Statistical Inference

Prof. Somesh Kumar

Department of Mathematics

Indian Institute of Technology, Kharagpur

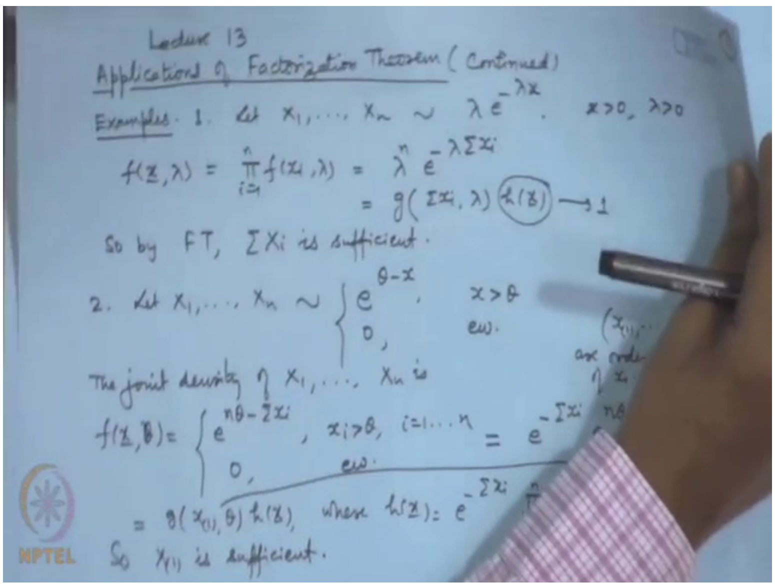
Lecture No. # 13

Sufficiency and Information

In the previous class, I have explained the concept of sufficiency. This concept is the concept which is called the principle of data relation. So, we have a random sample X 1, X 2, X n, but if we have a sufficient statistic T, then that is sufficient that gives the complete information about theparameter which is contained in the sample. So, we need not retain itwe have given one theorem, which is called factorization theorem and this is useful for deriving sufficient statistics in various probability models.

Yesterday, I have discussed the normal probability model and I have shown you that how, if we change the parameter space; that means, whether we have nu known or sigma square known or both are unknown in each of the cases, the sufficient statistics changes. So, sufficiency is the property of the probability model under consideration. Let me explain it though a few more examples andwe will use the concept ofthis factorization theorem here.

(Refer Slide Time: 01:29)



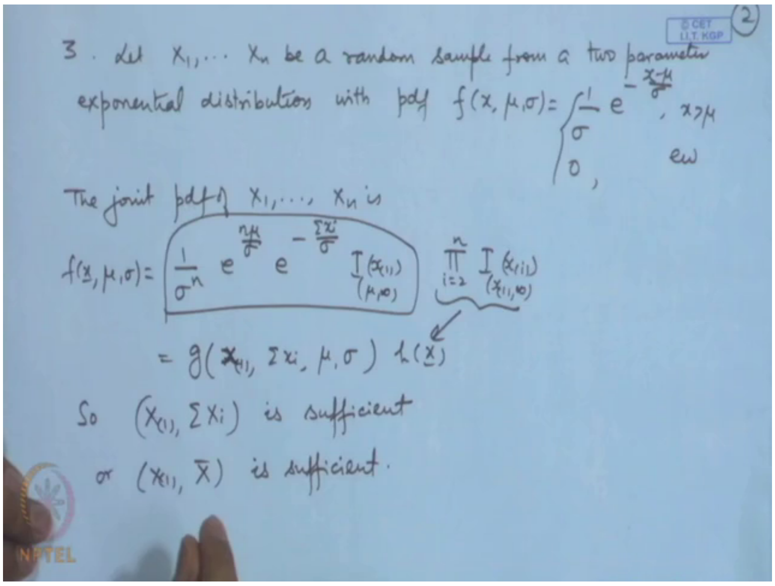
Let me start with exponential distribution, let X 1, X 2, X n follow exponential distribution say with parameter lambda. So, in the factorization theorem we need to write down the joint density of X 1, X 2, X n that is equal to lambda to the power n eto the power minus lambda sigma x i. Now, this whole thing we can write as a function of sigma x i and lambda and h x this h x, I am taking to be one itself the constant.

So, you can see by factorization theorem by factorization theorem sigma x i is sufficient, let us consider another exponential model in which in place of a scale parameter we will have a location parameter. So, let us consider say X 1, X 2, X n following exponential say, theta minus x where x is greater than theta zero elsewhere. Now, in this case the joint density of X 1, X 2, X n is f of x theta that is equal to e to the power n theta minus sigma x i; however, this description of x i greater than theta also place a role here.

Now, if we want to write it as a product here we will make use of the indicator function. So, we can write it like this e to the power minus sigma x i, e to the power n theta indicator function of the set X 1 from theta to infinity and indicator function of other x it's from 2 to n from X 1 to infinity. So, what we can consider we canwe can write it as g of X 1 theta into h of x where, h of x I, am writing as e to power sigma x i into product i is equal to 2, 2 n I of x I, X 1 to infinity.

So, here this X 1, X 2, X n they are denoting the order statistics of X 1, X 2, X n. So, g is this function this is a function of X 1 and theta. So, we conclude that X 1 is. So, X 1 is sufficient. Now, note here when we had lambda as the parameter and here we had a scale model the sufficient statistics was sigma X i although here, we are again we are dealing with the exponential distribution, but the nature of the parameter has changed therefore, the sufficient statistic is now, the minimum of the observations now, in a similar way let us take up the two parameter exponential distribution.

(Refer Slide Time: 05:18)



Let us take X 1, X 2, X n be a random sample from a two parameter exponential distribution say with density function f x, mu, sigma is equal to one by sigma e to the power minus x minus mu by sigma for x greater than mu and it is equal to 0 otherwise. So, once again the joint probability density function of X 1, X 2, X n this is now, 1 by sigma to the power n e to the power n mu by sigma e to the power minus sigma x i by sigma.

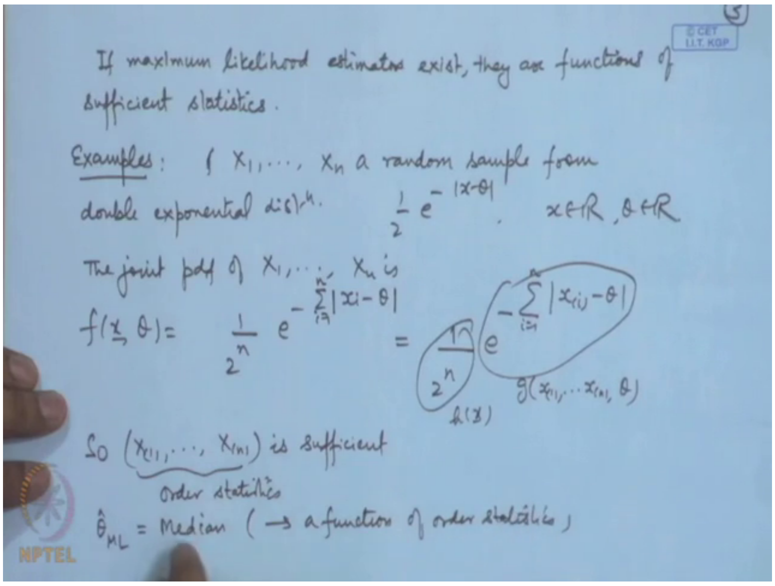
And once again the condition that each of x i greater than mu I can express in terms of the indicator function like x 1 is from mu to infinity and x i is other x i is they are from x 1 to infinity i is equal to 2to n. So, this portion I can write as g of x 1 sigma x i and mu sigma and this part is h X. So, here we concluded that x 1 and sigma x i is sufficient or we can also say x 1 and x bar because this is a one to one function of this, this is sufficient.

I also want to mention here we have earlier considered, the maximum likelihood estimators. Now, let us remember one maximum likelihood estimators for each of these problems for example, in this case the maximum likelihood estimator for lambda was one by x bar which is a function of sigma x i in this particular case.The maximumlikelihood estimator was x 1 that is a minimum of observations and it is sufficient here similarly, here you see the maximum likelihood estimator for mu and sigma, where x 1and x bar minus x 1respectively, which is again a one-to-one function of x 1, x bar that is the sufficient statistics.

So, we can observe that maximum likelihood estimator if it exists is actually, a function of the sufficient statistics the reason is obvious because in the factorization theorem, we are writing down the density as a function of the parameter and the sufficient statistics into a function, which is free from the parameter. Now, in the method of maximum likelihood estimator, we are maximizing the density function or the mass function with respect to the parameter.

Now, the part of the density which contains the parameter contains the variableonly through the sufficient statistics therefore, the maximization problem will give a solution in terms of the sufficient statistics alone.

(Refer Slide Time: 09:15)



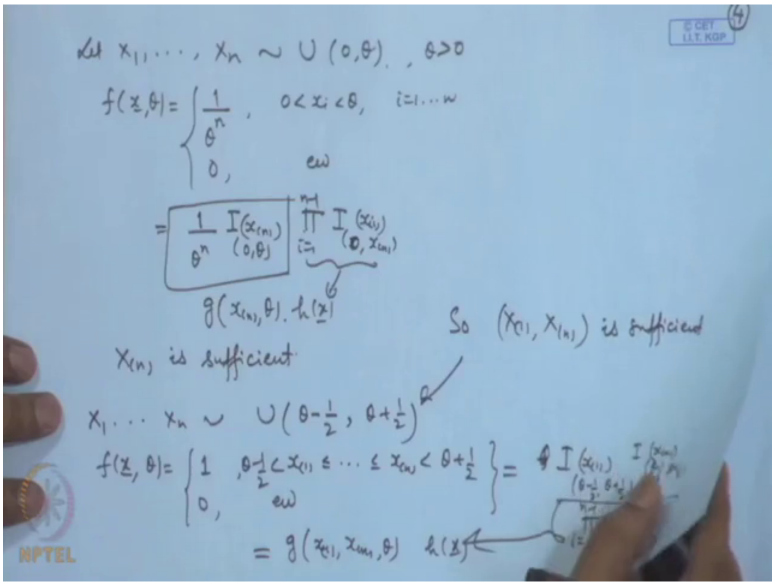
So, we have a general comment here that if maximum likelihood estimators exist, they are functions of sufficient statistics.Let us take some more examples here say for example, X 1, X 2, X n a random sample fromsay double exponential distribution half e to the power minus x minus theta, where x is any real number and theta is any real number. In this case, if we consider the sufficiency. So, the joint distribution of X 1, X 2, X n that is equal to one by two to power n e to the power minus sigma modulus x i minus theta, i is equal to one toone. Now, here you observe I cannot reduce it further as a function of parameter and a another variable here because each of the x i is appearing in themodulus sign and therefore, I cannot separate it out.

At the most I can consider the reduction as one by two to the power n e to power minus sigma modulus of x i order statistics minus theta. So, this function is now a function of the order statistics and theta and this you can. Now, call h x. So, we conclude that the order statistics X 1, X 2, X n is sufficient this order statistics. Now, remember here for this problem what was the maximum likelihood estimator. The maximum likelihood estimator was median.Median of the observations and median is afunction of this is a function of order statistics because if we have a odd number of observations say x 2 m plus one then x m plus one that is the middle of the observation was the median and if we have an even number of observations that is x 2 m then any number between x m and x m plus 1 and we can actually, consider say the middle of the two that is x m plus x m plus 1 by 2 as the maximum likelihood estimator.

So, this is a function of order of statistics in this case also. So, this statement is true in general another thing which I just now, pointed out that many times.When we are writing down the density function say in this case, we are we have to incorporate the region of the variable, which is dependent upon the parameter as a part of thejoint density function because if we do not include it then we cannot derive the sufficient statistics.For example, if we have written only this part then there is no sufficient statistics here because e to the power minus sigma x i can be separately written e to the power n theta can be separately written.

However, this is not a complete description of the density unless we include the reason x i greater than theta for all I and this is the way of including this a similar phenomena is observed in the uniform distributions also like in the uniform distribution.

(Refer Slide Time: 13:16)

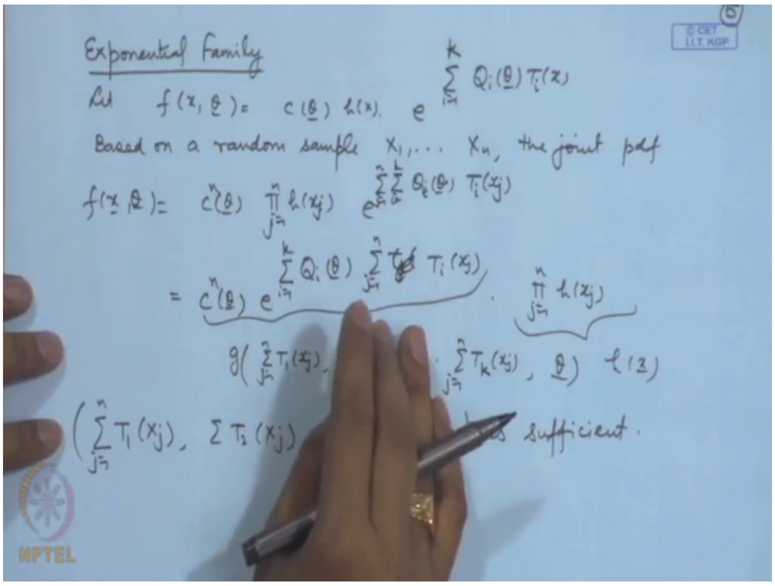


The range is dependent upon the range of variable is dependent on the parameter. So, let us consider say X1, X 2, X n say a random sample from uniform zero theta distribution. Now, in this case the joint density is equal to 1 by theta to the power nand once again each of thex i is between 1 to n and it is equal to 0 elsewhere. So, this part we will express as then one by theta to the power n I of say X n from 0 to theta and the other x i isare from 0 to X n for I is equal to 1to n minus 1. So, you can see this portion you can express as g of X n theta and this part you can write as h x. So, X n is sufficient; however, if we consider say a uniform distribution which ison a two sided interval here we have taken one side as 0.

Suppose, we consider say from theta minus say 1 by 2, 2 theta plus 1 by 2 in this case the joint distribution is simply 1 because theta plus half minus theta minus half is 1; however, each of the x i's they are between theta plus half and theta minus half. So, this part then you can incorporate as indicator function of X 1 from theta minus half to theta plus half and the indicator function of X n from theta minus half to theta plus half. And the remainingorder statistics lying between X 1 and X n I is equal to 2, 2 n minus 1. So, this you can see it is a function of X 1, X n theta into h x this part is h x and this part is a function X 1, X n and theta. So, here X 1, X n is sufficient although the parameter remains one dimensional here, but the order statistics contains two terms if you remember the maximum likelihood estimator.

The maximum likelihood estimator for this problem was any value between X n minus half to X 1 plus half. So, which is a functionof X 1, X n. So, the statement that the maximum likelihood estimators, if they exist they are functions of the sufficient statistics is satisfied here also now this factorization theorem is very useful.

(Refer Slide Time: 17:18)

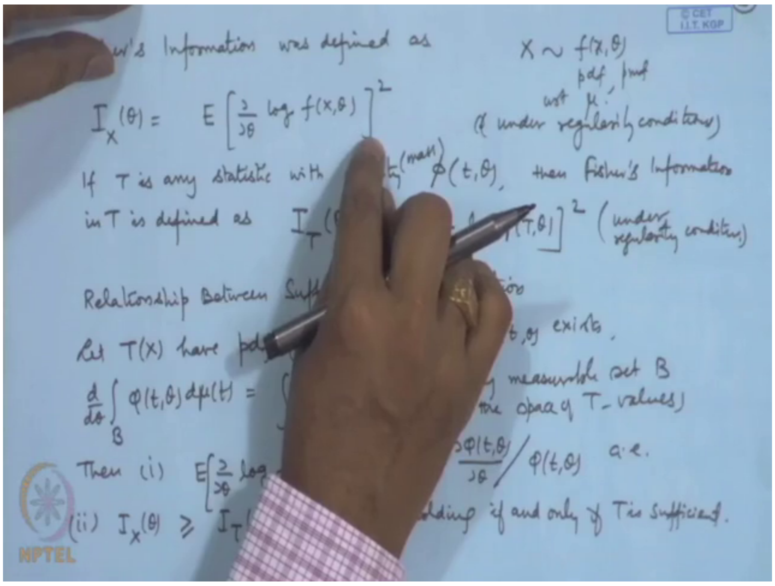


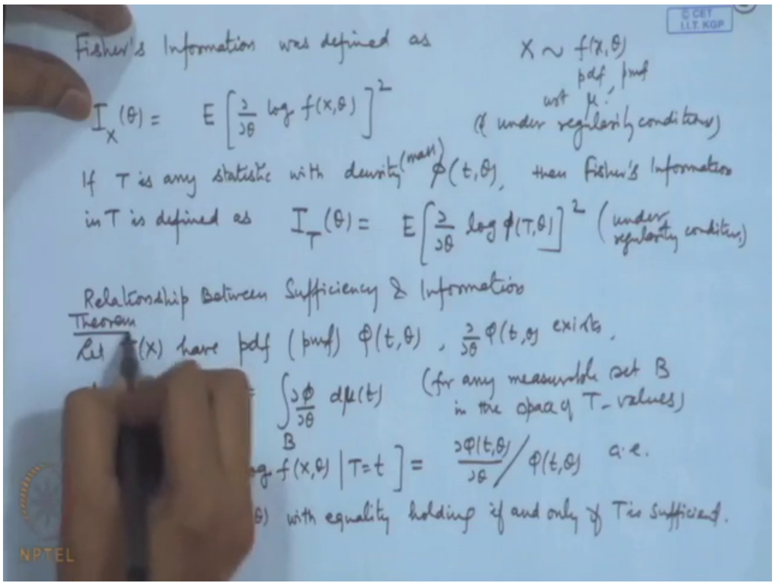
If we are considering a general distribution in an exponential family so, let us consider distributions in exponential family. So, if we are considering a k dimensional exponential family let f x theta be equal to c theta h x e to the power sigma say Qi theta T x i is equal to say 1 to k thisis called a k dimensional exponential family, provide the parameter space contains a k dimensional rectangle. So, based on a random sample X 1, X 2, X n the joint probability density function we can write as c theta to the power n product of h x i e to the power sigma Q i theta into T of. So, let me change here I j because I j is being used here.

I is equal to 1to k t, x j sigma j is equal to 1to n.Now, this I can write as c to the power n thetae to the power sigma Q i theta sigma T j t ofthere is a mistake it is T i x. So, T I, x j. So, this becomes what I am done doing is I am taking this summation inside. So, this becomes sigma of T i, X j, j is equal to 1to n I is equal to 1 to k into product of h x j, j is equal to 1to n. So, this part is now a function of sigma T 1, x j, j is equal to 1to n sigma T 2, x j, j is equal to 1 to n and.

So, on sigma T k, x j, j is equal to 1 to n and the parameter and this part we can consider as h x. So, we conclude that sigma of T 1, x j sigma of T 2, x j and so on.Sigma of T k, x j, j is equal to 1to n this is sufficient of course, when we write like this we assume that this q 1, q 2 etcetera are linearly, independent otherwise some of the terms can be merged together. Now, let me introduce the relationship between the Fisher's information measure and the concept of sufficiency.

(Refer Slide Time: 20:55)





So, if you remember the Fisher's information was defined as Fisher'sinformation was defined as I x theta is equal to expectation of del by del theta log of f x theta whole square, here the assumption is that the distribution of x is f x theta and of course, this could be p d f or p m f with respect to a measure mu andwe are making the assumption of regulatory conditions that is differentiation under the integral sign is allowed.

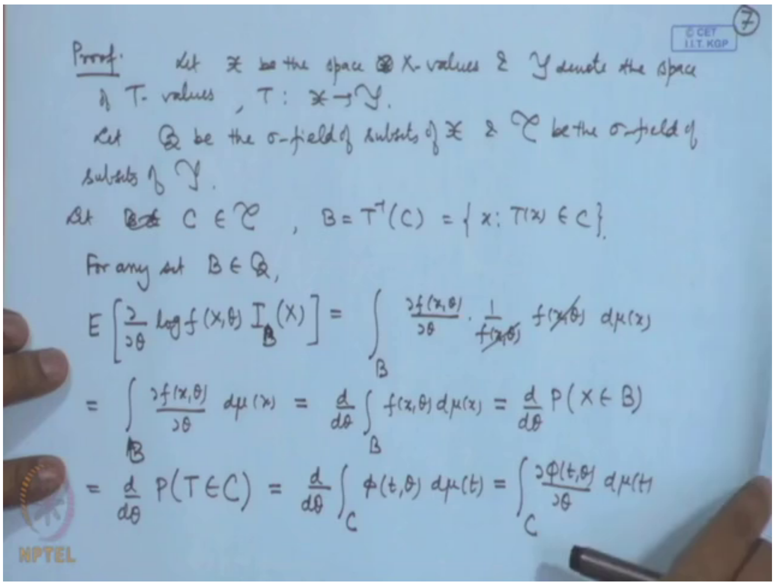
So, this is under regulatory conditions if the distribution of X is f x theta then the information measure Fisher's information in X about theta is defined asexpectation of del by del theta log of f x theta is square. Now, if t is any statistic and suppose the density of let me give the name as phi T theta,thenFisher's information in T is defined as I T theta is equal to expectation of del by del theta log of phi T theta square, once again we are making assumption about.

So, this could be p d f also or p m f and we should have the regularityconditions satisfied for phi also means, we should be able to differentiate the density with respect to that parameter, we should be able to differentiate under the integral sign. So, here also under regularity conditions we satisfy it under regularity conditions. So, we have the following result regarding relationship between sufficiency and information.Let T x have p d f or p m f phi t theta del by del theta phi t thetaexists d by d theta of integral phi T theta d mu T over an imageset b is equal to del phi by del theta d mu T for any measurable set b that is in the space of T values.

Then we have the following results first is that expectation of del by del theta log of f x theta given T is equal to t this conditional expectation is equal to del phiby del theta divided by phi e theta almost everywhere. Secondly, the information in the x is always greater than or equal to the information in any statistic T with equality holding if and only if T is sufficient. So, what we are saying suppose we have X 1, X 2, X n as a sample and T is any statistic. Then in general the information content in a statistic will be less than or equal to the information contained in the full sample; however, if T is sufficient then it will be the same and this is a necessary and sufficient condition.

So, this is what I was mentioning from the, that is the utilization of the information or the content of the information in the concept of sufficiency that sufficient statistic contains all the information, which is available in the sample because we are saying I T theta will become equal to I h theta. So, this is the physical meaning of the concept of sufficiency that if, we are considering this definition as the definition of information because this I x, theta we will call information in the sample. So, what we are saying is that there is no loss of information, if we consider a sufficient statistics. So, let me prove thistheorem here.

(Refer Slide Time: 26:57)

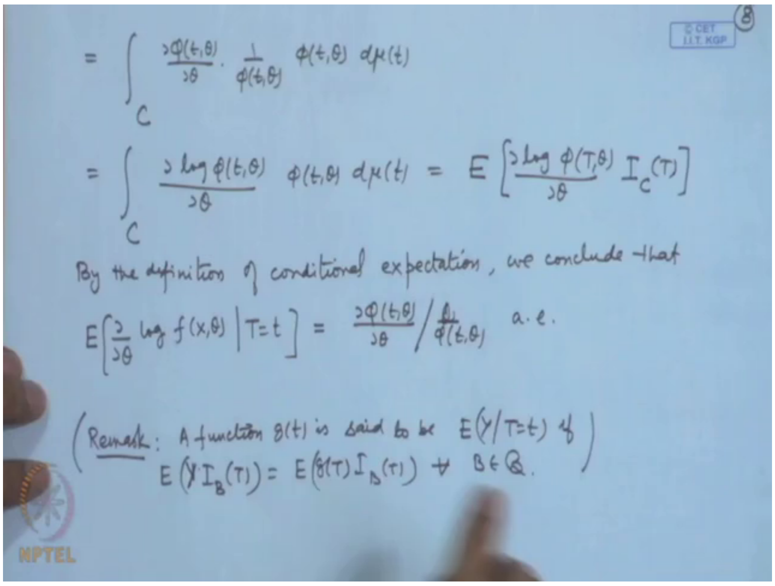


So, see we have say let x be the space of x values and say y denote the space of T values; that means, T is a function from x to say y naturally, we will be considering the sigma field's of subsets of x and similarly a sigma field of subsets of y also. So, let us use some notation say b, be the sigma field of subsets of x and say c be the sigma field of subsets of y which, we are considering here. So, now, let us consider b asay a set c in c then for that define say b is equal to t inverse c that is the set of x such that T x belongs to c. So, consider for any set say b belonging to script b, let us consider expectation of delta by delta theta log of f x theta over the setb.

So, this is equal to del f by del theta divided by f x theta. Now, this is expectation. So, it becomes f x theta d mu x over the set b. So, this f x theta and this f x theta cancels out, we are getting integral del f by del theta d mu x over a over b.Now, this we can consider because we have made the assumptionthat, we can differentiate under the integral sign. So, this equal to d by d theta of f x theta over b. Now,this the integral of the density of the random variable x over the set b. So, this is nothing, but probability of the set b.

Now, we have defined the set b to be the inverse function of or inverse image of T. So, b is the set where T x belongs to c. So, this probability of x belonging to b is same as probability of T belonging to c therefore, we can write it as d by d theta integral of phi Ttheta that is the density of a T with respect to the corresponding measure over the set c.Now, once again we have made the assumption that we can consider differentiation under theintegral sign. So, this becomes del phi by del theta d mu T over the set c now, we can divide and multiply by the density of t inside the integral sign.

(Refer Slide Time: 31:21)

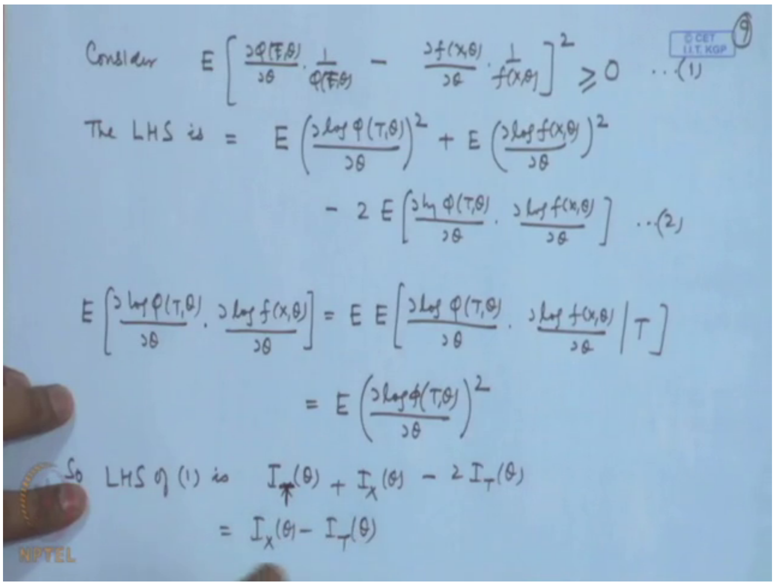


So, we will get here this term as equal to del phi by del theta 1 by phi t theta phi T theta d mu T over set c. So, now, this becomes nothing, but the derivative of log of phi t theta d mu T over the set c this is nothing, but the expectation of this expectation of del log phi T theta by del theta indicator function of the set c look at the statement that, we have proved. Now, we started with expectation of del by del theta log of f x theta I b x.

We are showing that this term is now, equal to this term is equal to expectation of del log phi t theta by del theta I c T.Now, what is the relationship between x and T and b and c t is a function of x and b is the inverse image of the set c therefore, by the definition of the conditional expectation, we conclude that by the definitionof conditional expectation we conclude that expectation of del by del theta log of f x theta given T is equal to T which is equal to del phi by del theta one by divided by phi of T theta, that is the statement given here of course, since we are obtaining this result from the expectation. So, we can say that this statement is true almost everywhere.

That means, the set we have this may not be true will have probability 0; that means, their set of values of small t for which this statement is not true then under the probability distribution of t that set will have probability 0. So, actually whatwe have used here is we have simply used the definition of the conditional expectation. In fact, let me write here remarka function g t is said to be conditional expectation of y given T if expectation of y I, b, T is equal to expectation of g t, I d, t for all b more measurable sets of b. So, we have used thisdefinition. So, what we have done is we have established a relationship in the, in the log likelihood or you can say the information content term in the density of the sufficient statistics and the original variable.

(Refer Slide Time: 35:17)

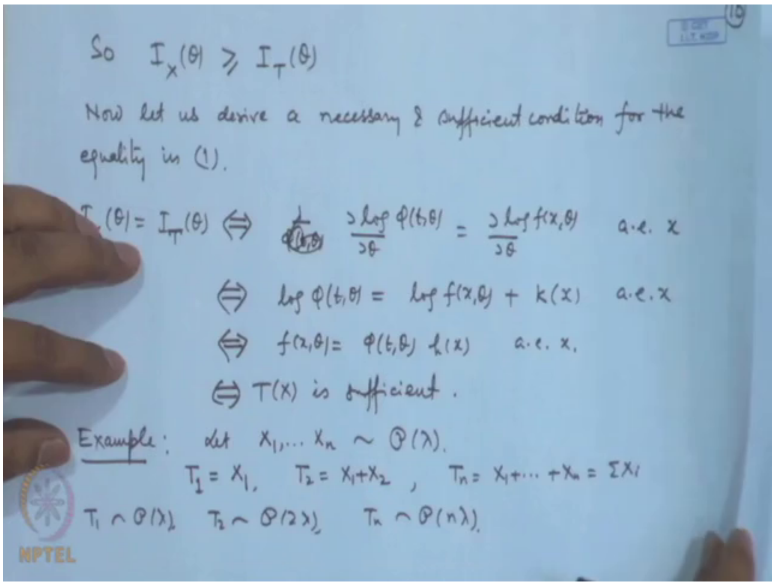


Let us look at the proof of the second part. So, consider here expectation of del phi by del theta 1 by phi t theta minus del f this will be capital here because you are considering expectation. So, this term is going to be greater than or equal to 0 because this is a perfect square term here. Now, let us expand the left hand sidethe left hand side is equal to now you expand this. So, this is becoming expectation of del log of phi t theta by del theta square plus expectation of del log f x theta by del theta square minus twice expectation and the product of these terms, that is del log phi t theta by del theta into del log f x theta by del theta.

At this stage you notice here that expectation conditional expectation of del by del theta log f x theta given T is this term that is dellog this term is nothing, but del log phi T theta by del theta.If I consider this expectation here I can write here it as expectation of expectation given T then this term becomes expectation of del log phi T theta by del theta into del log f x theta by del theta you can express as expectation of expectation del log phi T theta by del theta del log f x theta by del theta given T.Now, if we use the relationship which we proved in the first part that is this one thenthis conditional expectation becomes this term itself. So, this will become square of.

So, left hand side of one is then information in X, the first term is information in T plus information in X minus twice information in T that is equal to information in X minus information in T and the right hand side is it is greater than or equal to 0.

(Refer Slide Time: 38:54)



So, we conclude that so, I x theta is greater than or equal to I t theta. So, information in a statistic is always less than or equal to the information in the full sample now let us consider when we will have equality. Now, let us derive a necessary and sufficient condition for the equality in 1 now when will there be equality if we are considering I x theta is equal to I theta I t theta equal to 0. So, that equal to 0 will come if we have equal to 0 here now this is an expectation of a non negative quantity if we say that expectation is 0 then the quantity itself must be 0 with probability one . So, I x theta is equal to I t theta is equivalent to saying that 1 by phi t theta or you can say del log phi T theta by del theta is equal to del log f x theta by del theta almost everywhere.

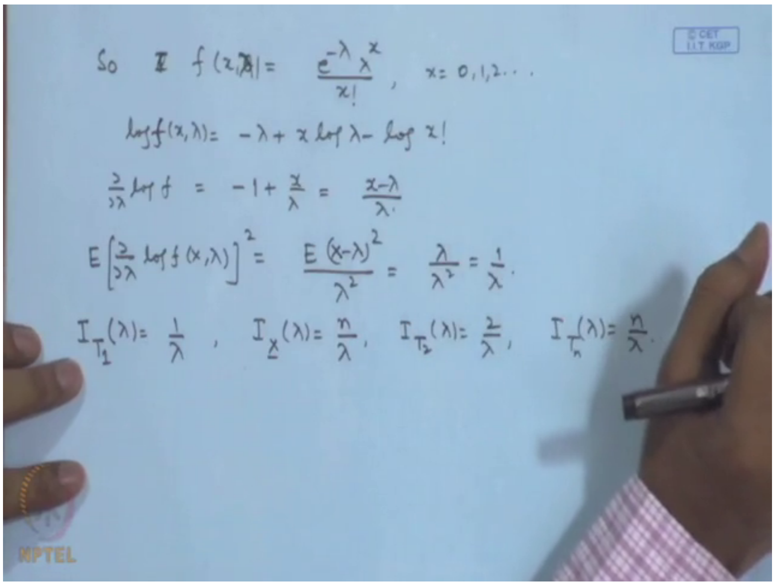
That means, the set of values of x where this is not true will have probability 0. Now, you integrate on both the sides. So, you will get log of phi t theta is equal to log of f x theta plus a function of say x because this integration is with respect to theta. So, this is equivalent to saying that if I consider f x theta, then it is equal to phi t theta into a function of x. Now, this is nothing, but factorization theorem. So, we are saying that T is sufficient.

So, the information in the statistic T is equivalent to equal to the information n x if and only if the random variable the statistic T is sufficient. So, this Fisher's information is measure is extremely important concept. In fact, in the current physics or in the information theory this is widely used one can look at the references physics of Fisher's informationthere is currently a book, which has come out and italmost establishes entire physics theory on the Fisher's information measure.

Let me give an example of a calculation of the information, we will show that this statement is true. So, let me take up say let us consider say X 1, X 2, X n following say Poisson lambda distribution, let us take several statistics let us take say T 1 is equal to X 1,T 2 is say X 1 plus X 2 and say t n is equal to X 1 plus X 2 plus X n that is sigma x i in the case of Poisson distribution, the we can easily derive the distribution. So, T 1 follows Poisson lambda T 2 will follow Poisson twolambda and T n will follow Poisson n lambda.

Let us, independently derive the information in T 1, T 2 and T n and also let us derive the information in the full sample.What is information in full sample? X 1, X 2, X n. So, let us derive all these things.

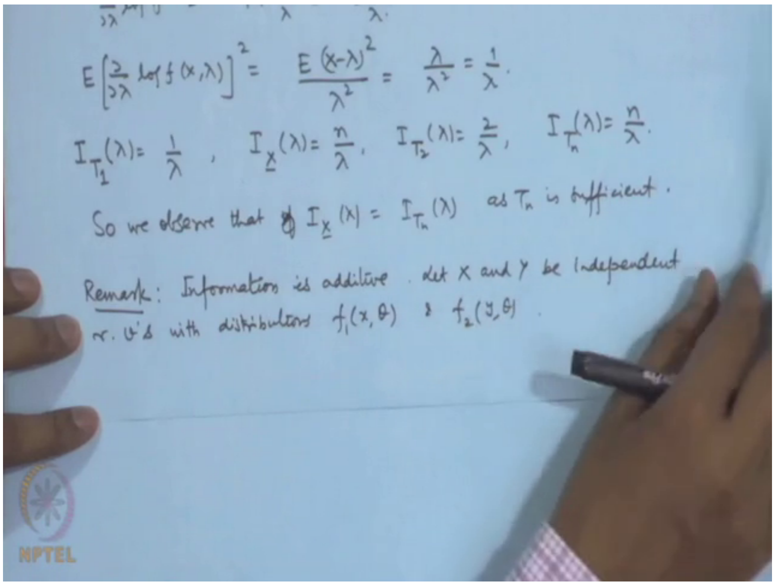
(Refer Slide Time: 43:39)



So, information in one of the x that is calculated.If I calculate the information in X 1(Refer Slide Time: 43:48) and if I take n times that information is easily we can see an additive function. So, the density function or the probability mass functionin the Poisson distribution is. So, log of this is f x lambda minus lambda plus x log lambda minus log of x factorial. So, del by del lambda log of f that is equal to minus 1 plus x by lambda, which we can write as x minus lambda by lambda. So, expectation of del by del lambda log of f x lambda square that is equal to expectation of x minus lambda square by lambda square.

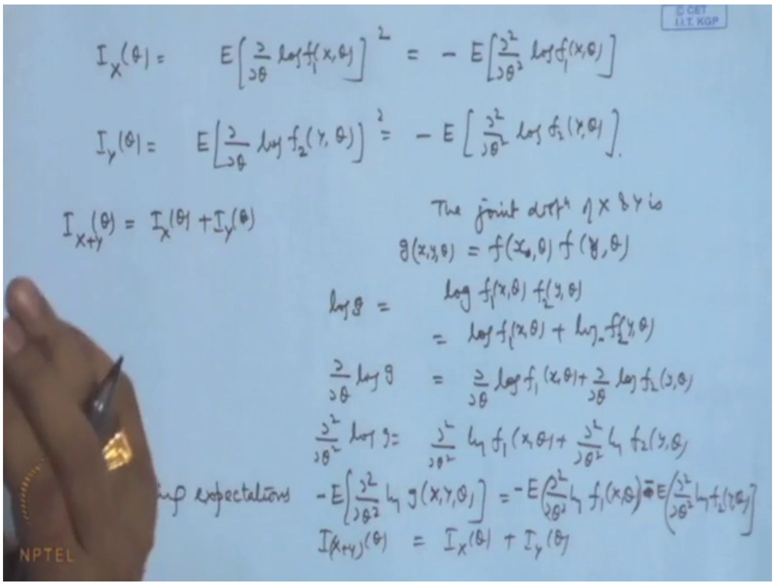
Now, this is nothing, but the variance of x because in Poisson distribution expectation of x equal to lambda. So, this is equal to lambda by lambda square that is equal to1by lambda. So, if I consider the information in T1 then that is equal to 1 by lambda.If we consider the information in say x itself, then it is additive. So, it will become n by lambda, if I consider information in T 2 that will be equal to 2 by lambda and if I consider information in T n that is also equal to n by lambda.

(Refer Slide Time: 43:39)



So, you can see this is less than this, this one is less than this; however, this one is equal to this and T n that is sigma x i in the case of Poisson distribution, we have shown that it is sufficient statistics. So, we observe that ifinformation of X is same as information in T as T n is sufficient.We write a comment here, that information is additive. So, suppose I am considering independent random variables let X and Y be independent random variables with distributions say f 1 x, theta and f 2 y, theta.

(Refer Slide Time: 47:02)

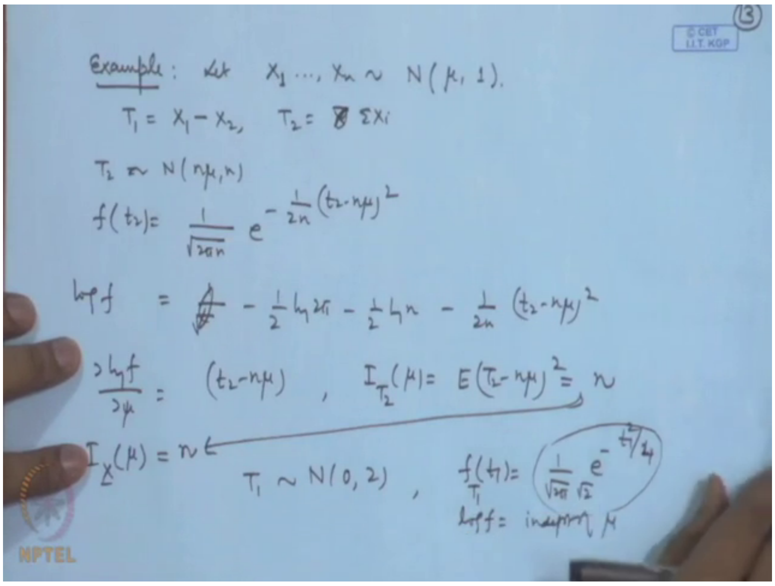


Then, let us consider information in x that is equal to expectation of del by del theta log of f x, theta whole square, which is also same as minus expectation of del 2 by del theta 2 log of f x, theta we have seen this relationship similarly, information in y that is equal to information expectation of del by del theta log of f y say, this is f 1 this is f 2 that we can also write as expectation of del 2 by del theta 2 log of f 2 y, theta. Information in x plus y so; that means, we will consider the joint distribution of the jointdistribution of X and Y is because the distributions are independent, it is equal to the product of the f x theta into f y, theta.

So, if I take log of f x, theta into f y, theta it is equal to log of f x, theta plus log of f y, theta. So, if I consider let me write thisnotation for this joint density here say g of x, y, theta, then log of g is equal to log of f plus log of log of f 1 plus log of f 2. So, if I consider del by del theta log of g that is equal to del by del theta log of f 1 plus del by del theta log off 2. So, if I consider second order derivative del by del theta 2 log of g that is equal to del 2 by del theta 2 log of f 1 plus del 2 by del theta 2 log of f 2.

So, if I take expectations here taking expectations.We get expectation of del 2 by del theta 2 log of g x, y, theta is equal to expectation of del 2 by del theta 2 log of f 1 x, theta plus expectation of del 2 by del theta 2 log of f 2 y, theta. So, if I put a minus sign on both the sidesthen this is becoming information in x plus y and this is becoming information in x and this is information in y. So, we have proved that information is additive, if I am considering independent observations, then information in this total will be equal to the information in one plus the information into the other one, but independence is used here. Let me explain theequality of sufficient statistic informationby means of another example here.

(Refer Slide Time: 50:53)



Let us considersay x, x 1, x 2, x n following say normal mu 1 distribution.Let us consider say T 1 is equal to x 1 minus x 2,T 2 is equal to x bar or sigma x i. So, what is the distribution of T 2. T 2 will have normal n mu n. So, if we want to write down the distribution of this that is equal to 1 by root 2 pi n, e to the power minus 1 by 2 n, t 2 minus n mu square.

That is equal to 1 by root 2. So, if I take log of f I get minus 1 by 2 log of 2pi minus 1 by 2 log of n minus 1 by 2 n, t 2 minus n mu square. So, del log f by del mu that will be equal to T 2 minus n mu because I get a minus n minus 2 n, here which will cancel out. So, if I consider information in T 2 that will be equal to expectation of T 2 minus n mu square that is equal to n.

Now, consider the information in the normal distribution we have already calculated it was equal 2 in 1 of the variables it was one. So, in n variables it is n here which is matching here information in x about mu thatwas also equal to n. So, you can see that these two things are same let us look at the information in T 1. Now,T 1 here is normal 0 2. So, the density function of T 1 will be free from. So, this is simply a constant because it is simply one by root 2 pi into root 2, e to the power minus T 1 by T 1 square by 2T 1 square by 4.

You can see there is no mu occurring here. So, if I consider log of this, this is independent of mu and therefore, if I consider derivative with respect to mu that is going to be 0. So, information in T 1 is simply 0. Now, we will define this conceptlittle later, if the information about the parameter is 0 in the statistic, it will be called ancillary statistic if the information is full; that means, whatever information in the whole sample is there and if that is equal then it is called a sufficient statistic.

So, thisconcept of information is very, very significantit actually tells the kind of statistic that we are considering and therefore, for what purpose it should be used. Now, I have also considered the cases that we can consider more than 1 sufficient statistics. So, we need to distinguish between different sufficient statistics in the sense that.What is the maximum reduction of the data that is possible?That is called the concept of minimal sufficiency. So, in the following lecture I will bestarting the concept ofminimal sufficiency.How to derive it? How to characterize the concept ofminimal sufficiency? So, these are the topic that I will be covering in the next lecture.