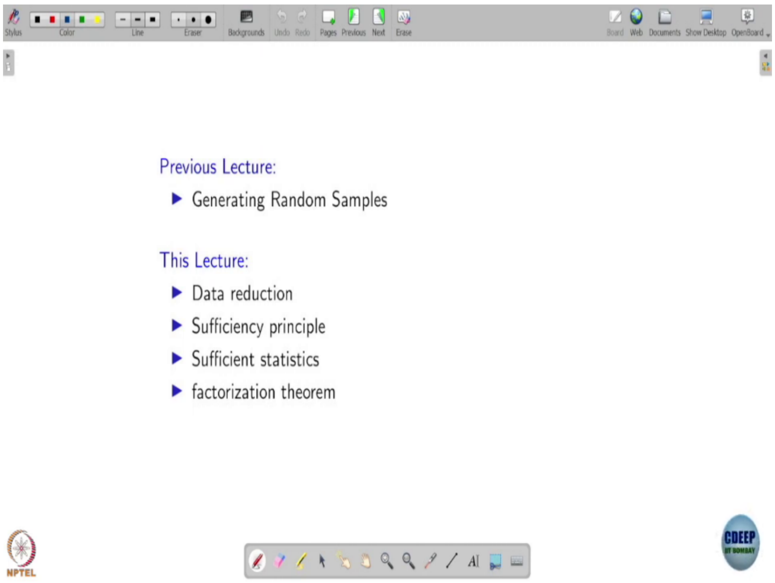
Engineering Statistics
Professor Manjesh Hanawal
Department of Industrial Engineering and Operation Research
Indian Institute of Technology, Bombay
Lecture 31
Sufficiency Principles and Sufficient Statistics

(Refer Slide Time: 00:15)



Previous Lecture:

- Generating Random Samples

This Lecture:

- Data reduction
- Sufficiency principle
- Sufficient statistics
- factorization theorem

So, last time, we talked about this generating random samples, where we talked about both direct and indirect methods. So, by the way, how many of you were able to solve the question on the indirect method in the mid sem, I think many of you got confused, where are deliberately ignored whether there is a term indirect or not there.

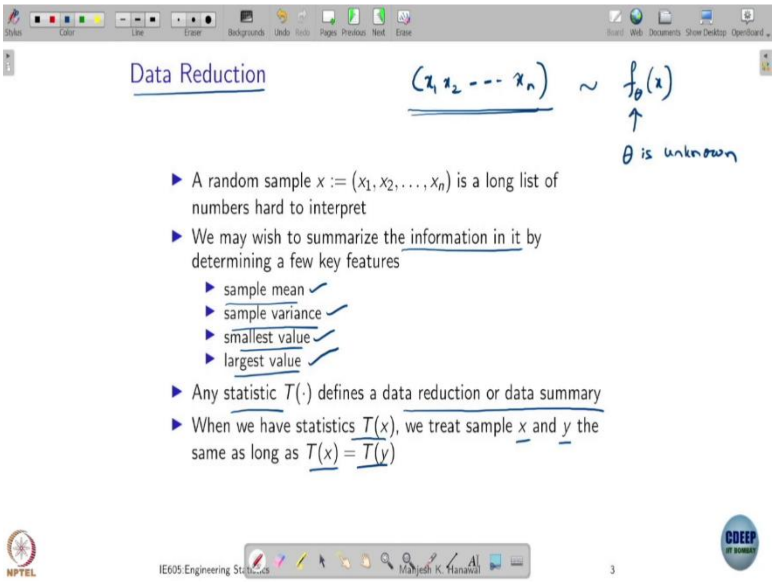
So, when I say direct method things become very simple? If I give you PDF or like a serious function, if I give you a cdf function that is it, when you just invert it direct method one can easily get. But the challenge there is we may not always be able to invert. And question I gave, that was related to which distribution?

F distribution. Was that easily invertible? No, right, it is a little complicated. That is what one has to go with indirect method. And that is what I deliberately wrote that word indirect there, but many of you are miss this and also some confusion, because the way TA is evaluated that but that is fine. So, basically, there we talked about if given a pdf or rather cdf how to generate the samples. And now, we are going to switch gears and start talking about data is already available to us. And from that, how to infer the underlying cdf which is going to generate and the most often what we will do is we will assume the class of the cdf itself.

We will assume that, this samples are coming from a Gaussian distribution, or Poisson distribution or something, but we do not tell you the parameters. So, unless you do not know the parameter, you do not know exactly what is the distribution. So, that is what we will do now, we will say that data is given to us.

And it we will be told a priori that this is going to come from this class, but I will not just tell you the parameters. And your job is to identify those parameters from the data. And for that, we are going to look something called data reduction. And there are different principles there. We will start focus most of our time on sufficiency principle, and something called sufficient statistics. And today, we will also cover factorization theorem.

(Refer Slide Time: 03:12)



Data Reduction

$(x_1, x_2 --- x_n) \sim f_\theta(x)$

↑

θ is unknown

- A random sample $x := (x_1, x_2, \ldots, x_n)$ is a long list of numbers hard to interpret
- We may wish to summarize the information in it by determining a few key features
  - sample mean ✓
  - sample variance ✓
  - smallest value ✓
  - largest value ✓
- Any statistic $T(\cdot)$ defines a data reduction or data summary
- When we have statistics $T(x)$, we treat sample x and y the same as long as $T(x) = T(y)$

IE605:Engineering Statistics Manjesh K. Hanawal 3

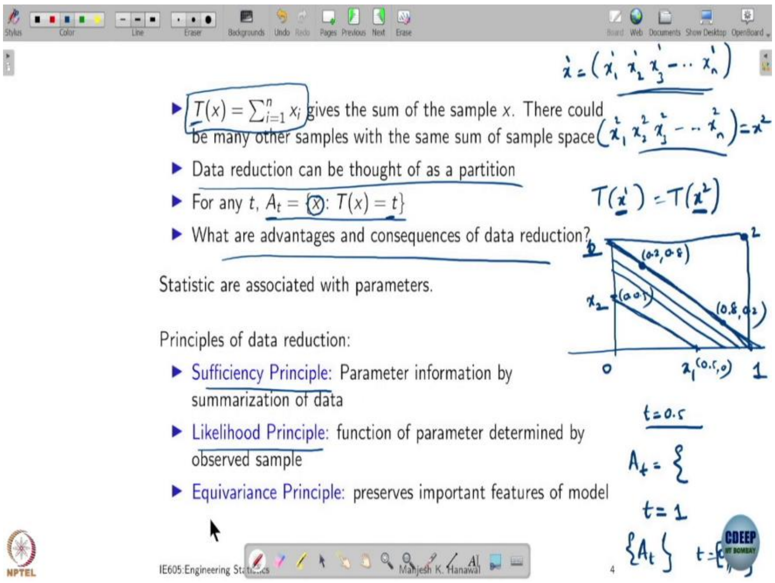
Now, for us, we are going to assume that data generated let us say X_1, X_2. These are the observed data and we are going to say that this data is going to come from some underlying population f_θ and I do not know what is θ, the structure of the f_θ may be known. For example, I may say that f_θ is a Gaussian distribution, then you know its structure. So, now let us say you how this data points. Now, this is just a bunch of data for us. Now, let us now focus on data reduction part. Data itself of no value to me, but what is important to me is the information provided by the data.

So, you may be interested in obtaining some key information from this data either by defining sample mean, sample variance, or like smallest value or largest value. And to define sample mean, I am going to use all the values. So, is sample variance. Similarly, to define smallest value, I am going to use all the values.

Now in general, we can talk about any function of the samples, which we earlier called it as statistic, T is any function which can operate on my data samples, we call it a statistic. And we said that sample mean, sample variance, smallest values, largest value, these are all examples of your statistic. And what they are actually giving you is a basically kind of data reduction or data summary.

Mean is telling you one summary, sample variance is giving you another summary, smallest value is one summary, largest value another summary. Now, if you are only interested in this statistic, and statistic is giving me the summary and if it so happens that if two samples x and y are giving you the same summary, T(x) and T(y) which happens to be the same, then from information point of view, both x and y are same for me, because they are giving me the same information when I use my statistic.

(Refer Slide Time: 06:11)



$x^1 = (x^1_1, x^1_2, x^1_3 - \dots x^1_n)$

- $\boxed{T(x) = \sum_{i=1}^{n} x_i}$ gives the sum of the sample x. There could be many other samples with the same sum of sample space $(x^2_1, x^2_2, x^2_3 - \dots x^2_n) = x^2$
- Data reduction can be thought of as a partition
- For any t, $A_t = \{\boxed{x} : T(x) = t\}$
- What are advantages and consequences of data reduction?

$T(x^1) = T(x^2)$

Statistic are associated with parameters.

Principles of data reduction:

- Sufficiency Principle: Parameter information by summarization of data
- Likelihood Principle: function of parameter determined by observed sample
- Equivariance Principle: preserves important features of model

$t = 0.5$

$A_t = \{$

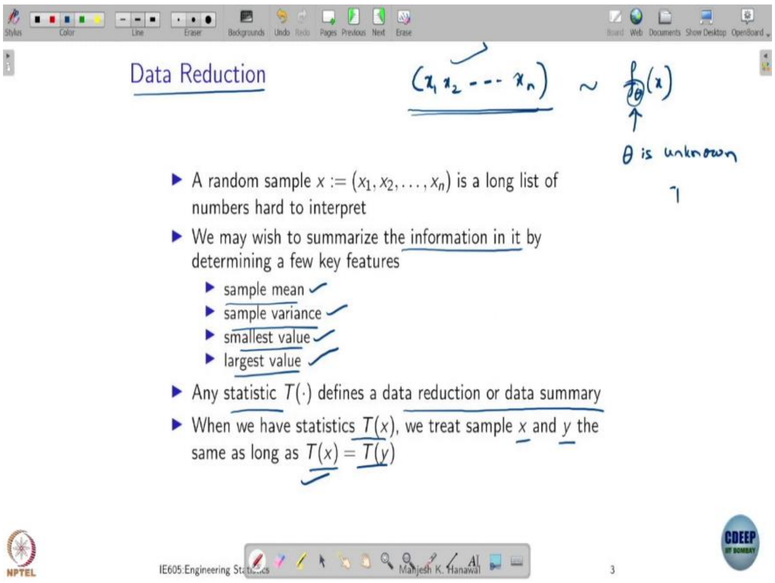
$t = 1$

$\{A_t\}$ $t =$

IE605:Engineering Statistics

Manjesh K. Hanawal

4



Data Reduction

$(x_1, x_2 \cdots x_n) \sim f_\theta(x)$

θ is unknown

- A random sample $x := (x_1, x_2, \ldots, x_n)$ is a long list of numbers hard to interpret
- We may wish to summarize the information in it by determining a few key features
  - sample mean ✓
  - sample variance ✓
  - smallest value ✓
  - largest value ✓
- Any statistic $T(\cdot)$ defines a data reduction or data summary
- When we have statistics $T(x)$, we treat sample x and y the same as long as $T(x) = T(y)$

IE605:Engineering Statistics Manjesh K. Hanawal 3

Now, thinking more about this. I am now going to take one particular example of statistic which is basically sum of all the values in my random sample. Now, it may happen that let us say, I have n samples and another bunch of samples. Let us call this maybe x and a simple x1 and let us call this x2.

It may happen that if you add all of them, they may add up to the same value. And both of them could be still coming, I mean, my assumption here is that both of them are coming from the same underlying distribution. So, what is now happening is; This samples I am summarising, by applying some statistic, in this case, some this samples also I am summarising by applying this statistic T.

Now, if they happen to give me the same value, let us say if I take $T(X_1)$ that is this and this happens to $T(X_2)$, we say that X_1 and X_2 they are, same for me because they are providing the same information. So, for me, like at this point, this random sample and this random sample, they are the same.

So, I may group all the random sample, which give me the same information. And that grouping can lead to partitioning of my space. So, last time, we said let us take a two dimensional case. Maybe this is 0, sorry, this is 1. This is 1. And I am interested in all the points. My Space is only this. This is my X_1 and this is X_2. Just let us take one hypothetical case. If I draw a line and any take one point, let us take this point. And let us take this point.

If I add the components of these two points, will they have the same value? You people understand this? Let us call this 1. Let us say this let us call this is 0.2, and 0.8. And this point here is 0.8 and 0.2. So, you take any point on this line, they will add up to the same value. So, from a data reduction point of view, if I sum that all these points on the line are the same for me, they are indistinguishable to me? So, like that, what I can do is I can now start off thinking this data reduction as the partitioning of my space itself. And in that all the points x, which will map to the same number, let us call T, I am going to collect them, and I will call that set as A_t. now let us take another point.

So, here the sum is 1, if I am interested in all the points, whose sum is going to be, let us say, 0.5, where how, on which line, they will lie. Let us consider all the points whose sum is going to be 0.5. So, here is one point, this is 0 0.5. Another point is going to be here, 0.5, 0. And so if I am going to connect this line, all the points there is going to have sum 0.5? So, like that, what I can do is? I, in this case, maybe let us take t equals to 0.5. Now if I am going to define A_t, this is going to be basically all the points on this line. And similarly, if I take t equals to 1, this is going to in this set is going to include all the points on this line.

Now, I can consider all these sets A_t. Now, what are the possible values of T? T here, what is the value of what is going to be here? It is going to be 2? If I am going to sum so maybe, I will be interested in taking T from 0. All the numbers, maybe t = 0 to 2. So, if I am going to construct all the sets where t is going to between 0 to 2.

And I take these sets A_t will they overlap or they disjoint. They are going to be disjoint? Because all of them like there will be on this particular line different lines depending on what t I am going to choose. So, that is why now if I am going to think of this, this set At now forms a partition. So, in a way the when I reduce data, the possible reductions I am going to get that is going to partition my space in this fashion.

Now, what are the advantage of this data reduction? We discussed last time right data reduction is one obvious thing is, if I tell you just the sum then I do not need to maintain the entire vector right X_1, X_2, up to X_n, I just has one value I have reduced it to one value, I just need to store that one value in that way that is going to be useful in terms of reducing but now our question will be always with respect to getting information about the underlying parameter.

So, that is our basic principle through which we started like when I started talking about this, we say that I have data what I want is extract information about this parameter theta but this data reduction came we are going to use this data reduction as a means to get information about that theta. Now, the question here is if so, fine data reduction is fine. You have this samples you can reduce it to one number.

The question always remains is when you reduce it, how well that reduced value is capturing information about your theta parameter. For that you we are going to look into different principles one is called sufficient principle, sufficiency principle is always going to talk about the statistic, when you are going to say that you are statistics is going to capture sufficient information about your parameter theta.

So, sufficiency principle will cover this and we are going to look into another thing called likelihood principle, which is going to write your data sorry the parameter that you are interested in as a function of your data observed and then it is tried to connect how your data points are governing or like what is how you are read parameters and the data points are related in the best possible way or most likely fashion and likelihood principle will try to capture that essence.

So, there is something called equivalence principle we will not go into that it is it is just like a relaxation of this condition here equivalence principle. So, in this case we said that, x and y will map to the same statistics we say that x and y are kind of same for us, but why that should be the case. So, maybe x and y are somewhat related, may not be exactly the same. So, that will be captured by equivalence principle but we will not dealing much with that in this course.