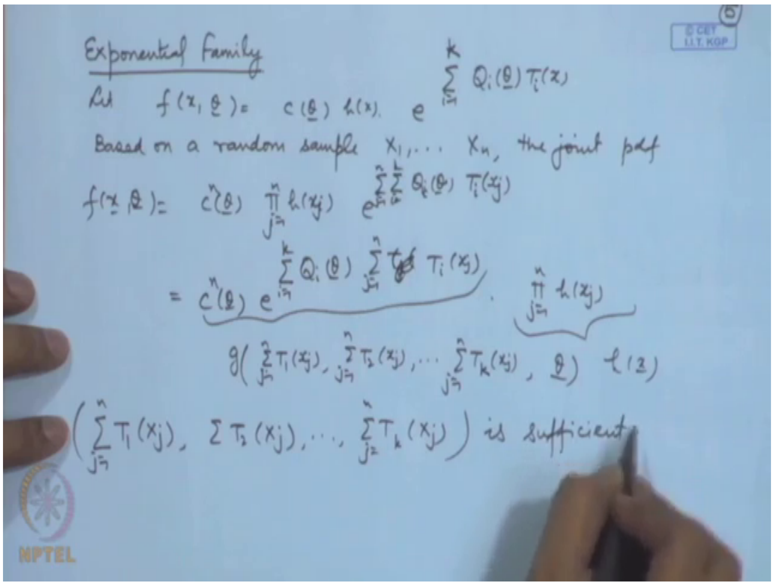
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 26
Sufficiency and Information –II

(Refer Slide Time: 00:24)

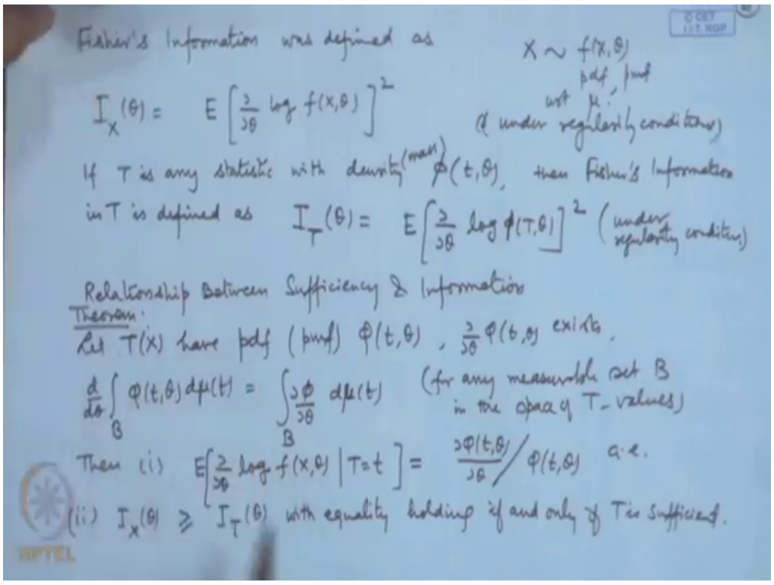


Now, this factorization theorem is very useful if we are considering a general distribution in an exponential family. So, let us consider distributions in exponential family. So, if we are considering a k dimensional exponential family let f x theta be equal to c theta h x e to the power sigma say Q i theta T x i is equal to 1 to say k. This is called a k dimensional exponential family provided 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 hx i e to the power sigma Q i theta into T of. So, let me change here i to j because i is being used here i is equal to 1 to k T x j sigma j is equal to 1 to n. Now this I can write as c to the power n theta e to the power sigma Q i theta sigma T j T of there is a mistake here is T i e i so, T i x j. So, this becomes what I am doing is I am taking this summation inside. So, this becomes sigma of T i x j j is equal to 1 to n, i is equal to 1 to k into product of h x j, j is equal to 1 to n.

So, this part is now a function of sigma T 1 x j, j is equal to 1 to 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 1 to 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: 04:02)

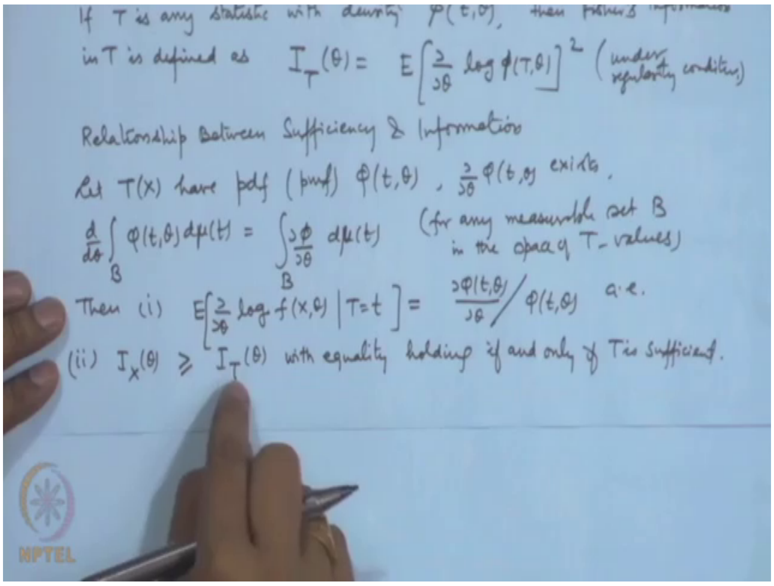


So, if you remember the Fisher's information was defined as, Fisher's information was defined as I X theta is equal to expectation of del by del theta log of f X theta whole square. Here the assumption is that the distribution of X is fx theta and of course, this could be pdf or pmf with respect to a measure mu and we are making the assumption of regularity conditions that is differentiation under the integral sign is allowed. So, this is under regularity conditions. If the distribution of X is f x theta then the information major Fisher's information in X about theta is defined as expectation of del by del theta log of f X theta square.

Now, if T is any statistic and suppose the density of let me give the name as say phi t theta then Fisher'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 pdf also or pmf and we should have the regularity condition satisfied for phi also; that 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 being satisfied under regularity conditions.

So, we have the following result regarding relationship between sufficiency and information let T X have pdf or pmf phi t theta del by del theta phi t theta exists d by d theta of integral phi d theta d mu t over any measurable set 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 phi by del theta divided by phi t theta almost everywhere.

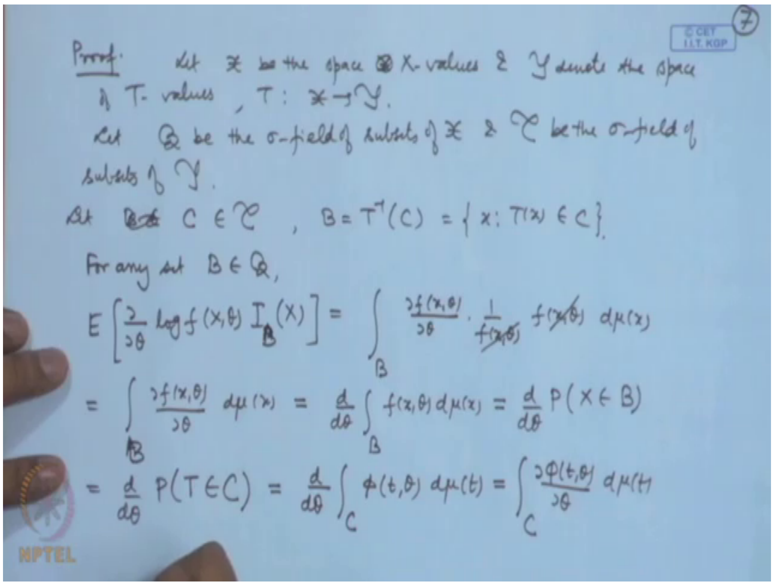
(Refer Slide Time: 08:23)



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 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 and 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 X 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 is 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 this theorem here.

(Refer Slide Time: 10:04)

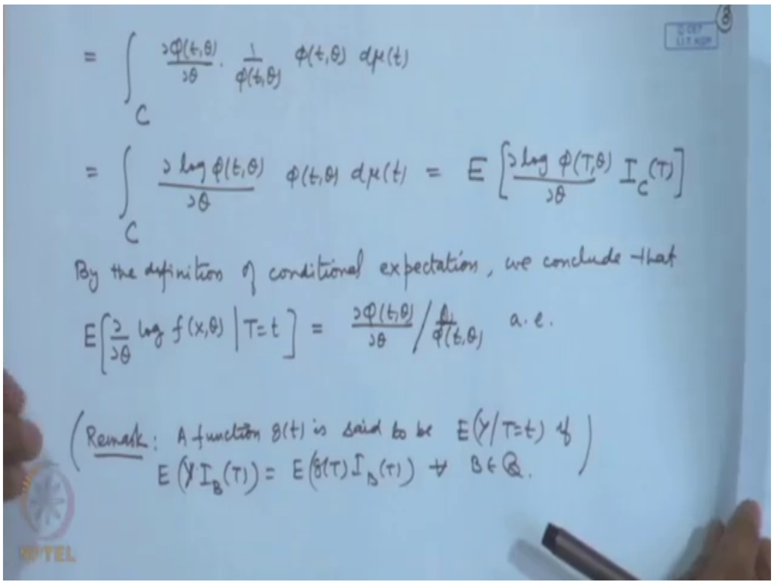


So, see we have say let x be the space of X values. And say y denote the space of T values; that means, T is a function from x to say y. Naturally we will be considering the sigma fields of subsets of x and similarly a sigma field of subsets of y also. So, let us use some notation say B be the sigma 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 a say 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 set B.

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 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 assumption that we can differentiate under the integral sign. So, this is equal to d by d theta of f x theta over B; now this is 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 t theta that is the density of 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 the integral 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: 14:28)

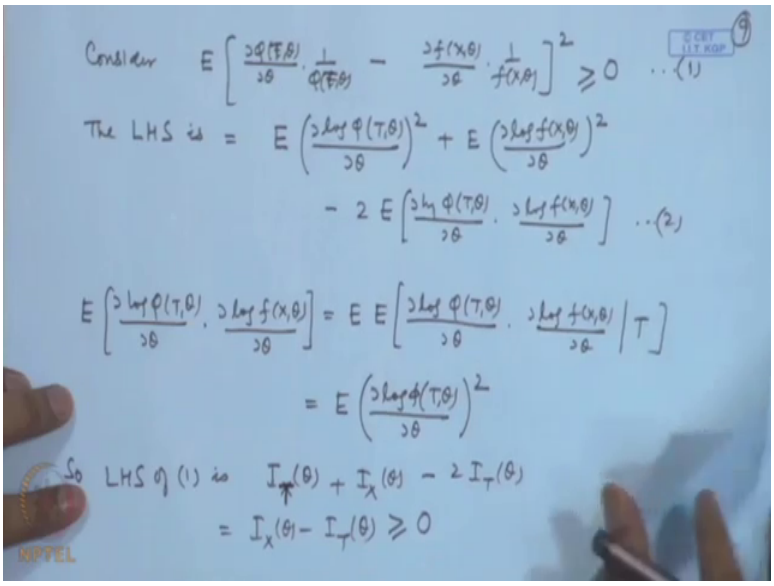


So, we will get here this term is equal to del phi by del theta 1 by phi t theta phi t theta d mu t over set C. So, now this becomes nothing, but the derivative of log of phi t theta d mu t over the set C; this is nothing, but the expectation of this expectation of del log phi T theta by del theta indicator 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 definition of conditional expectation.

We conclude that expectation of del by del theta log of f x theta given T is equal to t it is equal to del phi by del theta 1 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 where this may not be true will have probability 0; that means, they are 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 what we have used here is we have simply used the definition of the conditional expectation. In fact, let me write here remark a 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 B T for all B boreal measurable sets of B. So, we have used this definition here. So, what we have done is we have established a relationship in the condition 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: 18:24)



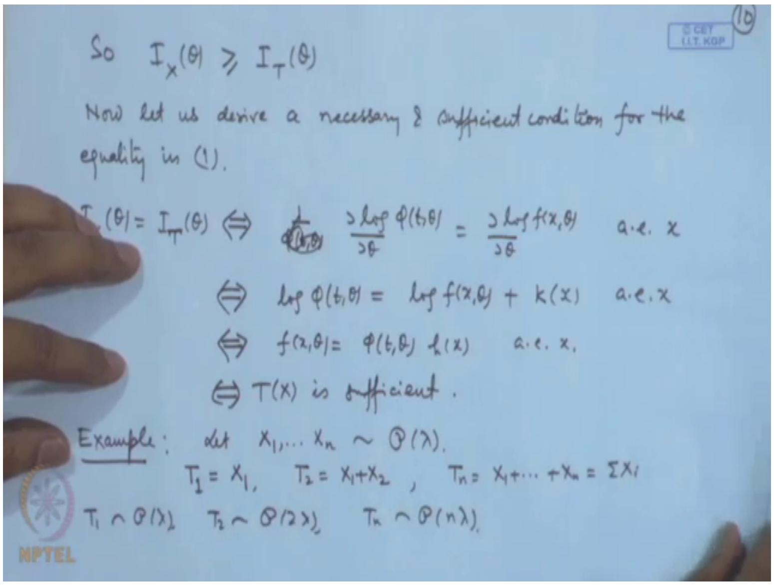
Let us look at the proof of the second part. So, consider here expectation of del phi by del theta 1 by phi T theta minus del f this should be capital here because we are considering expectation. So, this term is going to be greater than or equal to 0 because, this is a perfect square term here. Now let us expand the left hand side the 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 of x theta given T is this term that is del log this term is nothing, but del 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 we 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 then this conditional expectation becomes this term itself.

So, this will become a square of. So, left hand side of 1 is then information in x sorry 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 that it is greater than or equal to 0. So, we conclude that.

(Refer Slide Time: 22:01)



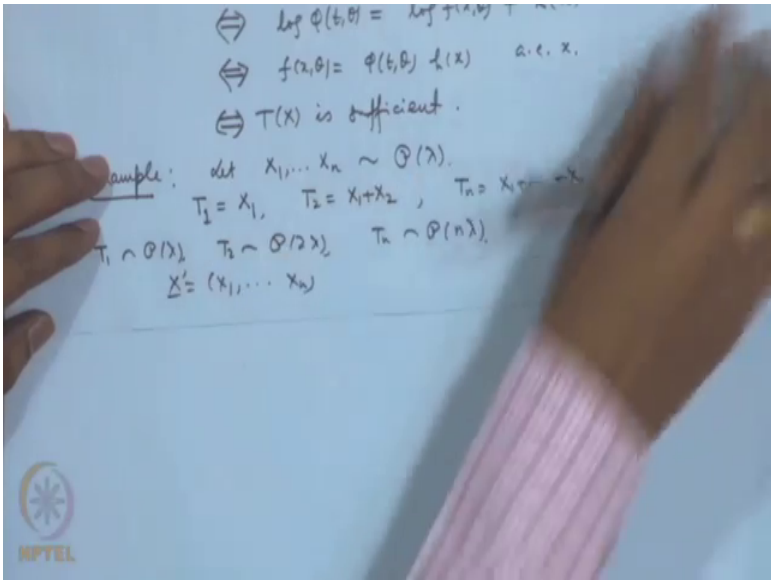
So, I x theta is greater than or equal to I T theta. So, information in a statistic is always less than or equal to the information in the full sample. Now let us consider when we will have equality; now let us derive a necessary and sufficient condition for the equality in 1. Now when will there be equality.

If we are considering IX 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 1. So, IX 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 in x if and only if the random variable the statistic T is sufficient here. So, this Fisher's information 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 differences physics of Fisher's information there is a currently a book which has come out and it almost establishes the entire physics theory on the Fisher's information measure.

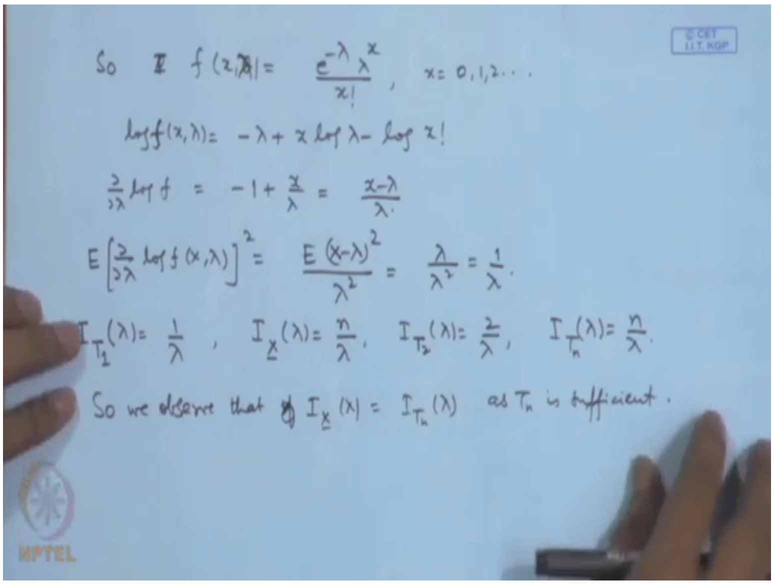
Let me give an example of calculation of the information; we will show that this statement is true. So, let me take up say let us consider say X 1 X 2 X n following say Poisson lambda distribution. Let us take several statistics let us take say T 1 is equal to X 1 T 2 is say X 1 plus X 2 and say T n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

(Refer Slide Time: 26:37)



What is information in full sample X 1 X 2 X n? So, let us derive all these things.

(Refer Slide Time: 26:45)

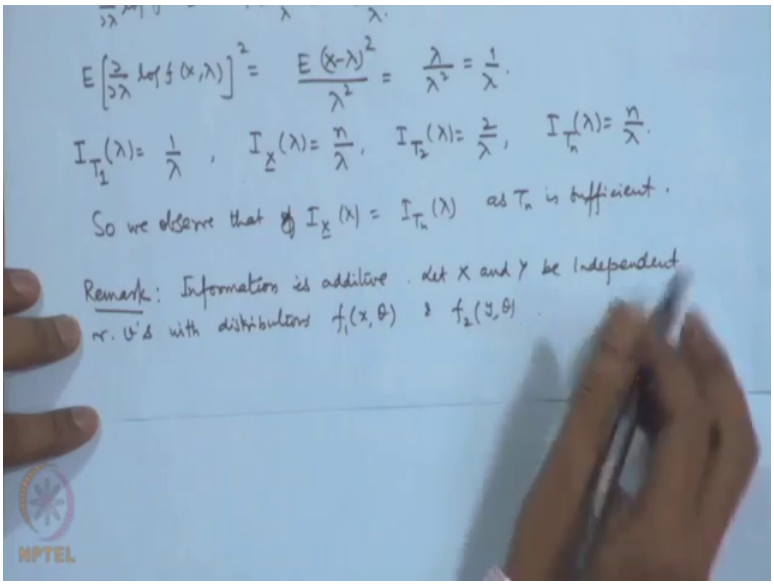


So, information in one of the x that is calculated if I calculate the information in X 1 and if I take n times that information is easily we can see an additive function. So, the density function or the probability mass function in the poison distribution is.

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 fx 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 is equal to lambda. So, this is equal to lambda by lambda square that is equal to 1 by lambda. So, if I consider the information in T 1 then that is equal to 1 by lambda if we consider the information in say X itself.

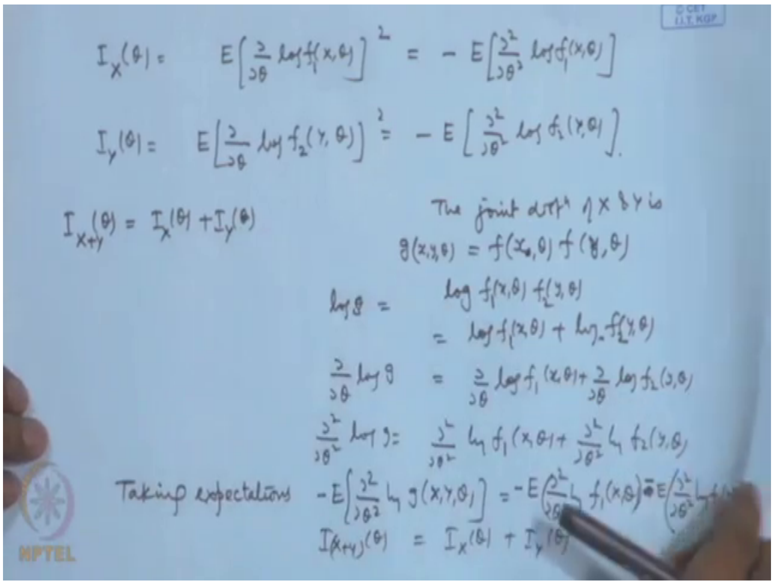
Then it is additive so, it will become n by lambda, if I consider information in T 2 that will be equal to 2 by lambda and if I consider information in T n that is also equal to n by lambda. So, you can see this is less than this one is less than this; however, this one is equal to this and T n that is sigma X I in the case of Poisson distribution we have shown that it is sufficient statistics. So, we observe that if information of X is same as information in T as T n is sufficient.

(Refer Slide Time: 29:20)



We write a comic 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: 30:09)

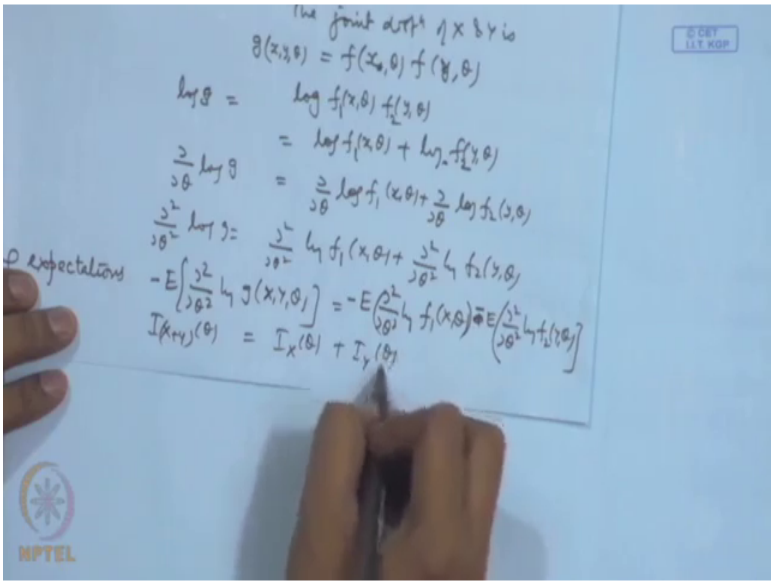


Then let us consider information in X that is equal to expectation of del by del theta log of f x theta whole square which is also same as minus expectation of del 2 by del theta 2 log of f x theta we have seen this relationship. Similarly information in y that is equal to information expectation of 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 joint distribution 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 this notation for this joint density 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 of f 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.

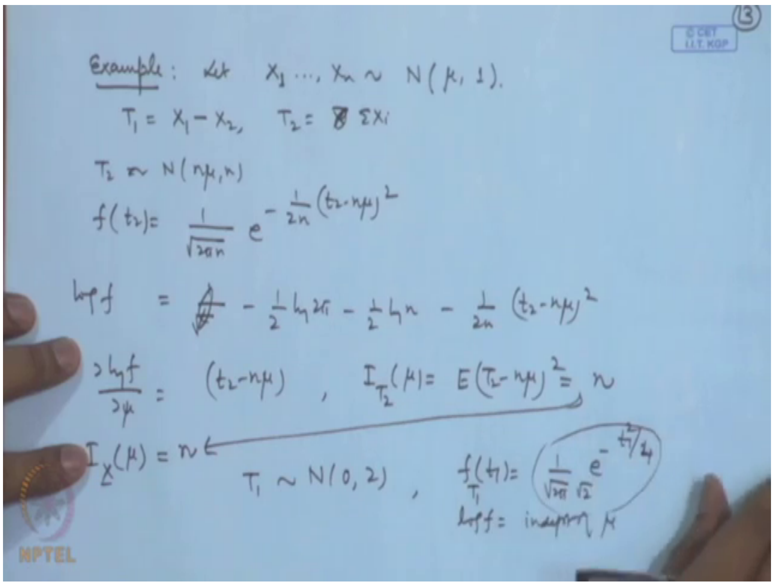
(Refer Slide Time: 32:56)



Expectation of del 2 by del theta 2 log of f 2 y theta. So, if I put a minus sign on both the sides then 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 1 plus the information into the other one, but independence is used here. Let me explain the equality of sufficient statistics information by means of another example here.

(Refer Slide Time: 33:59)

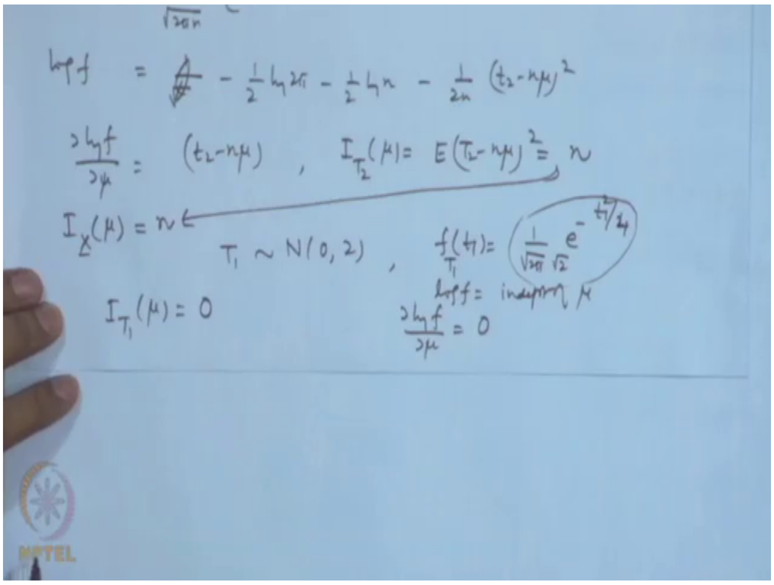


Let us consider say X 1 X 2 X N following say normal mu 1 distribution. Let us consider say T 1 is equal to X 1 minus X 2 T 2 is equal to X bar or sigma X i. So, what is the distribution of T 2? T 2 will have normal n mu n. So, if you want to write down the distribution of this that is equal to 1 by 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 2 by 2 log of 2 pi 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 to in one of the variables it was 1. So, in n variables it is n here which is matching here information in X about mu that was 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 1 by root 2 pi into root 2 e to the power minus t 1 by t 1 square by 2 t 1 square by 4, you can see there is no mu occurring here. So, if I consider law of this is independent of mu.

(Refer Slide Time: 36:55)



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 concept a little 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 statistics. So, this concept of information is very very significant it actually tells that 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 one 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 be starting the concept of minimal sufficiency, how to derive it? How to characterize the concept of minimal efficiency? So, these are the topics that I will be covering in the next lecture.