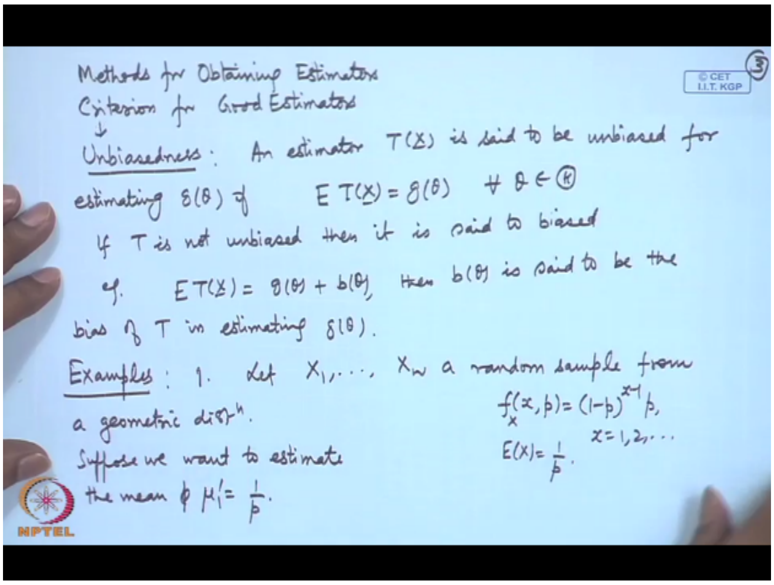
If one looks at the logic that mu also denotes the median of a normal distribution. In that case, one may consider the sample median or we may use say T 3, we may call it to be M mode, that means the mode of X1, X2, Xn. If I give the interpretation to mu that it is the point which is the mode of the density function. Now, the question comes that in a given problem, out of T1, T2 and T3, these may take different values and therefore which one should be used. I can give another analogy here.

Suppose I consider sigma square, now sigma square is the variance here. Now for variance, one may consider say sigma Xi-X bar whole square, the way we define variance 1/n days. So this could be my say U1X. but one may consider some other options also n place of this, somebody may consider the mean deviation about say mean. One may consider say mean deviation about median.

Once again, if I have choice of estimators, there are various estimators which are available, then what should be our method to analyze them. There is another problem. These are written in some sort of heuristic way. For example, if mu is the mean of the distribution, I consider sample mean. If I give the interpretation mu is the median, I take the sample median. If I take the interpretation mu is the mode, then I take the sample mode.

But, there may be some other parameters or parametric functions for which this direct interpretation may not be available. And in that case what should be our method of getting the estimators. So there are 2 problems.

(Refer Slide Time: 19:21)



One problem is method for obtaining estimators and another is the criteria for good estimators. So we will take up both of these topics. Let me firstly consider some criteria for good estimators and then when we obtain by some methodology, then we will check whether those criteria are satisfied or not. Now, one may argue in this way that if I consider an estimator. Now, this is based on the sample.

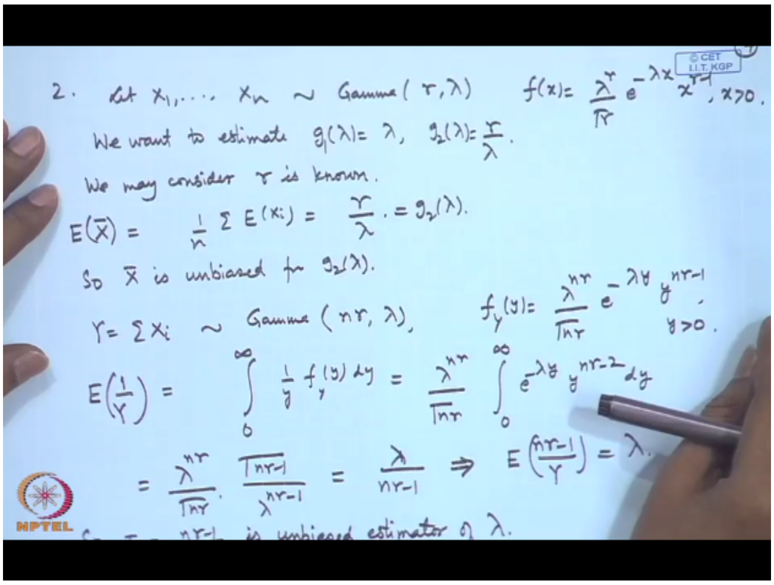
Now, if I take the sample at another point of time, or at another point of time or some other person also takes a sample and gets an estimate, then on the average, it should be same as the original value of the parameter. Now, this can be modeled in statistical terms as the criteria of unbiasedness. So, an estimator TX is said to be unbiased for estimating g theta. If expectation of TX = g theta for all theta.

That means on the average, this T, that is expected value of T is same as the original value. If T is not unbiased then, it is said to be biased. For example, if expectation of TX = g theta + b theta, then b theta is said to be the bias of T in estimating g theta. Let us consider some example here. Let us consider say X1, X2, Xn a random sample from a geometric distribution.

So, by geometric distribution we have introduced in the first module that we will consider the probability mass function as fxp = 1-p to the power x-1p for x = 1, 2 and so on. If this was a distribution, the mean was 1/p. suppose we want to estimate the mean that is let me call it mu 1 prime that = 1/p. then I can consider say TX = X bar. Then expectation of TX will be = expectation of X bar that = 1/n expectation of sigma Xi that = 1/n sigma expectation of Xi.

Now each Xi will have mean 1/p. So this is becoming n/n1/p that = 1/p. therefore this is unbiased estimator for estimating the mean of the geometric distribution. Let us take some more example.

(Refer Slide Time: 24:16)



Let X1, X2, Xn follow a gamma distribution with parameter say r and lambda. That means I am considering the density function to be lambda to the power r/gamma or e to the power -lambda x x to the power r-1, for x > 0. Now, suppose I want to estimate lambda here, let me call it g1 lambda. Suppose this is lambda. Suppose I also want to estimate g2 lambda, actually the mean of this distribution will be r/lambda. So suppose I want to estimate r/lambda also okay.

We may consider here say r is known. If we consider that and let us consider say X bar. So expectation of X bar that = again if I apply the same argument, it will be = 1/n sigma expectation of Xi. and each Xi has mean r/lambda. So this will become r/lambda, that = g2 lambda. So, X bar is unbiased for g 2 lambda. Now, to consider the estimation of lambda, let us consider in a slightly different way. Let me consider for example Y = sigma Xi, then the distribution of Y that will be gamma.

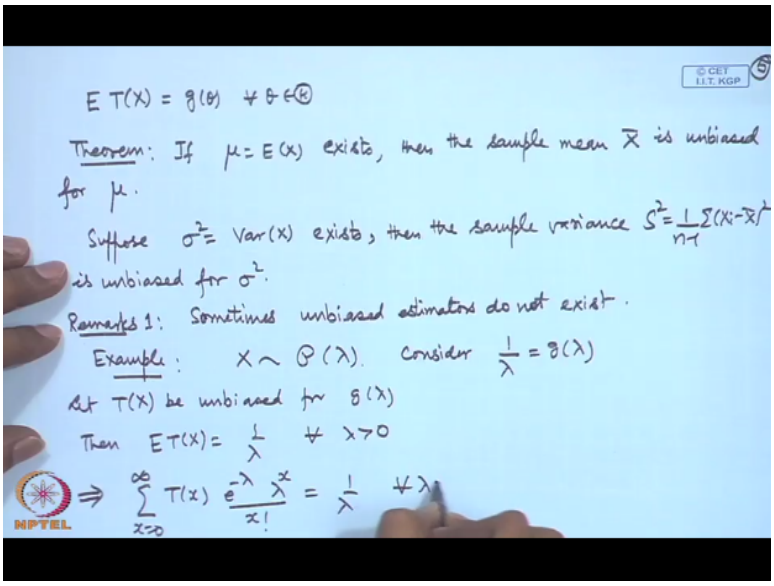
In the previous module, I have mentioned the gamma distribution if the scale parameter is kept fixed, then it is additive. Therefore, this will become gamma nr lambda, that means if I write down the distribution of Y, that will be lambda to the power nr/gamma nr e to the power

- lambda y, y to the power nr - 1, for y positive. If I consider say expectation of 1/Y, then it = 1/y fydy 0 to infinity.

Now, that is equal to lambda to the power nr by gamma nr e to the power - lambda y, y to the power nr-2dy. So that = lambda to the power nr by gamma nr, gamma nr-1/lambda to the power nr-1. So that gives me 1/nr-1 lambda. Since I am assuming r to be known here, I can adjust this coefficient, that means I will get expectation of nr-1/y = lambda. So this gives say T2 that = nr-1/sigma Xi is unbiased estimator of lambda.

So the problem of unbiased estimation can be solved by suitably choosing the functions. However, this method is more heuristic in nature, a more general form could be to actually consider.

(Refer Slide Time: 28:28)



That means you need to basically write down an equation of the type expectation of TX = g theta for all theta and solve this equation. However, you have some general result which you have seen. For example, I considered geometric distribution and as well as I considered the gamma distribution. If I want to estimate the mean of the distribution, I can take the sample mean and it is unbiased.

So, if the moment exists, that means the mean is defined, then we can actually say that, the sample mean is always unbiased for the population mean. So we can consider, if mu that = expectation of X exists, then the sample mean X bar is unbiased for. In fact, we may prove a little more general result also, we may also consider suppose sigma square = variance of X, this exists, then we define something called the sample variance.

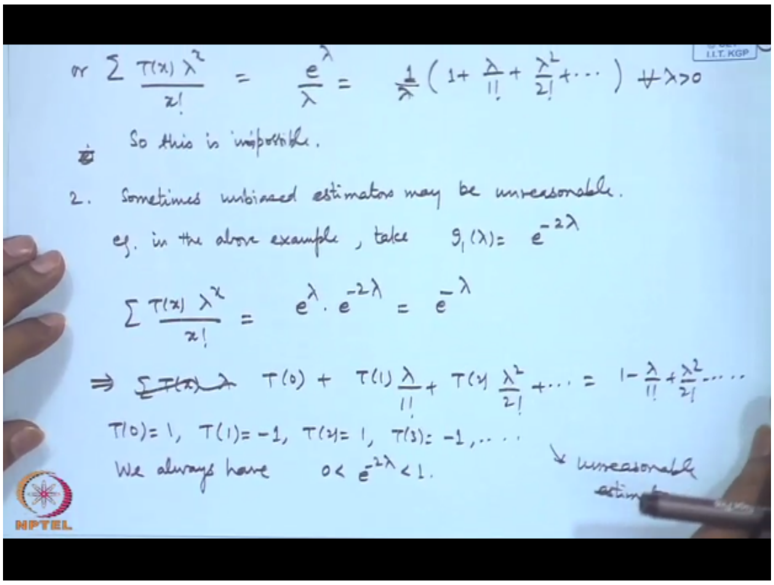
Then the sample variance, I call it S square that = 1/n-1 sigma Xi - X bar square. Then this is unbiased for sigma square. These results are true irrespective of any distribution. The only condition is the existence of mu and sigma square here respectively. So these results are pretty useful and they are used to derive heuristic estimators. For example, in this case of gamma distribution, I could easily derive the estimator of r/lambda.

However, now, to get an estimator lambda I considered an improvisation because, if I am considering a function of sigma Xi for 1/lambda, then it will be reversed for that and that is why I considered 1/Y here or 1/sigma Xi here. And here, some sort of distribution theory is used here, that means the sum of the gamma distributions. Similarly, in the case of geometric distributions, suppose I want to estimate the probability p here, probability of success in individual trial, then I will have to consider 1/sigma Xi here.

Of course 1/sigma Xi will follow negative binomial distribution. So I can use the property of that negative binomial distribution and I can construct in a similar way. So although unbiased estimators are heuristically having a nice justification that on the average, the estimated value will be equal the true value of the parameter. But there may be sometimes the problems. So, you may have for example, sometimes, unbiased estimators do not exist.

To take a very simple example, let us take say X following Poisson lambda distribution. Now, consider 1/lambda as g lambda. Let T be unbiased, for g lambda. Then we should have expectation of T = 1/lambda, for all lambda. Now this will imply sigma Tx e to the power - lambda, lambda to the power x/x factorial that = 1/lambda, for all lambda positive.

(Refer Slide Time: 33:04)



We can further rewrite this condition as sigma Tx lambda to the power x/x factorial = e to the power lambda/lambda and this right hand side I expand, it is 1 by lambda 1+lambda/1 factorial + lambda square by 2 factorial and so on. So what we are saying is, this is for all lambda > 0. If you look at this term carefully, the left hand side is a power series in lambda, the right hand side is a power series in lambda but there is also a term 1/lambda here.

Now, so basically you have a Laurent series also coming here. So the 2 series cannot agree on an interval, because if they have to agree on an interval, all the coefficients must match. So, this is impossible. That means the unbiased estimator of 1/lambda does not exist. A second thing is that sometimes, unbiased estimators may be unreasonable. For example, let us consider say in the same in the above example, take say g1 lambda = say e to the power - say 2 lambda.

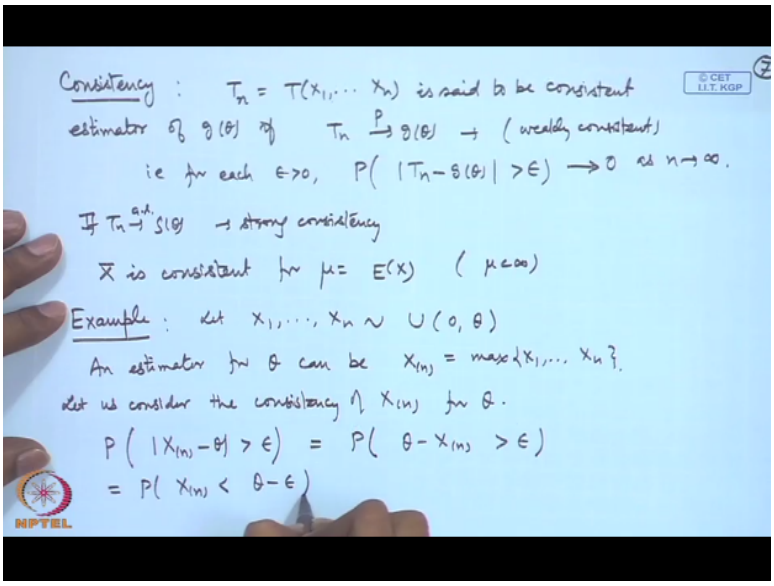
Now, let us look at this condition here. Sigma Tx lambda to the power x/x factorial = e to the power lambda into e to the power - 2 lambda. That = e to the power-lambda. So, let us write down the terms here, so sigma Tx lambda, so this will give me T0+T1 lambda/1 factorial+T2 lambda square/2 factorial and so on. That = 1-lambda by 1 factorial + lambda square by 2 factorial and so on. Now let us look at here. This gives me T0 = 1, T1 = -1, T2 = +1, T3 = again -1 and so on.

Now look at the problem, this g1 lambda function, it is e to the power -2 lambda, since lambda is positive, we always have 0 < e to the power -2 lambda < 1, this is an unbiased estimator now. But it is taking values always +1 or -1. So this is unreasonable estimator. That means to estimate a parametric function which is lying between 0 and 1, I use the values +1 and -1, because depending upon what is the observation x, you will either use it as 1 or you will use as -1.

So this is not a good criteria or good estimator here. So, now one thing I would like to mention right here, see we have introduced a criteria of unbiasedness. So it is based on the reasonable assumption that if we consider the sampling a large number of times, then on the average the estimated value should be equal to the true parameter value.

So this the only one criteria we have also seen that sometimes we may not have done by estimator or sometimes even if it exists it may not be reasonable. Therefore, we have some other criteria also. Let us look at one or two some such things.

(Refer Slide Time: 37:37)



Another criteria is that of consistency. Now, we are considering T, let me write the estimator as Tn because it is dependent upon n observations. So this is said to be consistent estimator of g theta, if Tn converges to g theta in probability. Now, if you remember in the previous lecture, we have introduced the concept of convergence and probability. That is for each epsilon > 0, probability of modulus Tn-g theta > epsilon, this goes to 0 as n tends to infinity.

So if we give the physical interpretation of this condition, it means that as the sample size increases, that means if the estimator is based on a larger sample, then the probability of this being closer to the true value increases. Because the probability of this being away from the true value is decreasing to 0. So, consistency is again a large sample property. That means if we are taking more and more observations, then we are approaching towards the true value at least in the sense of probability.

Now, if we remember the strong law of large numbers and the weak law of large numbers, in the weak law of large numbers we assume that if the mean of the observations is mu, then the sample mean is consistent for the population mean. So, in fact we also saw the strong law of large numbers, so we call it then weak consistency and the if Tn converges to g theta almost surely, then it is called strong consistency. This is we can say weakly consistent, just to keep in analogy with the weak law and the strong law of large numbers.

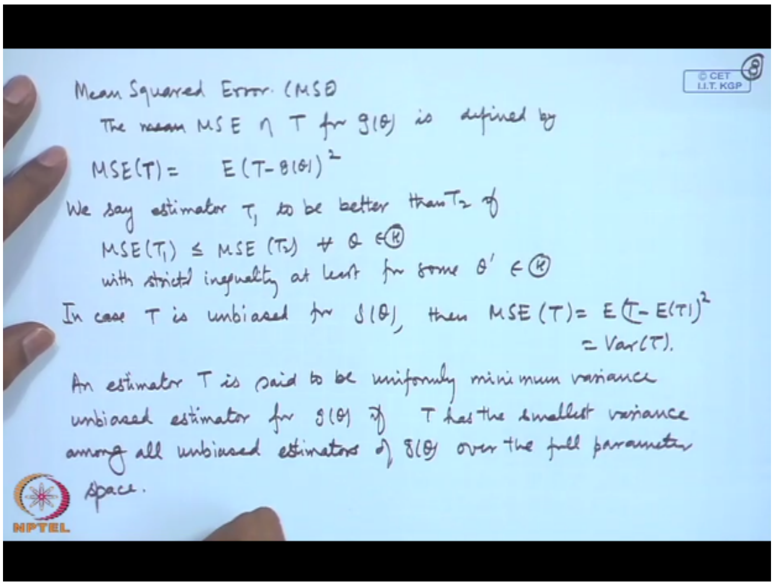
So, X bar is consistent for mu, that is the mean of the population. Of course we have to assume that mu should exist. Sometimes we may be able to apply this weak law of large numbers or strong law of large numbers, then one may try directly. Let me take an example here, which is different from the consistency of the sample mean. Let us consider say, X1, X2, Xn following uniform distribution on the interval 0 to theta.

Now, if we are considering the uniform distribution on the interval 0 to theta, all the observations are lying between 0 to theta. An estimator for theta can be say Xn, that is the maximum of X1, X2, Xn. let us consider the distribution of Xn, the consistency of Xn for theta. So let us consider probability of modulus Xn-theta > epsilon. Since all the observations are between 0 to theta, theta is bigger than Xn.

So, this probability is same as theta-Xn > epsilon. Which = probability of Xn < than theta-epsilon. Now this is equivalent to probability of each of X1 < theta-epsilon and so on Xn less than theta-epsilon. Because if the maximum of the observations is < theta-epsilon, then each observation will be less than theta-epsilon. So, because of the independence, it simply becomes, the Fx at theta-epsilon to the power n, that = theta-epsilon/theta to the power n.

Naturally this goes to 0 as n tends to infinity. So that means, this Xn is consistent for theta. So, in this case, we have not used the law of large numbers, rather we have gone for a direct verification of the result. Now, we may also have a situation, where there are 2 estimators, say both may be unbiased, both may be consistent, or may be biased and one may be unbiased and then, how to compare. So, we introduce the mean squared error criteria.

(Refer Slide Time: 43:39)

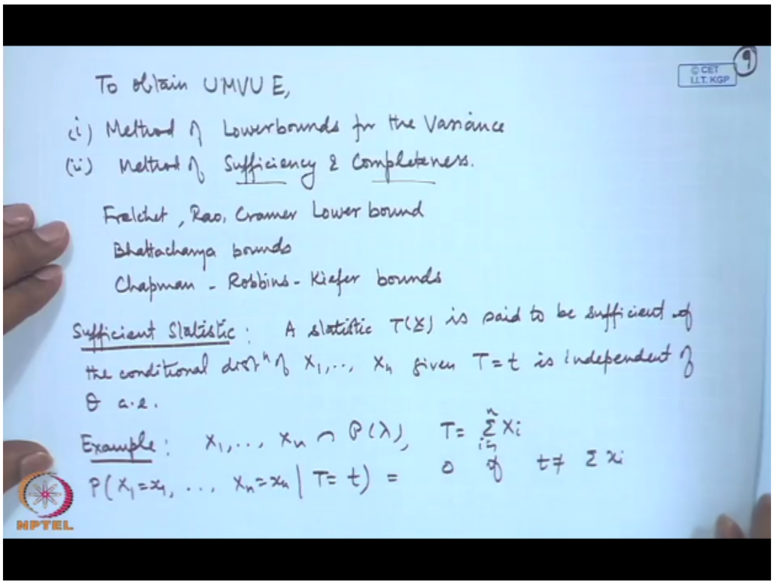


So, the mean squared error, that is we call it MSE. MSE of T for g theta is defined to be say MSE T, that = expectation of T-g theta whole square. And in the terms of this, we can give the comparison, because the smaller the mean squared error, the better the estimator can be. So, we say estimator T1 to be better in the sense of mean squared error than T2, if mean squared error of T1 <= mean squared error of T2, for all theta with a strict in equality at least for some theta prime belonging to theta.

Now in case, T is unbiased for g theta, then the mean squared error of T is becoming expectation of T-expectation T whole square. That is nothing but the variance of T. So, an estimator T is said to be uniformly minimum variance unbiased estimator for g theta, if T has the smallest variance among all unbiased estimators of g theta over the fill parameter space.

Now the question is that, how to determine the minimum variance and by estimators or how to obtain the estimator which have less mean squared error. For that there are certain other techniques.

(Refer Slide Time: 47:03)



For example, to obtain uniformly minimum variance unbiased estimator, we have method of lower bounds for the variance, then second method is the method of sufficiency and completeness. We will not be discussing in detail these methods in this particular course as they have been discussed in another course on statistical inference, which is also available on NPTEL. Here I will briefly mention about the concept of sufficiency and completeness, regarding the lower bound I will mention one lower bound here.

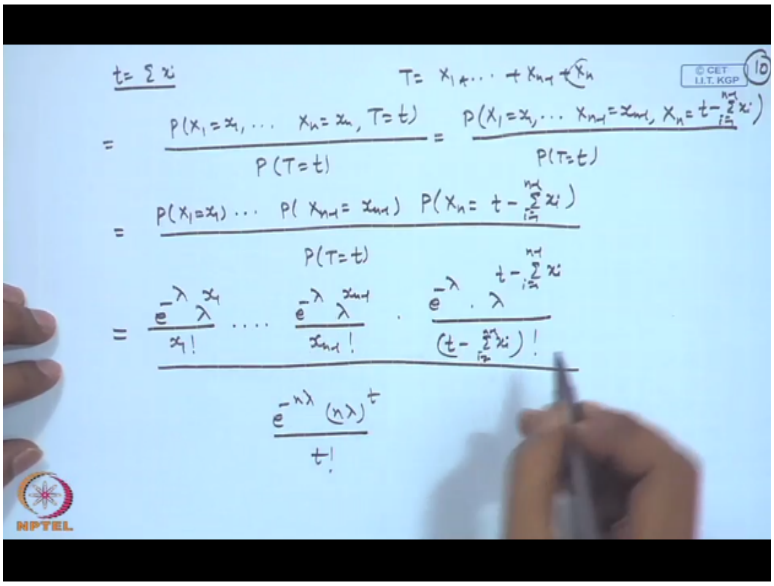
So, but this telling various conditions on the density function. So, that becomes a bit theoretical. Since, this is a course on statistical methodology. I will just mention the method here. So, in the method of the lower bounds for the variance what we do if I say variance if I say that T is an unbiased estimator for g theta, then variance of T should be always >= certain number. Now, if that is so then, if I am able to obtain an estimated T which is having variance = to that bound, then certainly it will be minimum variance unbiased estimator.

So, there are methods like we have Frechet, Rao, Cramer lower bound then we have Bhattacharyya bounds, we also have Chapman-Robbins, Kiefer bounds. For detailed discussion about these bounds, you may look at the lectures on statistical inference. Let me briefly mention about the concept of sufficiency and completeness here.

So, sufficient statistic, so we have the same setup that we have a random sample X1, X2, Xn from a population with distribution FX theta, so a statistic TX is said to be sufficient, if the conditional distribution of X1, X2, Xn given T = say small t is independent of theta, almost everywhere. Now, to give a simple example, suppose I consider say X1, X2, Xn follow Poisson lambda distribution. I consider T = sigma Xi i = 1 to n.

Then if I considered the conditional distribution of X1, X2, Xn given T = t, then certainly this = 0, if t is not = to sigma Xi.

(Refer Slide Time: 51:16)



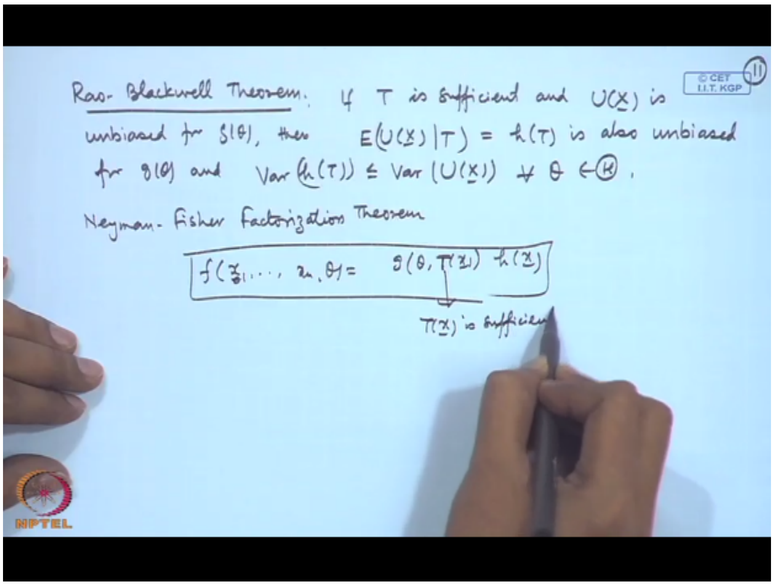
If we consider t = sigma Xi, then this probability = probability of X1 = x1 and so on. Xn = xn, T = t divided by probability of T = t. This can be then further simplified as probability of X1 = x1 and so on Xn-1 = xn-1. Now this T is X1+X2+Xn-1+Xn. So, if the value of T is fixed, then the value of Xn is also fixed. So we can write this as Xn = t-sigma Xi, i = 1 to n-1 divided by probability of T = t, that is equal to now we can use the independence here.

So, probability of X1 = x1 and so on probability of Xn-1 = xn-1 probability of Xn = t-sigma Xi i = 1 to n-1 divided by probability of T = t. Now, we also have seen the additive property of the Poisson distribution. If X1, X2, Xn are independent Poisson lambda, then sigma Xi will follow Poisson n lambda distribution.

So, we can make use of this fact here in the calculation and this one then becomes e to the power-lambda, lambda to the power X1/X1 factorial and so on e to the power-lambda, lambda to the power Xn-1/Xn-1 factorial e to the power-lambda, lambda to the power t-sigma Xi i = 1 to n-1/t-sigma Xi i = 1 to n-1 factorial divided by now this capital T follows Poisson n lambda. So this becomes e to the power-n lambda n lambda to the power t/t factorial.

So easily you can see that lambda is simply canceling out and we are getting here t factorial divided by X1 factorial and so on Xn-1 factorial*t-sigma Xi factorial i = 1 to n-1. So, this is independent of lambda. So T that = sigma Xi is sufficient statistic here. Now, we have a very strong result in a given decision problem or in a given estimation problem if the sufficient statistics exist, we can make use of this to create better estimators and ultimately it will lead to minimum variance and estimators.

(Refer Slide Time: 54:34)



So we have the result called Rao Blackwell theorem. Say if T is sufficient and say UX is unbiased for g theta, then expectation of UX given T, let me call it say hT is also unbiased for g theta and variance of hT will be <= variance of UX for all theta. We will show that by combining this concept with another concept of completeness, we can actually get the minimum variance and by the estimator in estimation problems where the unbiased estimators exist.

So, before that we also mention another result, this method of proving the sufficiency is somewhat complicated in this particular case, I was able to already guess an estimator or guess a function which I could prove that it is sufficient. But in general problems, it may not be so easy and another thing is, this involved the computation of the conditional distribution here, which again may not be easy. For example, here we were dealing with the discrete distribution and therefore, it is easy to derive the conditional distribution.

But suppose we are dealing with a continuous distribution, then this interpretation will not be there and the calculation of the conditional density may involve lot of algebraic calculations.

So, there is another result called Neyman-Fisher factorization theorem which involves writing down the joint distribution of the observations that is f X1, X2, Xn as a product of two terms. One is g theta say TX * another term called hX.

If this factorization is available, then that means this term is involving the parameter and the statistic T and the second term is free from theta, then we say TX is sufficient. Now, this is necessary and sufficient condition of course under certain conditions and therefore this can be easily used for calculation of the or obtaining sufficient statistics in a given problem. In the following lecture, I will also introduce the concept of completeness and how these can be used to derive the uniformly minimum variance and by estimators.

I will also introduce the concept of method of movements and the maximum estimator for deriving the estimators.