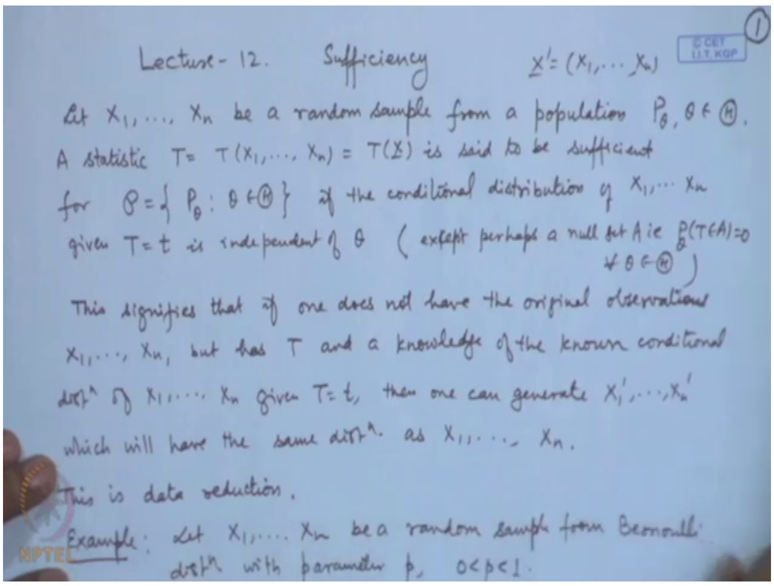
Statistical Inference
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 23
Sufficiency – I

Now, I start with a new concept that is called sufficiency. In the context of statistical inference, there is a concept which is useful to retain the necessary data without losing any information. What is the literal meaning of the word sufficiency? The literal meaning of the word sufficiency is that it is enough; sufficient means enough. So, usually we are dealing with the statistical model that we deal in the inference problem is that we say let X 1, X 2 X n be a random sample meaning thereby that we have data on n observations or you can say n data points are available to us.

Now, in many of the practical problems, it becomes difficult to retain the data because, it may occupy lot of storage space whether it is on computer or it is in the form of hard copy of the data and then there is a danger of losing the data. It will be always interesting to say that let us keep the minimum things such that whatever information or whatever useful inferences, we want to make we are not suffering in that. That means, we do not lose any important part of it.

(Refer Slide Time: 01:55)



A formal specification of this concept is called sufficiency or sufficient statistic in the context of statistical inference.

So, let us introduce the formal definition of sufficiency as before we have a random sample. So, let X 1 X 2 X n be a random sample from a population say p theta theta belonging to say script theta a statistic. So, a statistic we have already defined a statistic means a function of observations. So, T that is T X 1 X 2 X n, which we also write as T X; that means we are denoting X as X 1 X 2 X n. So, T X is said to be sufficient. Now what do you mean by sufficient for what? So, we usually mention the word sufficient for the family of probability distributions.

In loose turns we also say sufficient for the parameter theta meaning thereby that whatever be the parameter under consideration, many times in the problems we will have 1 dimensional parameter 2 dimensional parameter etcetera. In that case, we will have to consider a specifically what parameter is being considered. So, the formal definition, I am writing for the family of probability distributions; meaning thereby that whatever parameters are under consideration this could be a scalar or a vector parameter.

So, this is said to be sufficient if the conditional distribution of X 1 X 2 X n given T is equal to say small t is independent of theta. Of course, except perhaps a null set A that is on a set a where T takes probability 0. So, this is a exceptional case, but in general the distribution of the random sample given the statistic, if it is independent of the parameter then we say that this t is independent, then we say that this t is a sufficient statistic.

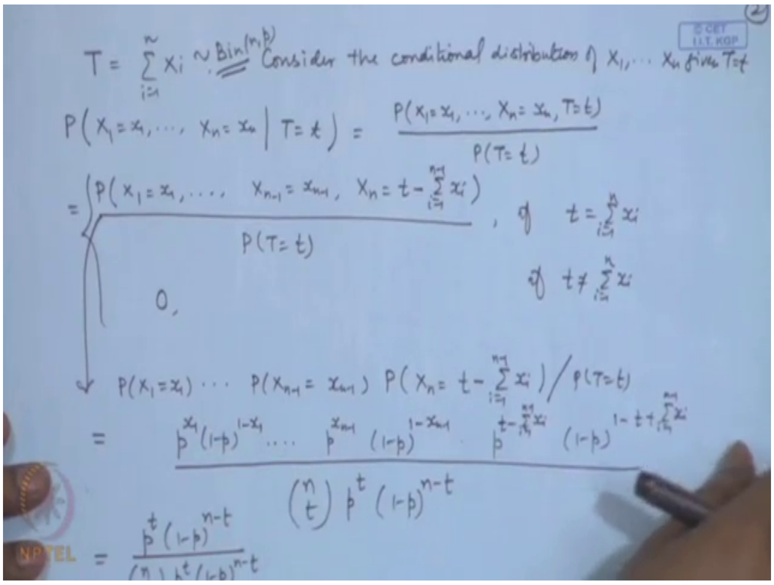
Now what is the physical interpretation of this definition that the distribution of X 1 X 2 X n is free from theta and then we said is sufficient what does it mean? It means that now if the distribution is free from theta; that means, the distribution of X 1 X 2 X n given T is completely known. So, suppose we know T, we know the distribution of T now this conditional distribution of X 1, X 2, X n given T, since it is free from theta then that is also known. Therefore, if I merge these 2 distributions that is the conditional distribution of X 1, X 2, X n given T and the distribution of T I get the joint distribution of X 1 X 2 X n and T from there I get the distribution of X 1, X 2, X n.

It means that even if I may not have the initial X 1, X 2, X n with us, but we can generate that distribution once again, because of the information or we can say the distribution of X 1 X 2 X n given T being free from the parameter and T is known to us. This signifies that if one does not have the original observations X 1 X 2 X n, but has T and a knowledge of the known conditional distribution of X 1, X 2, X n given T then one can generate say X 1 prime, X 2 prime X n prime, which will have the same distribution as X 1 X 2 X n.

So, this is called data reduction as we will show later on that in most of the practical problems this sufficient statistics will become like 1 dimensional or 2 dimensional thing although, you may have any number of observations. So, this data reduction is helpful and we will show a statistically also that basing our decisions on the sufficient statistics is also useful. That means, if there is any inference made in the terms of estimation testing of hypothesis etcetera. If I am making inference based on the sufficient statistics we are better off.

So, let me explain this example say a binomial distribution example let me take. Suppose, I have X 1, X 2, X n be a random sample from say Bernoulli distribution with parameter p a p lies between 0 and 1.

(Refer Slide Time: 09:09)



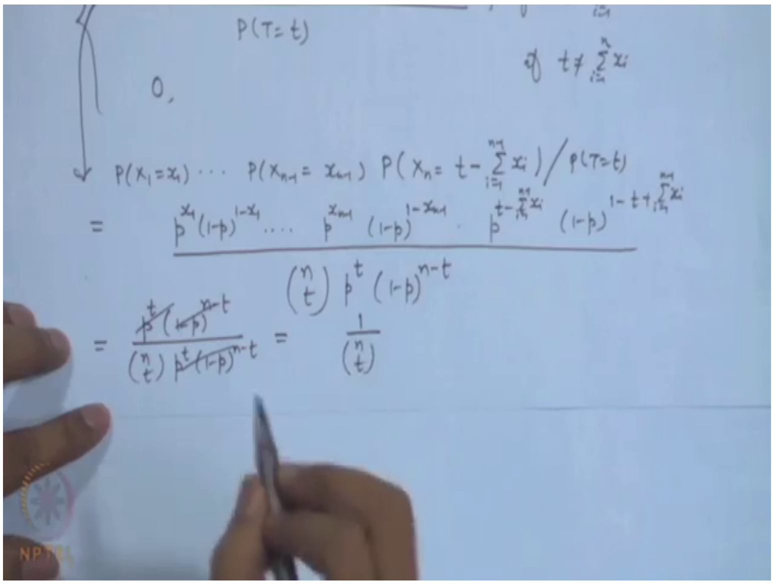
Let us consider say T is equal to sigma X i i is equal to 1 to n. Let us look at the conditional distribution of consider the conditional distribution of X 1, X 2, X n given T that is equal to X 1 is equal to X 1 and so on, X n is equal to X n given T is equal to t that is equal to probability of X 1 is equal to small x 1 and so on X n is equal to small x n T is equal to t divided by probability of T is equal to t.

Now, that is equal to since T is equal to sigma X i, if small x 1 plus small x 2 plus small X n is equal to t then only this probability will be calculated. In other cases, this will be simply equal to 0 so that is equal to probability of X 1 is equal to small x 1 and so on, X n minus 1 is equal to small x n minus 1 and X n is equal to t minus sigma X i i is equal to 1 to n minus 1, if t is equal to sigma X i i is equal to 1 to n otherwise this is 0.

Now here, we can make use of the fact that X 1, X 2 X n are independently distributed Bernoulli random variables. So, if they are independent this probability of the joint occurrence will be equal to the product of these probabilities. So, this term let me write this term is anyway 0. So, this term is equal to probability of X 1 is equal to small x 1 and so on X n minus 1 is equal to small x n minus 1 probability of X n is equal to t minus sigma X i i is equal to 1 to n minus 1 that is equal to p to the power X 1 1 minus p to the power 1 minus X 1 and so on; p to the power X n minus 1 1 minus p to the power 1 minus X n minus 1 p to the power t minus sigma X i i is equal to 1 to n minus 1 1 minus p to the power 1 minus t plus sigma X i i is equal to 1 to n minus 1 and divided by probability t is equal to t.

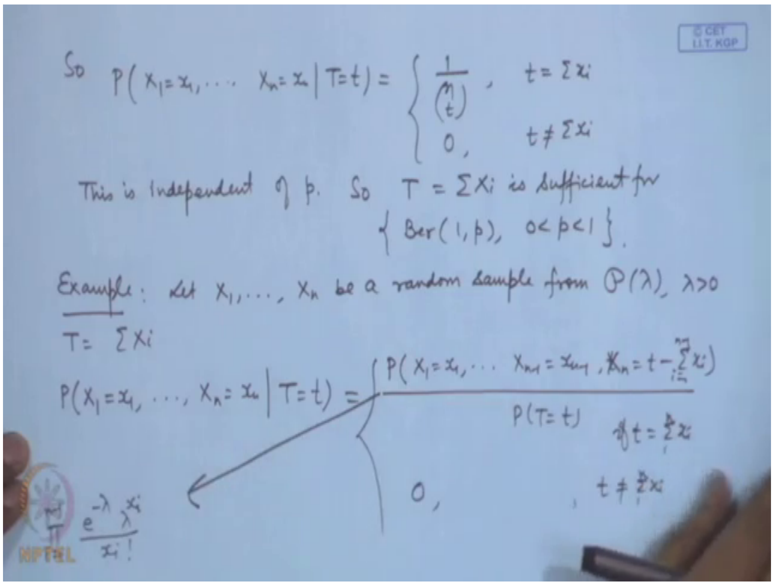
Now, what is the distribution of t? If X 1 X 2 X n are Bernoullis independent then this will be binomial n p. So, probability t is equal to t that will be equal to n c t p to power t 1 minus p to the power n minus t. Now we can easily see these terms this p to the power terms if you add, you will get p to the power t. Similarly, if you add 1 minus p exponents, you will get n minus sigma X i so, that will cancel out with plus sigma xi, you get n minus t. So, you get it as p to the power t into 1 minus p to the power n minus t divided by n c t p to the power t into 1 minus p to the power n minus t.

(Refer Slide Time: 13:19)



Now, this term simply cancels out. So, we get it as 1 by n c t. So, this conditional distribution then we can express as.

(Refer Slide Time: 13:31)



Probability of X 1 is equal to small x 1 and so on X n is equal to small x n given T is equal to t that is equal to 1 by n c t for t is equal to sigma x i and it is equal to 0, if t is not equal to sigma xi. You look at this term there is no theta and no parameter appearing here p is not appearing here. So, this is independent of p. So, T is equal to sigma X i is sufficient for the family of Bernoulli distributions.

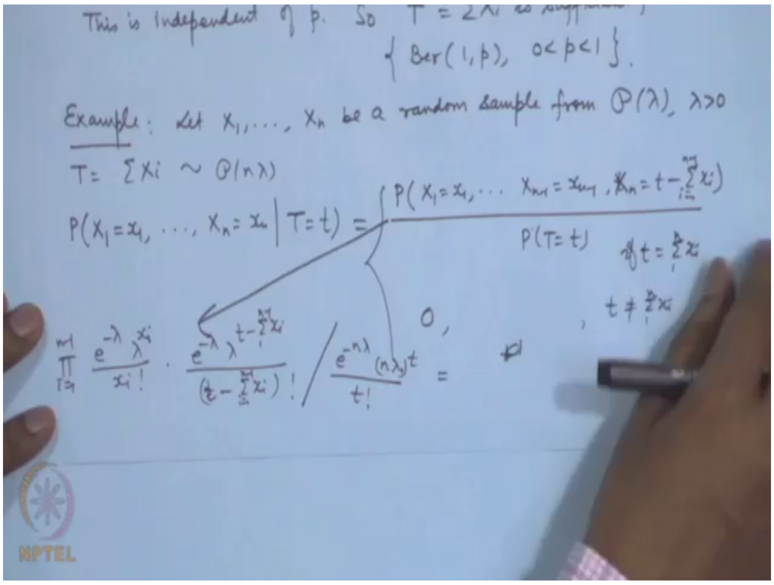
We may also say it as that T sufficient for p here. Now note here, the physical significance of sufficiency, if we are observing X 1 X 2 X n as p independent Bernoulli random variables. Ihat means, they are observations related to success or failure in an Bernoulli and trials. For example, you are looking at a game of say dart and we are considering hitting a target and we make n aims at the target then what is important, whether individual hits whether this say second one hit correctly. Third one did not hit correctly, is it important information or out of n total attempts, how many are correct? That means, X that is some of exercise

Now, here you see in the concept of sufficiency exactly sigma X i is turning out to be sufficient. Therefore, this is the relevant information and whatever individual information about X 1 X 2 X n is there that is not necessary to be written. In fact, now if we know this and we know the distribution of t that is binomial n p, we can generate another random sample let us call it say X 1 prime X 2 prime X n prime, which will have Bernoulli 1 p distribution.

Let me explain through another example say, let X 1 X 2 X n be a random sample from Poisson lambda distribution, where lambda is positive once again let us define T is equal to say sigma X i. Now you can proceeding the same way like in the binomial case, we can consider X 1 is equal to X 1 and so on X n is equal to X n given T is equal t. So are going as before we get it as X 1 is equal to X 1 and so on X n minus 1 is equal to X n minus 1 X n is equal to t minus sigma X i i is equal to 1 to n minus 1 divided by probability t is equal to t, if t is equal to sigma X i 1 to n it is equal to 0, if t is not equal to sigma X i 1 to n.

So once again, this term will be equal to e to the power minus lambda lamda to the power X i by X i factorial for i is equal to 1 to n minus 1.

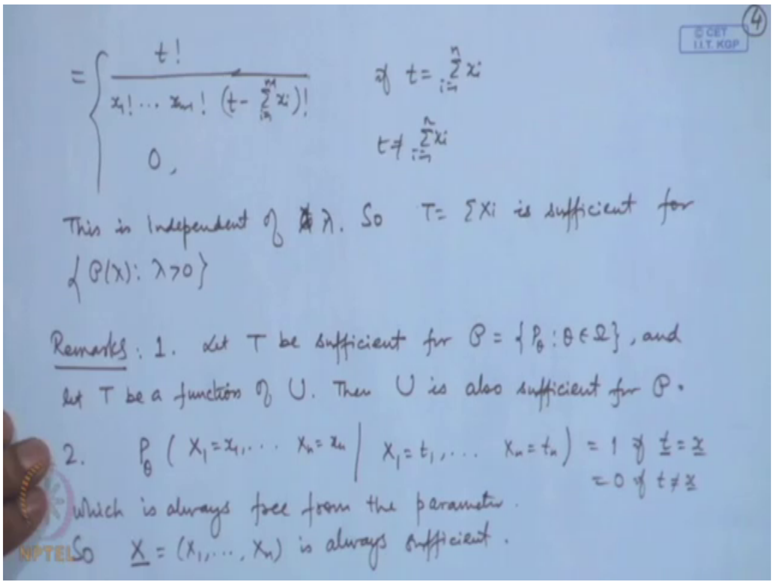
(Refer Slide Time: 17:47)



And the last one is e to the power minus lambda lambda to the power t minus sigma X i 1 to n minus 1 divided by t minus sigma X i i is equal to 1 to n minus 1 factorial. Now this will be coming e to the power minus n minus 1 lambda and then e to the power minus n lambda and also we have in the denominator t. Now, this will follow Poisson n lambda because, Poisson distribution is additive. So, if we are considering a random sample each one following Poisson lambda then sigma X i will follow Poisson n lambda. So, we can write e to the power minus n lambda n lambda to the power t by t factorial.

So, this e to the power minus n lambda cancels out and if you look at lambda to the power X 1 plus X 2 plus X n minus 1 that cancels here, you get lambda to the power t and in the denominator also we have lambda to the power t here. So, what we get here? This t factorial will go in the numerator. So, let me write it here.

(Refer Slide Time: 19:01).



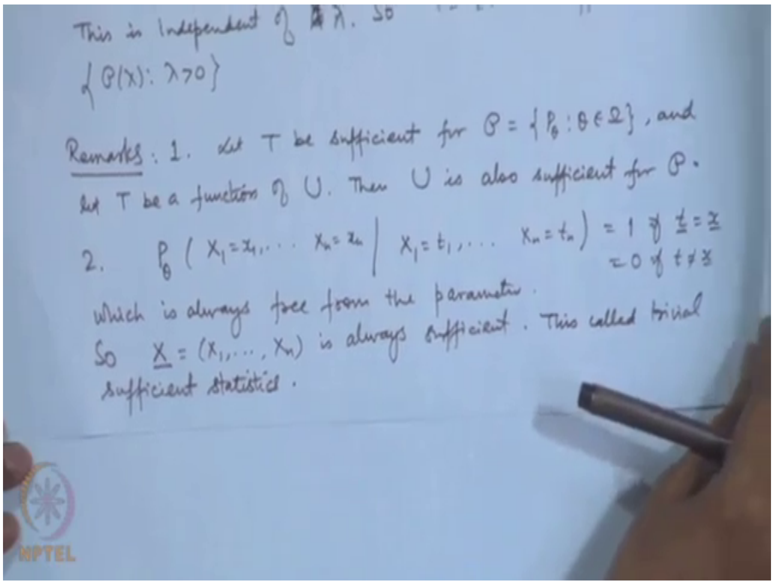
This is equal to t factorial divided by X 1 factorial X 2 factorial X n minus 1 factorial t minus sigma X i i is equal to 1 to n minus 1 factorial. If t is equal to sigma X i and it is equal to 0, if t is not equal to sigma X i i is equal to 1 to n

Once again, you notice here that this is independent of t. So, T is equal to sigma X i, this is independent of lambda sorry. So, T is equal to sigma X i is sufficient for the family of Poisson distributions, we may also say that sigma X i is sufficient for the parameter lambda. Now we can make certain statements here if, I am considering conditional distribution of X 1 X 2 X n given t and suppose t is a function of u. Then if I considered the conditional distribution of X 1 X 2 X n given u then that will also be free from the parameter. Because, if that is not free from the parameter then the conditional distribution of X 1 X 2 X n given t will also not be free from the parameter. Therefore, if t sufficient and t is a function of u then u is also sufficient and of course, if I have a 1 to 1 function of t then that will also be sufficient.

So, let me give some remarks here, let T be sufficient for a family of distributions and let T be a function of U then U is also sufficient for P. Another point that you notice here that if I considered the conditional distribution of X 1 X 2 X n given X 1 is equal to a small x 1 x 2 is equal to a small x 2, x n is equal to a small x n. Then that is always independent of parameter, we can write conditional distribution of say X 1 is equal to X 1 X n is equal to X n given, say X 1 is equal to t 1 and so on, X n is equal to t n this is equal to 1. If this t vector is same as x vector otherwise it is 0. So, this is naturally free from the parameter free from the parameter. So, the sample X is always sufficient.

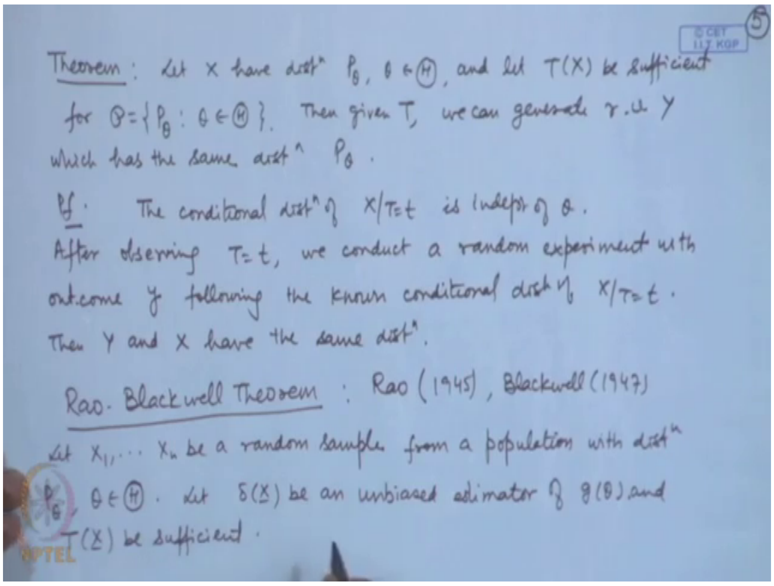
So, the full sample is always sufficient. We will be interested in getting some sort of reduction over there.

(Refer Slide Time: 22:59)



This is known as trivial sufficient statistics; trivial sufficient statistics.

(Refer Slide Time: 23:25)

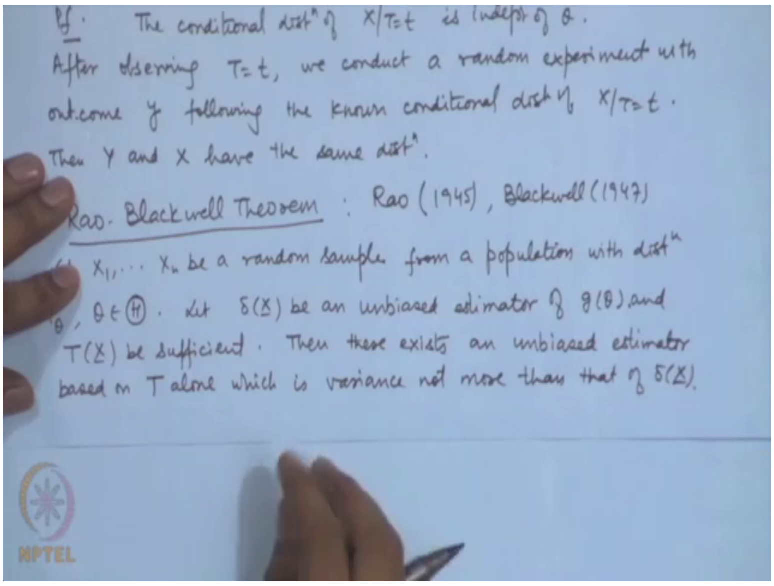


Let me formally prove that given a sufficient statistics, you can generate the original sample. So, let X have distribution say P theta theta belonging to say script theta and let T X be sufficient then given T. We can generate random variable y, which has the same distribution p theta that is the same distribution of X. So, the conditional distribution of X given T is independent of theta. So, after observing T is equal to t, we conduct a random experiment with outcome, say y following the known conditional distribution of X given T then Y and X have the same distribution.

Now, another important significance you can say of sufficient statistics is that if we are considering any unbiased estimator. I can have another unbiased estimator, which is based on the sufficient statistics, and it is variance will be less than or equal to the variance of the initial estimator this famous result is known as Rao Blackwell theorem.

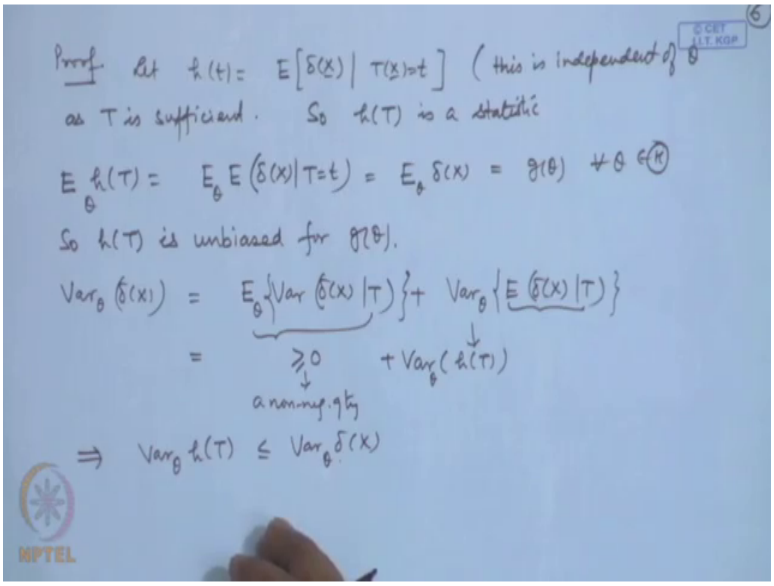
It is named after the Indian statistician CR Rao, who proved this result in 1945 and David Blackwell 1947, let X 1 X 2 X n be a random sample from a population with distribution p theta theta belonging to say a script theta, let delta X be an unbiased estimator of parametric function say g theta and T be sufficient.

(Refer Slide Time: 27:05)



Then there exists an unbiased estimator based on T alone, which has variance not more than that of delta X. Now this is a very very significant statement in a given problem, if I have a sufficient statistics then I can always base our unbiased estimators on that statistics. So, that I will do better than if I do not base it. That means, I will be utilizing the full information in the sample for making my statistical inference the proof is in fact, not very difficult.

(Refer Slide Time: 28:15)



Let us consider h t to be expectation of delta X given T, X is equal to t. Since, we know the conditional distribution of X given T is independent of theta; therefore, this expectation is going to be a function of t alone ok. This is independent of theta as T is sufficient. So, h t is a statistic and I can consider it for my estimation purpose, let us consider expectation of h t. Now expectation of h t is simply expectation of expectation delta X given t. Now this is nothing, but expectation of delta X that is equal to g theta.

So, this new estimator that I have used h T is unbiased. So, this is unbiased for g theta further. Let us consider say variance of delta X. Now this variance of delta X, I can express as expectation of variance delta X given T plus variance of expectation delta X given T. Now, this is equal to this quantity, if you see this is a non negative quantity and expectation of delta X given T be you are defined to be h T. So, this is equal to variance of h t. So, what we are getting? Variance of delta X is equal to variance of h plus a non negative quantity a non negative quantity. That means, variance of h t is going to be less than or equal to variance of delta X.

We will give applications of this result a little later.