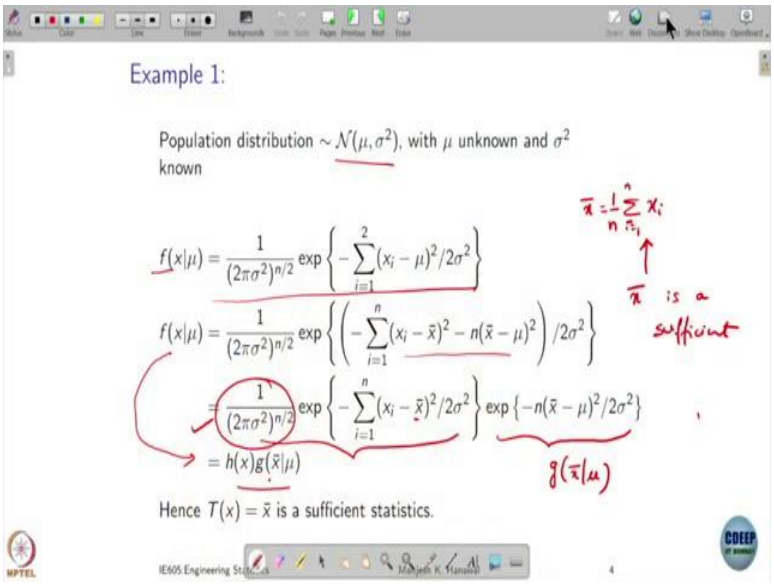
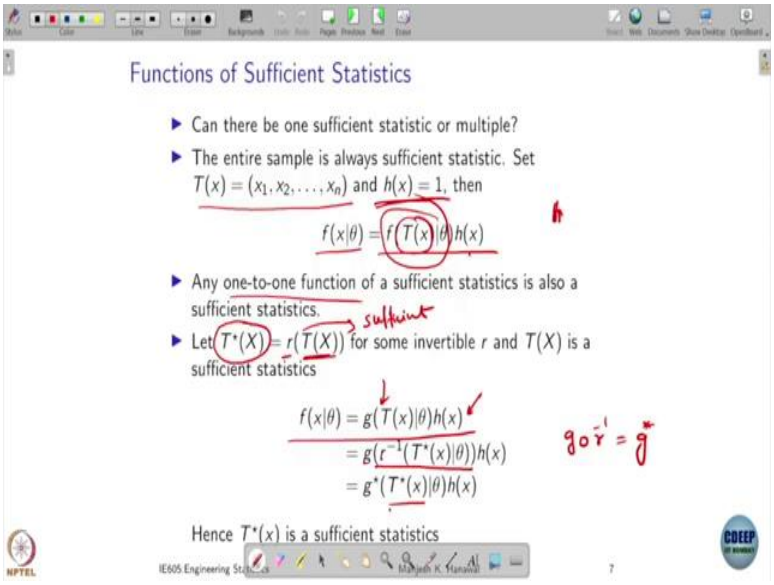
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay
Lecture 35
Minimal Sufficient Statistics

(Refer Slide Time: 00:22)

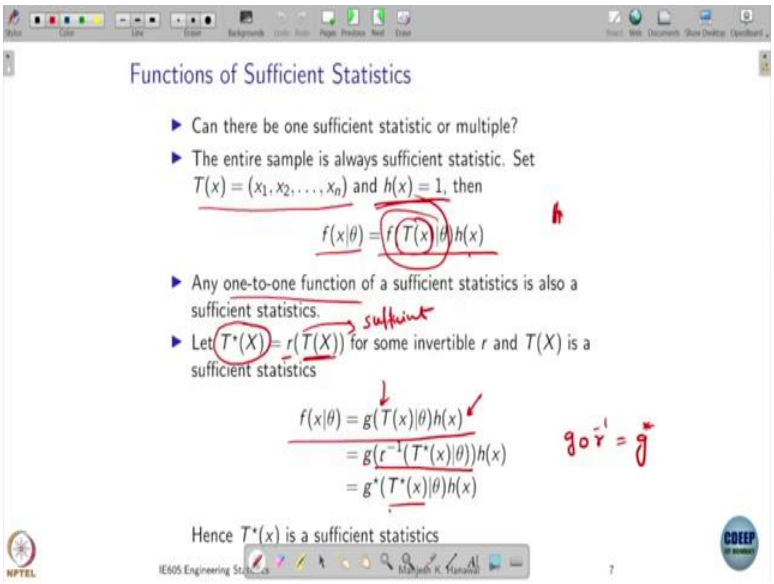




So, any questions from the last lecture, what are we discussed about this factorization theorem, how to use this factorization theorem to come up with some statistics which are sufficient and then if you know already some function is belongs to exponential family, all we need to do is

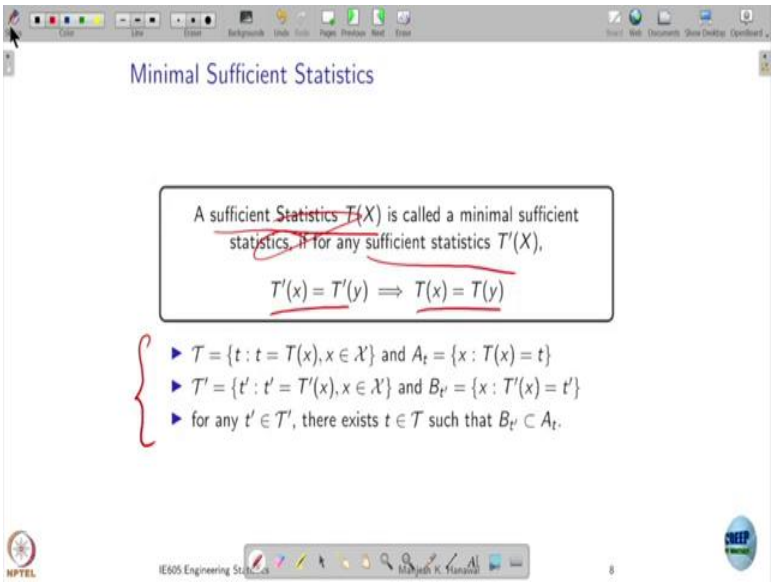
write it in its in this form, and they are all the ti functions will already give me the sufficient statistics for that.

(Refer Slide Time: 00:50)



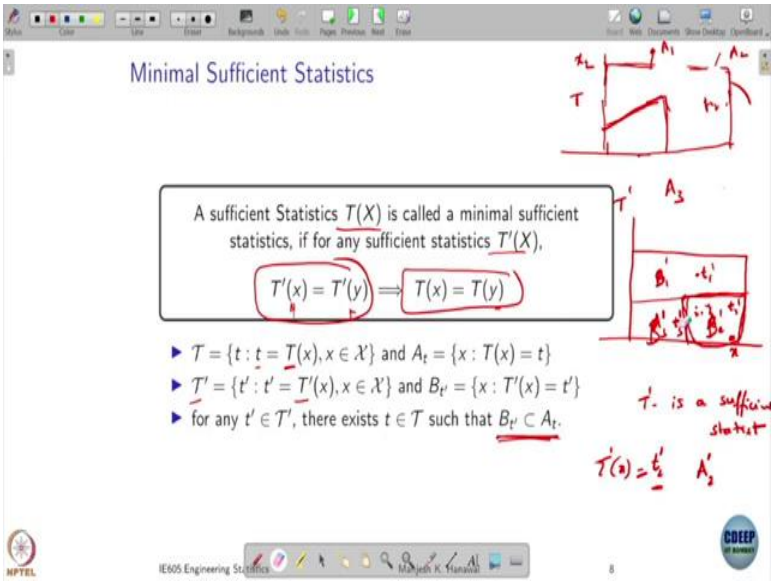
We also discussed that sufficient statistics need not be unique there could be multiple ways of getting sufficient statistics. And if you give me one sufficient statistics, I may derive multiple sufficient statistics by using some invertible function on it.

(Refer Slide Time: 01:09)



Then we started talking about minimal sufficient statistics. So, let us repeat some of the part here what you are already done. So, I am slowing myself by repeating things. If you want to further slow, you should interact or ask something related to this, I mean, I am trying to incorporating the bringing the slowness by repeating things, but if you also want to do from your side, you should act do not simply sit.

(Refer Slide Time: 01:41)



So, by definition, we said that we are going to call a statistic T or rather a sufficient statistic T to be minimal sufficient statistic, if you have any other sufficient statistics T' and that T' take same value on 2 samples x and y. Then it implies that my T also takes the same value on those 2 samples x and y.

That means if two samples are indistinguishable under sufficient statistics t, then they are also indistinguishable under the sufficient statistic, or minimal sufficient statistic T. Now, let us quickly understand this suppose, let us say what I have written in text, let us understand that through diagram.

Let us hypothetically just make a 2 dimensional case x1 and x2. Let us say this is my region for some reason. And hypothetically assume that this region gets partitioned into 3 parts under my hypothesis T. What I mean by that? Take any point here, let us say x1 comma x2, this is going to

map if you apply T on that, that your function t on that it is going to give you a value small t and that value small t is the same in this region.

You understand that part, all the points in this region map to the small t and similarly all the points in this also mapped to 1 T and like this I have made it into 3 partition. And let us call this region as A1, this region as A2 and this region as A3. So, all the points in the region A1, they have the same value. When you apply T on that it they will map to the same point.

Now, let us take another T'. Sorry, let us take another statistics T' this another statistics. So, let us call this is T and let us call let us do another partition under this this is under T'. Not necessary that this T' will also result in the same partition. It may result in some other partition. Let us call that as simply like this, under this let us call this A1'. And let us call this A2'. And let us call this A3'.

In fact, it may result in more than 3 partitions. It may have more than 3 or maybe less than 3, whatever. I am just for representation purposes I am just writing it as 3. Now, suppose T' is a sufficient statistic let us assume that and let us call the point to which all these maps here as called, let us call T1', let us call this as T2' and let us call all the value it maps to all the points in this region maps to as T3'. And similarly, here that is called T1, and this is T2, and let us call this T3.

Now, from this region, I have all the points gets either mapped to T1', T2', or T3'. Now, arbitrarily, let us pick T2'. And let us take some point. Let us, let us, let us know I know let some point here. Let some point let us call this point some x again, that consists of x1 and x2, let us say this T1' x maps to T2'. In fact, all points in this region are going to map to T2'. Now, this region let us call this whatever this where this T2', let us call this A2', I know that all these points here in this region, they are all getting mapped to the same value of T2 point. Now, if I look into this portion here will they also under T will they also map to the same point? Not necessarily not map to the same point same as T2', but all their values will be same?

Student: No.

Professor Manjesh Hanawal: If T happens to be minimal sufficient statistic so, all these points are taking same values, let us say any point you take x and y here, they are going to take the

same value here, but if T is a sufficient statistic by definition, they should also be taking the same value. So, all these points that are here they are also taking the same value under T even though I have partitioned incorrectly it is not necessary. So, then there should be region here under T which incorporates all these points some region here, which incorporates all these points that are mapped to the same value under t prime. And that is what we are saying.

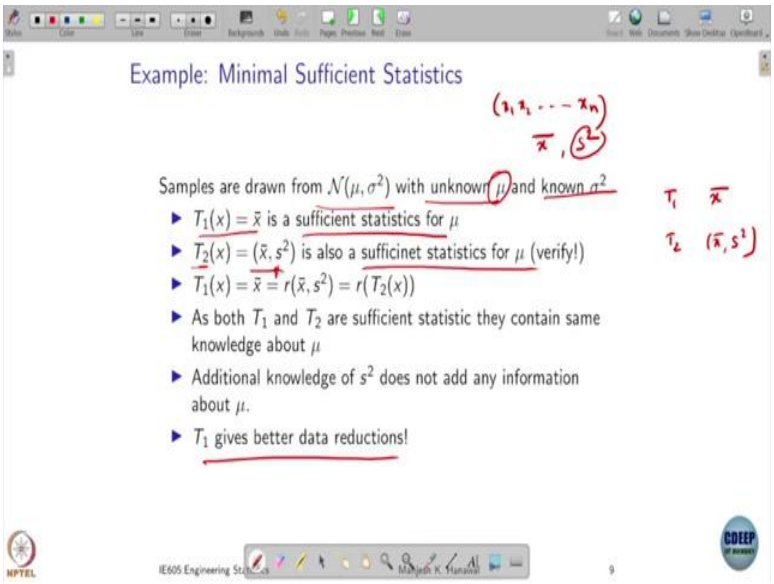
So, this region here, which has become a partition, this is going to be a subset of another partition under T maybe ideally, I should have drawn this like this it would have been if I had to draw this first figure with respect to this this should have been this maybe this like this, not even this and this, so how many partitions had got now?

So, this has 123, 123 and here, I wanted to have only maybe I want only this much now it has only 2 partitions in this, actually all these points are also here and that is now incorporated in a bigger partition. We just level a presentation. So, if this happens to be sorry, here it I call him b in this set. I am calling them as b, b1, b2 and b3. So, this b2' here is going to be part of a2 a2 here just by this definition that if T is a sufficient statistic every point that are mapping to the same they should also be mapping to the same value and under my T. So, obviously the number of partition under minimal sufficient statistics is going to be large or small compared to another sufficient statistics?

Student: Small.

Professor Manjesh Hanawal: It is going to be smaller, if this is some sufficient statistics we saw it is going to partition my space into 3, but if T is sufficient statistics the number of partition it is going to have is going to smaller than this. So, in this way among all the statistics with statistics is going to partition your space into a small number of subsets? Minimal sufficient statistic.

(Refer Slide Time: 10:53)

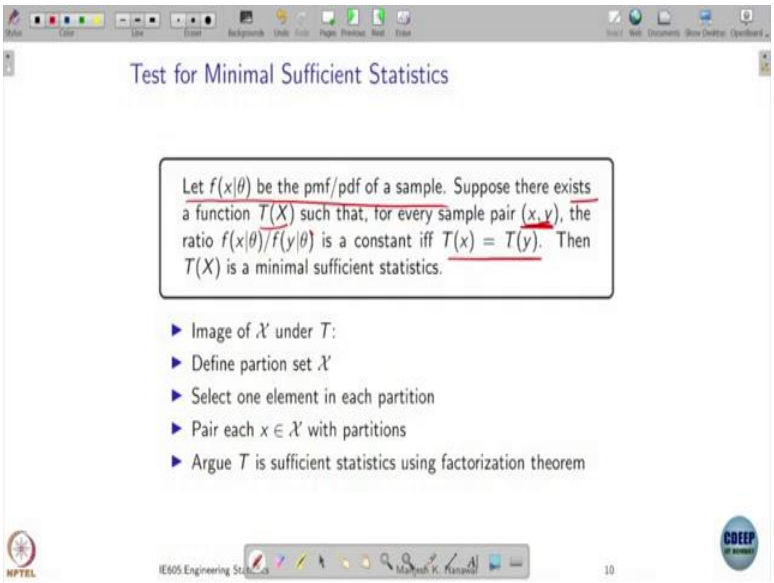


So, this example we also discussed last time suppose, we have a samples which are drawn from a Gaussian distribution with parameter mu and sigma squared with unknown mean mu and known parameter sigma square. So, only thing that is unknown to me is mu now, I know that sample mean is going to be a sufficient statistic, now if I am going to take another sufficient statistic another statistic will actually come consists of 2 components sample mean as well as sample variance.

Now, actually this is also be sufficient statistics, because every information all the information about my parameter mu is contained in this and in fact I can ignore x square and S bar I can just written which already you know is a sufficient statistics for mu but I know that T1 is a function of T2 if you give me T2 by just dropping the second component I can recover T1.

Now, even though both this statistics or sufficient statistics contain same amount of information about my unknown parameter mu T1 is a offers may a better reduction. So, because of that it compresses information better that should also give a kind of intuition that why the number of partitions is less under minimal sufficient statistics compared to a just an ordinary sufficient statistic.

(Refer Slide Time: 12:56)



So, now, how to find minimal sufficient statistics how to test a sufficient statistics is a minimal sufficient statistic. So, first of all we started statistics is any function that is fine check whether something is sufficient statistic that was not obvious, but we come up with a method, what is our method to check whether a statistic is sufficient statistic?

Student: (())(13:28)

Professor Manjesh Hanawal: So, one thing was factorization theorem which is give a four sufficient necessary condition for something to be sufficient statistic and if you have to just to do with simply verify whether something is a sufficient statistics we had some other thing also that ratio test we had, see whether the ratio of unconditional distribution and of your samples and unconditional distribution of your statistics is independent of data.

So, now we have minimum sufficient statistics something more than sufficient statistics how to test that? So, for that we have this result which again gives some complete characterization because it is giving both sufficient and necessary conditions. So, what it says is suppose, let us say f of x by theta is your population density.

Now, if there exists a statistic T, such that you take any pair any pair x and y two samples random samples and if it so happens that the ratio of have those pdfs at those two points x and y happens to be constant, when I say constant here, independent of x sorry independent of theta

here the parameter there are only 2, 3 things that are x y and theta here, when I say it is a constant it is independent of theta.

This is going to be constant if and only if, and if those two points x and y are such that my statistics use the same value on those 2 points, if this is the case, then my T is going to be a minimal sufficient statistics. First before we will just briefly discusses proof anybody has any difficulty in understanding this statement or interpreting this statement? What does this mean?

What it is saying is you give me a statistic. And what I am going to do is I take two points, two random points and check the ratio of these two points check the PDF of these two points. If this is independent of theta, provided on these two points, T of x equals to t y only then it should happen. If this is going to be independent, or theta on points x and y where T of x is not equal to t of y, then you are failing this statement. If that is the case, then T is going to be a minimal sufficient statistic.