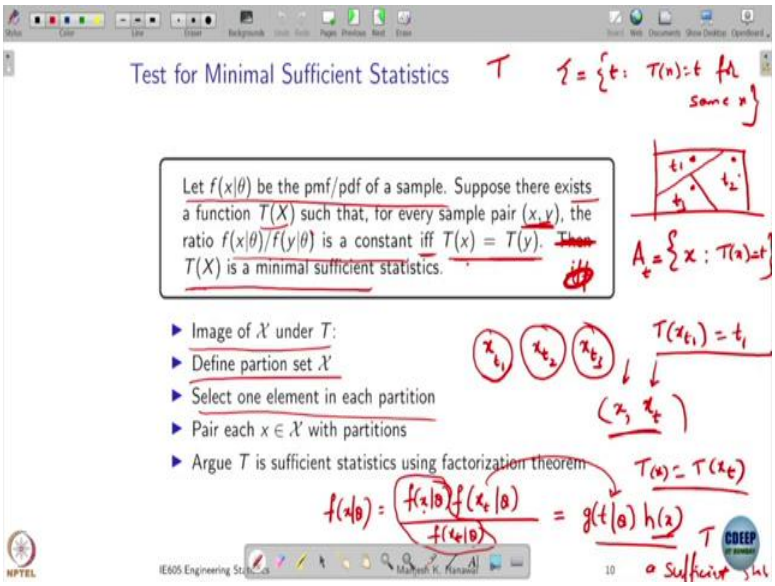
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operations Research,
Indian Institute of Technology, Bombay
Lecture 36
Test for minimal sufficient statistics with examples, Ancillary Statistics

(Refer Slide Time: 0:15)



Let us quickly and go through its proof. I just want to go through the proof because…

Student: (())(0:23).

Professor: So how you have to interpret? Let us say our claim final claim is here what whether T is a minimal sufficient statistics overall. This is our final claim. Now what we are checking is for any pair T of x, y, is this constant is, this ratio is going to be constant if and only if T(x) equals to T(y), if this is the case then we are going to talk about T is minimal sufficient statistics.

So here basically you are going to check some condition. And that checking condition itself has if and only if in it. And if that ratio is going to constant that means, so first of all that ratio is constant if T(x) equals to T(y) and on the other hand if T(x) equals to T(y) that ratio should be constant. If that is happening, then we are going to say that, so in a way you are right, like this is kind of giving a sufficient conditions in terms of an another condition, which is necessary and I mean, which has both forward and reverse direction.

if T of x equals to T y, if this is the case then we are going to talk about T is minimal sufficient statistics.

So here basically you are going to check some condition. And that checking condition itself has if and only if in it. And if that ratio is going to constant that means, so first of all that ratio is constant if T of x equals to T y and on the other hand if T x equals to T y that ratio should be constant. If that is happening, then we are going to say that, so in a way you are right, like this is kind of giving a sufficient conditions in terms of an another condition, which is necessary and I mean, which has both forward and reverse direction.

It itself is not like a complete characterization, it is basically giving a condition, sufficient condition under which T is a minimal sufficient statistic, but what is that condition, it has implication both forward and reverse direction. All of you see his point, see I am not saying this, I am not saying if and only if here, I am only saying then, that means if this condition holds then T is a minimal sufficient statistics.

That condition holds means basically I am giving a sufficient condition and that sufficient condition itself is characterized in terms of both forward and reverse condition. Now let us understand this, this is the first step. You give me a statistic T, what I did is I construct a partition, what I do is I look into all T such that T(x) equals to t for some x. What I did is this was my space like last time maybe let us take these are my three.

All these points here they mapped into let us say t2 all this mapped into t1 and let us call all these points here mapped into t3. So this is my image. So t1, t2, t3 are my image of my space under statistic t. Let us call this as capital tau. And now partition, we know that this already defines a partition. You take one point t here, then I know, then I have t equals to set of all points such that T(x) goes to t.

For example, you give me t1, then that set is all the points, which are mapping to the t1. Now I am going to, so in this case I have three ranges, A_{t1}, A_{t2}. What I am going to do is I am going to select three point from each one of this. What are these three points? Let us call them x_{t1}, x_{t2}, x_{t3}. So x_{t1} is coming from here, x_{t2} is coming from here and x_{t3} is coming from here. Suppose if I take x of t1, my claim is this is equals to t1.

All of you agree with this, because every point here, this x_{t1} is coming from this space, every, this point when I applied T, capital T, this has to give me. So these are, let us say these are, I am fixing these three points, let us take them to be three points. Now what we will do is take

any point x, somewhere in this, one of this, I do not know what it is and I know that this point I can pair, like with it let us call this sent some point of representative of each of this set.

Suppose if x comes from x_{t1}, then I am going to associate it with this x_{t1}. If this x come from this region, I am going to associate it with x_{t2} like that. So for each x I can associate this with some t, let us simply call it some t, which is one of these points. You give me any x, I am able to find what, which group it belongs to and find out the representative point from there and I can associate like this.

Can I do this? Now let us see, so on this point I know that T(x) equals to $T(x_t)$, you agree. So x is a point and I have paired it with this representative point from that set, so I know that if x and xt both belongs to the same set, if that is the case t will be assigned the same value to both of them. Now I want to see given a point x and theta, what I will do is I am going to divide this guy by x of xt, the representative point there, theta divided by f of xt by theta.

So what, now if I want to apply the factorization theorem here to conclude that this is a sufficient statistic, what I want to do here is I want to write it in the form of g of t given t and h of theta like this. So now if this t is a sufficient statistics, I need to write it like this and now what I am going to assume, I am going to assume that this guy is a constant, because t of x equals to t xt. So if that is the case can I say that this factor here is a constant, in the sense it does not depend on theta? Can I take that as my h(x)?

If does not depend on theta it better depend on x quantity we have. And this quantity here, now it depends through t and now I can take it as g of t given theta. Now if t of x equals to t y, what I have just argued is this ratio is a constant. And I mean, yeah, this sorry, I have used the fact that if this T(x) equals to T(y) this is a constant under that condition and I have just demonstrated that.

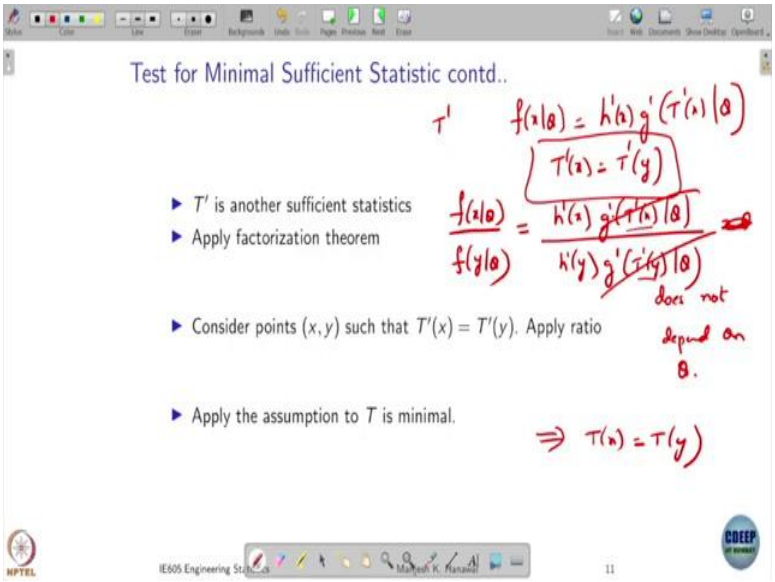
If that is the case it is a sufficient statistic. Now all I have done is I have established the fact that T is a sufficient statistic. Now let us use the other fact. What we will do is, we will now assume that this is a constant. Now I have partitioned in such a way that you give me a point x, you give me a point x and I came up with another pair with it. I mean, I came with another pair, another point to pair with this.

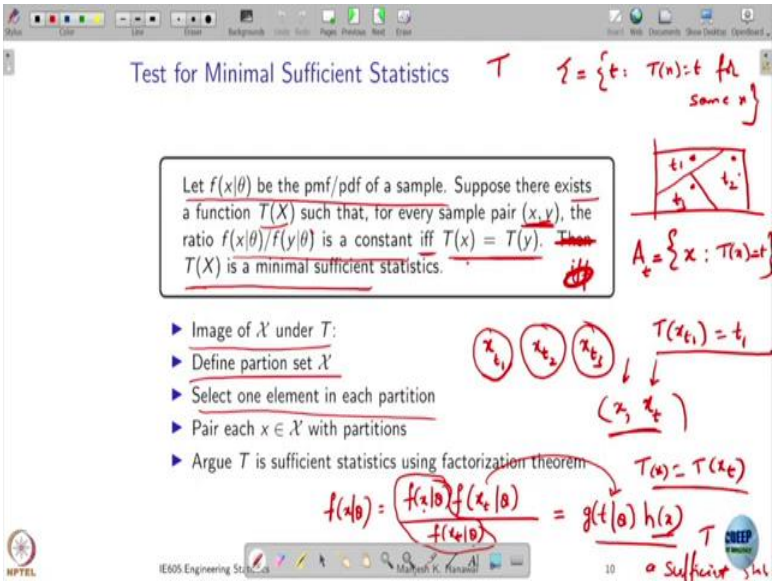
Now I know that these two pair will be such that t of x equals to t of x t. Now under this, this is, this ratio is a constant if and only if, now I have point x and y for which this is same. Because of this I can, now this is, I am assuming that this is a constant, that is what exactly I

am doing this. So I have written f of x theta here, I need to show that this PDF factorizes into g and h function, so to get that I have divided and multiply them by f of xt.

And now I am exploiting the fact that this ratio when I look into the pair x and xt that does not depend on theta that is when I use the fact and then it is… Now what I want to do is, always said that when this, under this condition I am able to demonstrate that if, I have a t which is satisfying this, I have shown that t is a sufficient statistic.

(Refer Slide Time: 11:22)





Now let us say how to show that sufficient statistics is going to be a minimal sufficient statistic now. So now let us take another sufficient statistic T' and we know that if T' is a sufficient statistics f of x theta I should be able to find a h' into g'(T'|theta). Is this my factorization theorem correctly applied here? I have simply applied my factorization theorem here.

Now what I will do is assume now I am going to assume this holds, let us say T'(x) equals to T'(y) on this sufficient statistics T' and now I am going to apply my factorization theorem and see how does this behave, x of theta divided by f of y of theta and I know that this is equals to under this h'(x).g'(T'(x)| theta) divided by h'(y).g'(T'(y)| theta).

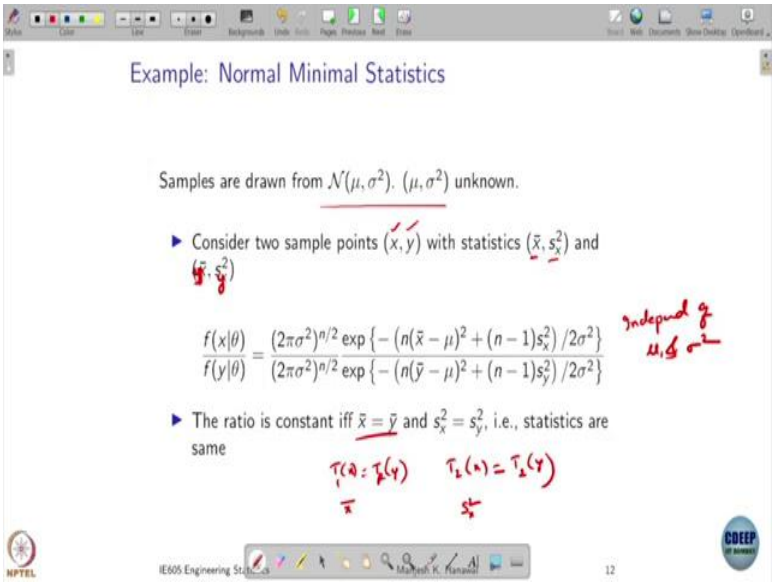
But now this points extend y such that T'(x) equals to T'(y). So this quantities this quantity are the same. g'(T'(x)) and g'(T'(y)), because both T'(x) and T'(y) are the same, so knock off it. Now it is now, this ratio is h'(x), h' does not depend on theta and I have h'(y) which also does not depend on theta. Now can I argue that this ratio does not depend on theta?

And we know that if this is the case, now we have just demonstrated that this ratio is a constant does not that depend, so this should imply T(x) equals to T(y), on these two points. So we just demonstrated that there exist a sufficient statistic, so first of all we argued that if there is a statistics, which satisfies this condition in the forward direction, T is a sufficient statistic.

And I said that if there is another sufficient statistics for which this ratio holds, I just argued that this ratio does not depend on theta. And now if this ratio does not depend on theta we have this condition T(x) equals to T(y), this implies now what, we started with this and we argued that this implies T(x) equals to T(y). So if T'(x) equals to T'(y)we are arguing that T(x) equals to T(y).

What does this mean? T is a minimal sufficient statistics. So do not get lost in this, that is what I did not discuss this yesterday, but go back and you just see that like how we are exploiting the sufficient status, sorry, factorization theorems here in deriving all these properties.

(Refer Slide Time: 15:17)



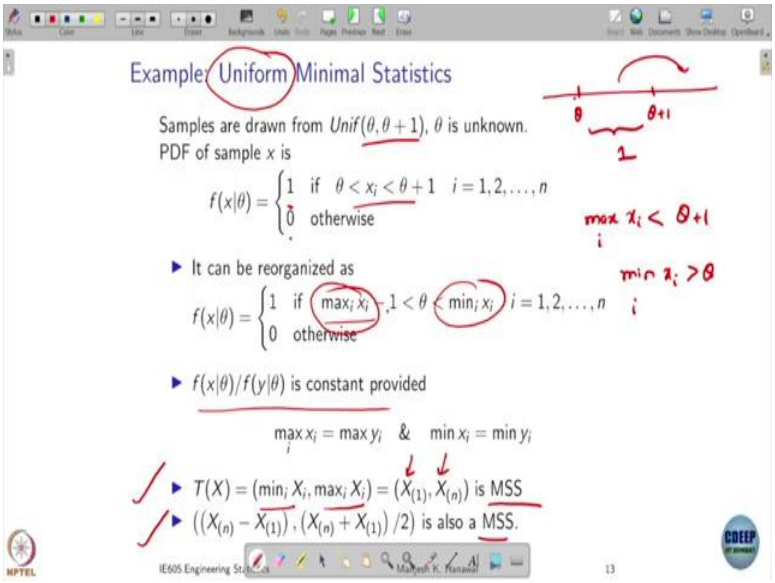
Let us quickly discuss this application. Now suppose I have some samples drawn from Gaussian distribution and where both mu and sigma square is unknown and what is one

possible sufficient statistics for this, sample mean and sample variance. Let us say I have two points x and y for which I have this sample mean and sample variant and another, this should have been y here.

And now let us see if I have these two, I have the sufficient statistics, which is computed and the two points x and y and let us compute this ratio. If I take this ratio we know how to write the joint PDF in the case of Gaussian distribution here. You can write this expression and you will see that the only way this guy you can make it independent of mu and sigma square by setting x bar equals to y bar and Sx^2 equals to Sy^2.

If they are not equal there is no way you can make it independent of mu and sigma square. So this ratio becomes independent only when x bar equals to 1, that is your T of x, the first part is, sorry, T1 equals to y and the second part is also same on those two points. So here T1 is your x bar and T2 are sample variance.

(Refer Slide Time: 17:18)



Now one more quick example we will talk about uniform minimal statistic. So this instead of Gaussian now we will be looking into uniform. Let us say I have a uniform distribution but the parameter is theta and theta plus 1. Now I have this uniform distribution, but now I am looking into uniform, the interval length is 1, but it is starting at theta and ending at theta plus 1. Now I can write the joint PDF as it is going to be 1 if all the samples are between theta and theta plus 1. If any of this sample is outside this theta then the joint PDF is going to be 0, that is obvious.

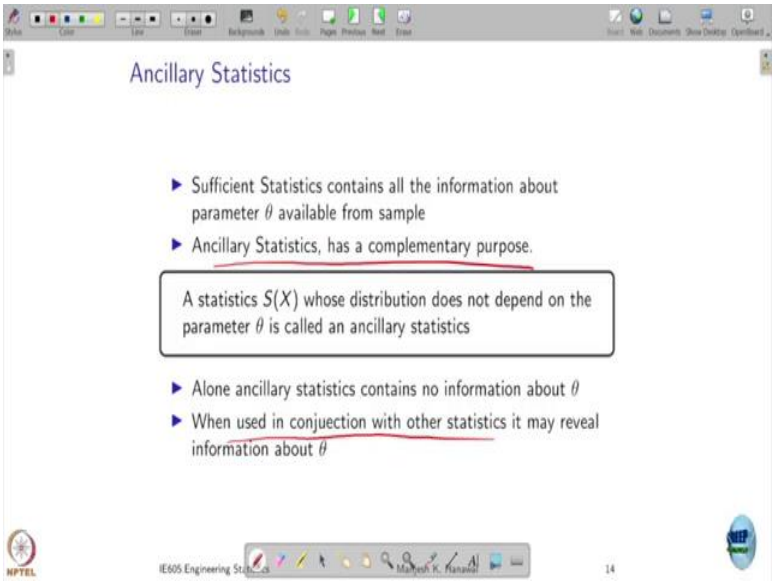
Now I can manipulate this. If I have to take all the samples and take max value of them, do you think the max value is going to be outside theta plus 1? No, it has to be below, so that is why this max value of xi is going to be less than or equals to theta plus 1. So this max value has to be less than theta plus 1.

And similarly if you take the minimum value of all the samples, can this be smaller than theta, no it has to be greater than theta, so we can write theta to be minimum these samples. Now you can check that if you take the ratio of this joint PDF at point x and y, they will be constant if and only if these two quantities are the same at points x and y. So this will give me a hint about what could be the possible minimum sufficient statistics for this uniform distribution.

It says that if you take T(x) equals to minimum of this maximum of that, that is what is minimum of xi, we called it as first order statistic and max value as, that is the n'th order statistics, that is the last one, then that becomes your minimal sufficient statistics. And you can also check that instead of looking at the minimal, maximal, you can look into the difference between these two that is x(n) – x(1) and their average of the maxima minima value that is also minimal sufficient statistics.

So I am saying that this is also sufficient statistic, this is also sufficient statistic. I am giving you two sufficient statistics for uniform distribution, then minimum sets, then what does it imply? Can minimal sufficient statistics be unique? No, so there could be multiple sufficient statistics and there could be multiple minimal sufficient statistics.

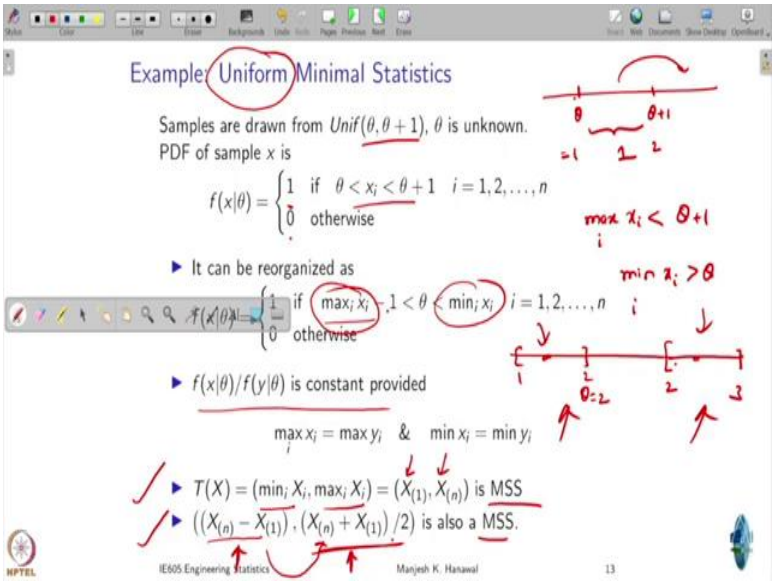
(Refer Slide Time: 20:29)



This last one slide we will cover about this statistics and this will conclude our discussion on sufficiency principle. So we just said that all our focus so far have been constructing a statistics which contains all the necessary information about the parameter of interest. But here ancillary statistics something which is not providing information about the parameter of my interest, but something complement about that.

So our statistics S(X) whose distribution does not depend on the parameter is called sufficient, sorry, ancillary sufficient statistic. So if I just give ancillary sufficient statistics, it does not provide maybe any information about your parameter theta, but when it is used in conjunction with other statistics, maybe it will provide more information.

(Refer Slide Time: 21:34)

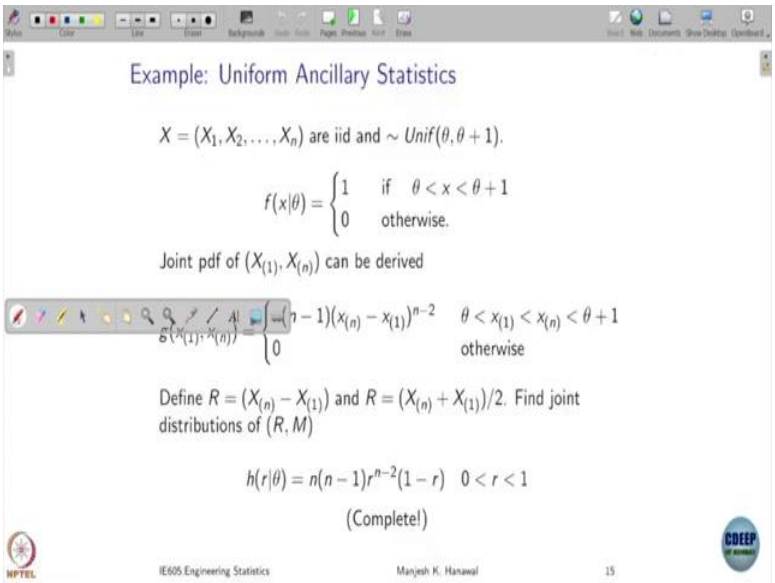


Just to give a quick example, I said, what does X(n) - X(1) give you range, but if I just give a range does it tell you information about theta here. So here let us say theta equals to 1 and then it becomes 2. I could take theta equals to 2. So in one case the range is 1 to 2 and in other case range is 2 to 3. So the first one is telling you just the range information, though sample could be coming from here or here, but on the other hand what this is giving you.

X(n), what this part is giving you? Average. So will this average, if I am going to take the average of this sample it is always going to lie in this and if I am going to take the samples of this interval they are going to align somewhere here. So to know if I give a sample whether it belongs here or here, is it enough to give me the range information. No, I need to give you some mean value also where potentially it can lie.

So here this X(n) minus X(1), can I treat it as a ancillary statistics about my parameter theta, but if I conject, if I use it in conjunction with X(n), this average here, it is providing me full information about my parameter theta and about the distribution. So in fact, it becomes minimum sufficient statics in this case. So that is the difference between your ancillary sufficient statistics.

(Refer Slide Time: 23:28)



So there is an example. I will leave you to look into this example because it has some more computation in this. So with this we will conclude our discussion on the statistics. So I hope we will all have understood what is the statistics, what is the sufficient statistics, what is sufficient statistics, how to check for sufficiency and then how to check for minimal sufficient statistics and at the end is ancillary statistics. Thank you.