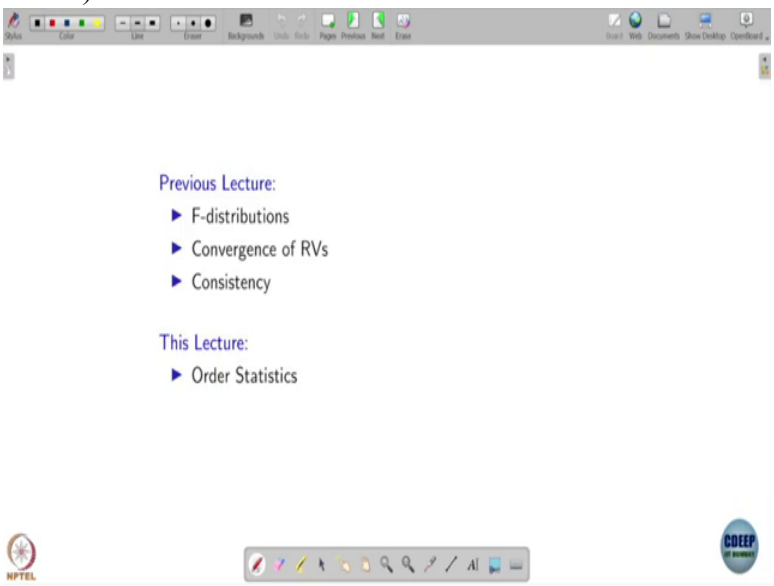
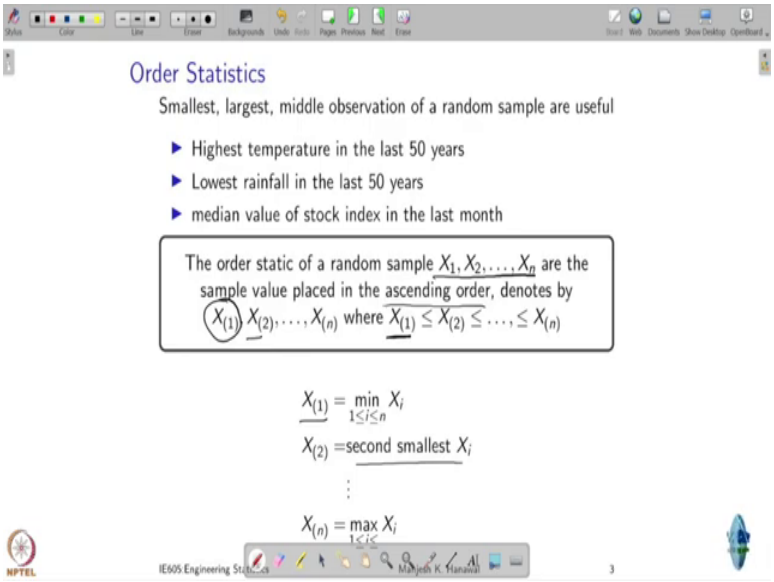
Engineering Statistics
Professor Manjesh Hanawal
Industrial Engineering and Operation Research
Indian Institute of Technology Bombay
Lecture 26
Order Statistics, Median and Percentiles

(Refer Slide Time: 00:15)



Now, I want to talk about something called order statistics.

(Refer Slide Time: 00:19)



So, as the name indicates, here I am interested in ordering of the samples. And look into the statistics of this ordered sample. So, where I should worry about ordering? So, maybe I will be

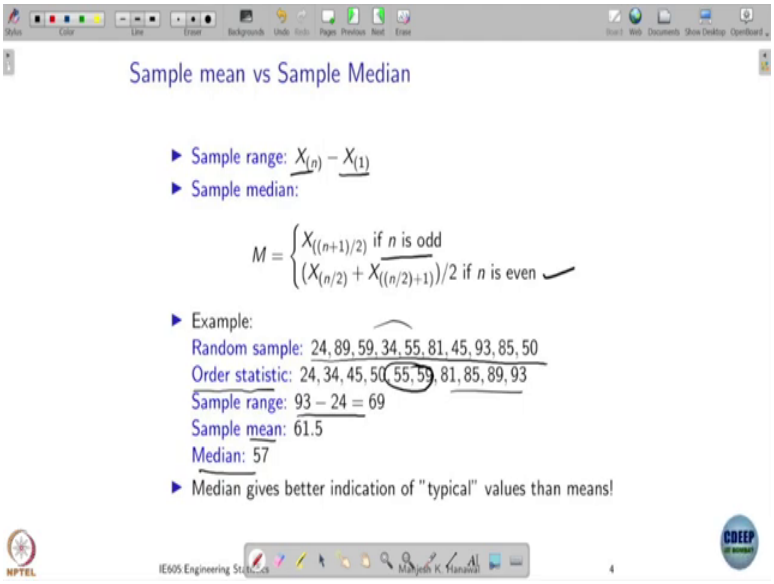
offered, if you just give me samples, I may want to find out which is the smallest, which is the largest, which is the middle of them. If I have to do this, I need to order them.

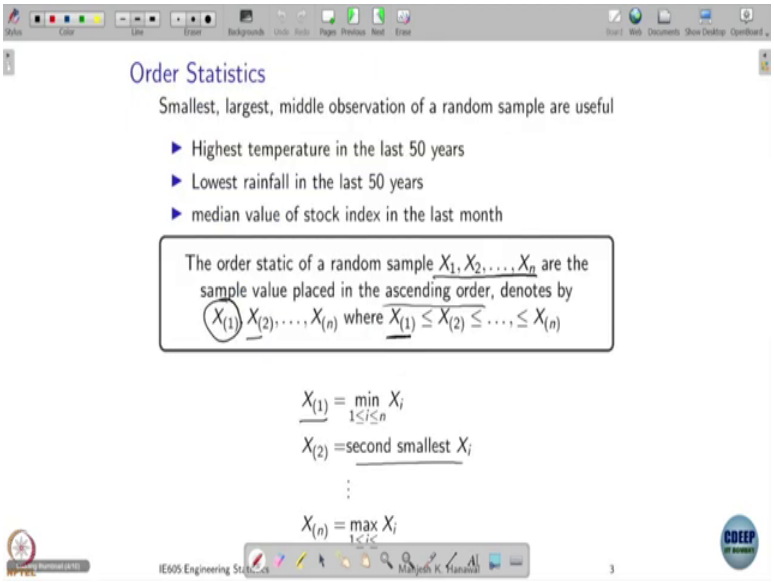
If you are just to give me samples, and if I want to find which is the smallest, maybe I want to order them increasing or decreasing way and from that I will find out which is the smallest and which is the largest of them. And this is also important in many applications, for example, which is the hot day in the last 50 years or, which year got the lowest rainfall. So, you may have data like a health, this metrological department may have recorded this data of each day of each month like that, that is like a samples for you.

And from that you want to identify which was the smallest. That identifying the smallest is easy, but what we want to do is understand its distributions. Now, because now, we have to order we will use now one notation to put our samples into orders. What we will do is let us say I have n samples given to me I want to put them in an ascending order. What I will do is now I am going to take the smallest among X1 X2 Xn as X1.

Notice that, now I would deliberately put a parentheses on 1. So, this X 1 is minimum of all my samples and the next X2 is the second smallest among all my samples and that I have calling it as X2 and like that. And naturally this X of 1 with the superscript has to be smaller than X of this 2 in parentheses like that. So, I have n samples, I have simply took the first smallest one put it first, took the second smallest one put it second like that.

(Refer Slide Time: 03:12)





Now, putting them in this way will already help me find some information about the data. First thing is sample range or what range my values are being taken. So, if I take my largest value and smallest value, it will give me indication of the range that is right. Largest is this, smallest is this matlab my samples are taking value on this range.

Next something called median. So, suppose let us take n, my n is odd number. So, n plus 1 by 2 is, is going to be an even number. But n plus 1 by 2 is going to be exactly the middle, exactly the middle in my sequence. And that is what is going to be called as median. Like when I in, order my samples in increasing order and one at the middle we are going to call it as median.

And when n is even n plus 1 by 2, I cannot get an integer. So, for that what we will do is we will take the average of the two middle terms. So, when n is even I can take two middle one and then take the average, that is what I am going to call it as median.

So, in a way median gives me information about what is the middle value of my samples, at 50 percent, what is the 50 percent value of my samples. So, I have an example. I will just put some numbers here some random samples, ten samples I have here. And their order statistics, I will just put them in an increasing order, I have just put them in increasing order. And from this it is easy to compute what is the sample range. So, 93 minus 24 and sample median, sample median. So, how did I get 61.5 here? Oh, no, this is sample mean. So, how to get this sample mean?

Student: Add on.

Professor: Just add this 10 numbers and divided by 10 that is 61.5. And what is the median here? So, to apply find out the median I just to first find out whether I have odd number of samples or even number of samples. In this case, I have even number of samples.

So, I am in this category, I will just take the two of the middle ones and their average is going to be 57. And that is why I got median. And usually, this median gives better indication of the typical values of my samples rather than the mean. So, mean here is giving you 61.5 whereas the median is giving you 57. Now, our what we are saying is if I have to take the typical value of this random samples I have, we are going to, we are saying that median is a better representation of that rather than mean value. So, why is that?

Student: Because the last four terms are different than in term (())(07:28) (())(07:37) that is why median is the right value because it gives a, it given an information about the central point, whereas the mean (())(07:50).

Professor: So, is there a case where I should go with mean, rather than taking median? And is there a case where I should prefer median over mean?

Student: In case where the data is not (())(08:06) you can use in.

Professor: Kyu right like? Let us say, I am interested in the average, I am interested in the performance of this class. And I know that some of you are going to do very well, close to 100 percent. So, you are going to boost the score of this class? And if somebody is going to just ask me tell me, what is the average, then I will just tell the average even though some people failed, but average is very high.

Student: Sir, that is why, that is the reason we cannot gather solution of majority of the class because some people (())(08:43) majority lies, understood these four in the whole scheme.

Professor: So, for that, why should I go with median?

Student: Because that gives a more, accurate idea about the most of the trans. That is why, so we should go with the mean, because the mean will give the false (())(09:06) find that (())(9:08)now.

Professor: But let us suppose now. So, maybe like that is what like if I want to just to make a binary decision, if somebody who is getting above mean value pass, below mean is fail, then you will fail?

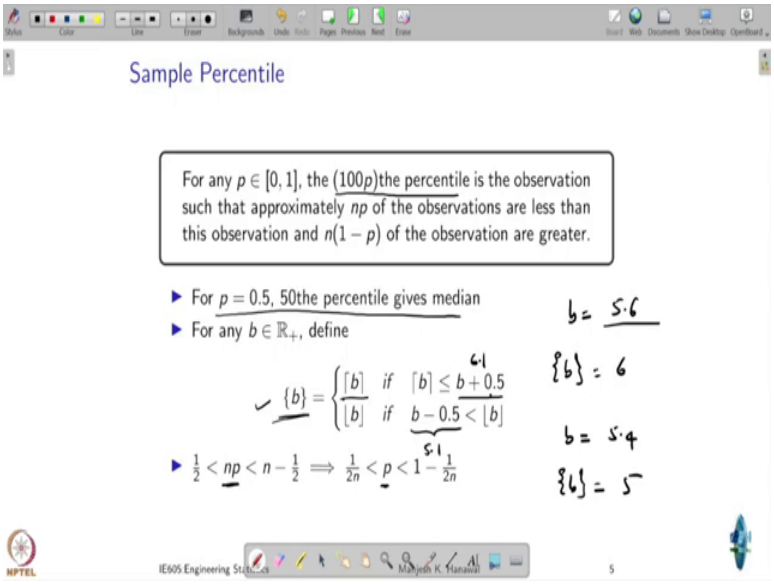
And if I say median then you will, you will fail or pass? So, that is how like come in if I want to that is what like in a classes like this, where I have to look into the overall. I mean, where I care about majority, not like only few, maybe median could be a better score. But on the other hand, let us take an another example. Instead this class let us say I put money, I may stock, I am interested in investing in stocks, and I invest in three, four stocks. All I care is on an average how much money I make. In that I should prefer mean or median?

Student: Mean.

Professor: In that case mean reward, I am interested in just reward like how the average, I will just put everybody I do not care like how they are performing individually, what matters is together how much return they give me. In that case, maybe I will go with mean. In this case, like where I have to worry about social things like, I mean, whether most of them will pass or not, I have to worry such kind of maybe median and I will go with.

So, depending on your application, mean may be suitable or median and may be suitable. So, you can choose as per your applications. But why to focus on 50 percent? We can focus on anything you want. So, that is where the sample percentile comes into picture.

(Refer Slide Time: 11:11)



Now, for any p between 0, 1, 100 pth percentile is observation such that approximately, np of the observations are less than this observation and remaining n 1 minus p of the observations are greater. And it is just like a generalization of the median we focused about. Median actually

looked into the 50th percentile. But I should, I may be worried about 40 percentile, 30 percentile, 99 percentile, or 98th percentile, and all of that.

So, that is what this is just like a generalization of that. And to just make this more formal, just like additional definition, we are just introducing this another term flower bracket b, this is going to take the value of c if it is b plus 0.5. So, we will just it is an indication that it is like suppose, let us apply a definition. Let us take b equals to some 5.6. So, what will be the value of b? As per this definition? What is a b plus 0.5? B plus 0.5 is?

Student: 6.1.

Professor: 6. 1, and b minus 0.5 is 5.1. And what is a floor of b? No sorry, ceil, ceil is going to be 6, is 6 less than 6.1?

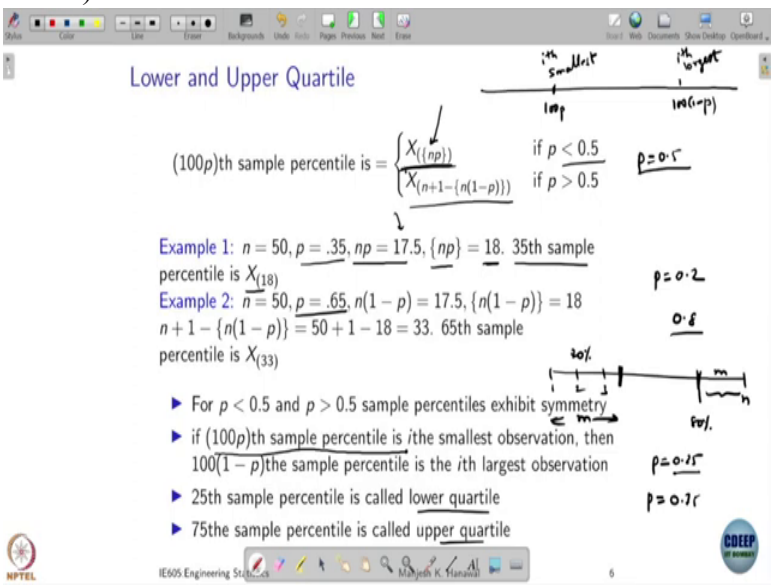
Student: Yes.

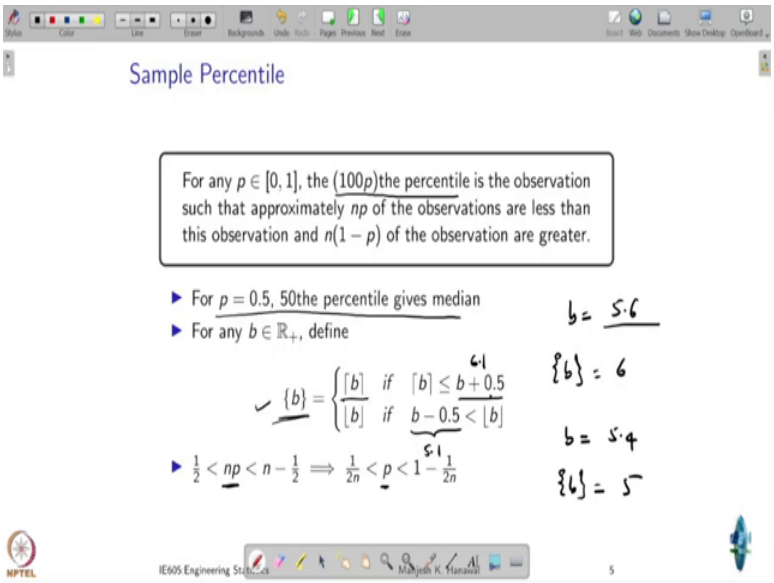
Professor: So, I am going to take this to be?

Student: 6.

Professor: 6, so, it is just like if it is more close, this is more closer towards 6. So, I will take it as 6 and if this is 6 and if this is 0.4, then I would have taken b to be 5, it is simple notation for that. And now, I want to identify suppose you have some p given and you have n samples. If np has to be between half and n minus half then you can find out that that P is simply 1 upon 2 n 1 minus 1 upon 2n. If you want the p if you are np to be between half and 1 minus half. Let us say I do not know why we are all using this.

(Refer Slide Time: 14:07)





Now using this, if I am interested in the 100th p sample, our percentile or pth percentile, it depends on whether p is between 0.5 or less than 0.5 or greater than 0.5. If your p is less than 0.5, p over value is going to be exactly this order statistics. And you notice already I have defined what is flower bracket of np? It is approximation to the next immediate integer. And if it is greater than 0.5, again I have to go with this adjustments, to make sure that it, it is more towards the right hand side.

And if it is p is less than 0.5 it is more towards left hand side and if p is greater than 0.5 more towards the right hand side. So, I have to make up for great corrections here. So, this is just a

definitions formal way of putting it. But you I hope you understand how to compute this 100p percentile now.

Now just simply put p equals to, now, this is for p 0.5 is strictly less than 0.5 and p greater than 0.5. And what how to find it for p equals 0.5? Just use the definition of median, median is exactly for p equals 0.5. And here I have just given one example. Suppose, n equals to 15 and you want to compute 35th percentile, then first we will compute what is np. And if you compute this flower bracket of np even though notice that it is in the middle, it is going to take the next 18, next value 17.5 is approximated to 18. And the 35th sample is simply x of 18^{th}. The 18th sample when you have ordered them 1 to 50.

And similarly, you can also compute what is the 65th percentile of this sample. In this case It so, happens that 65th percentile is going to be x of 33, 33rd element in my order statistics. So, this p less than 0.5, p greater than 0.5 samples percentiles exhibit symmetry, I hope you understand what is symmetry. By symmetry I mean suppose you take 0.2 and another value 0.8. So, 0.2 sample is certain numbers away from your left side. The same number of samples it will be away from right hand side for the 80th percentile.

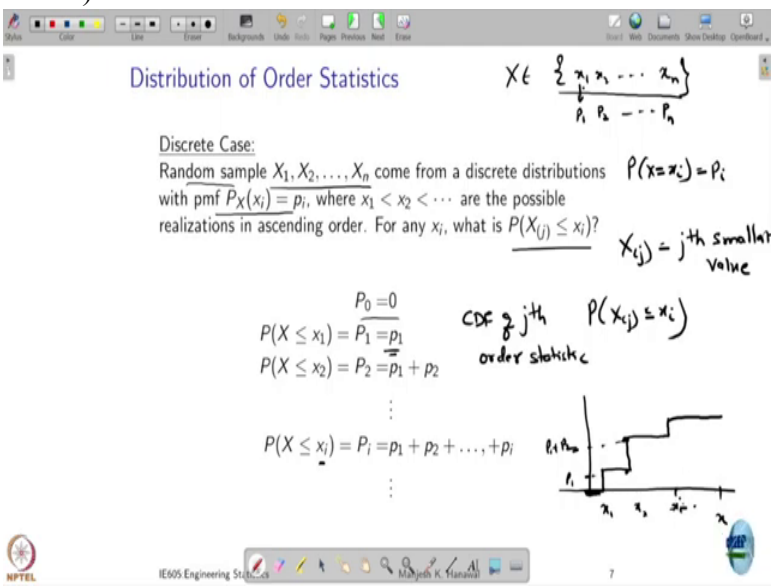
So, if you just take this or let us say 1, 2, 3, up to let us say some n and this is some here, if this is at 80th percent and it is m sample away and if it is a 20th percentile I mean at this point that means that will be also m sample away. In that way they exhibit symmetry.

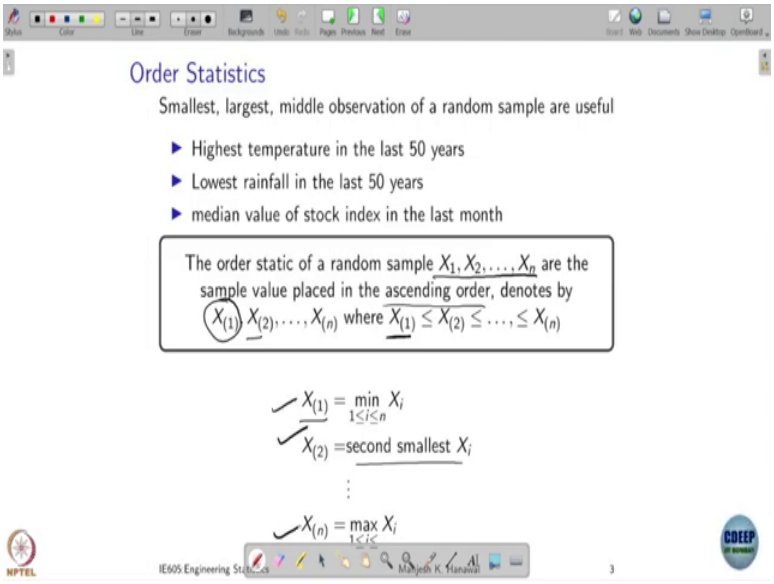
And if you want to generalize it, if suppose, for a given p 100pth percentile happens to the ith smallest observation like 100pth sample happens to be the ith, let us say if you have this and this is like 100pth sample percentile sample and this happens to be the ith smallest. Then if you look into 100 into 1 minus p here, assuming that the p is less than 0.5 here then this is going to be what? This is the ith largest element. In that way, they are also showing some symmetry here.

And now the special cases for us are something called lower quartile and upper quartile. So, when you are going to take p equals to 0.25, that number is called lower quartile. And when you take p equals to 0.75 this is called upper quartile. And notice that by our definition, they are symmetric. If p equals to 0.25 corresponds to the ith smallest, then p equals to 0.75 corresponds to the ith largest element in my order statistics.

And after maybe we will see later something called boxplot. When you have given a bunch of samples, you want to see that what samples lies within 25 times 75th percentile. And that means there is the ones where majority of the samples I am interested in. Maybe others are either too low or too large. So, I may be interested in just between something between 25 to 75. So, that is why this lower quartile and upper quartile comes. And maybe in Python, there will be already ready function. If you give a bunch of samples and tell identify 25th quartile, or 75th quartile, they will immediately find the sample value for you.

(Refer Slide Time: 20:25)





In the next five minutes, we will just discuss how to find the distribution of this order statistics. Now let us focus on this some calculations. So, what we want to now do is suppose, consider my discrete random variable. And I am interested in n samples which are coming from the same population mean. And assume that those are going to follow some common probability mass function.

So, all my X are taking value X1 X2, let us say some X n value n samples they are taking. And I am also assuming that this value is taken by my random variables are themselves ordered. So, this is the smallest value, it can take this is the next smallest value and like this. Now I am interested in what is the probability that my jth order statics is going to be taking value less than or equal to Xi. So, I hope you understand it what I am trying to say here.

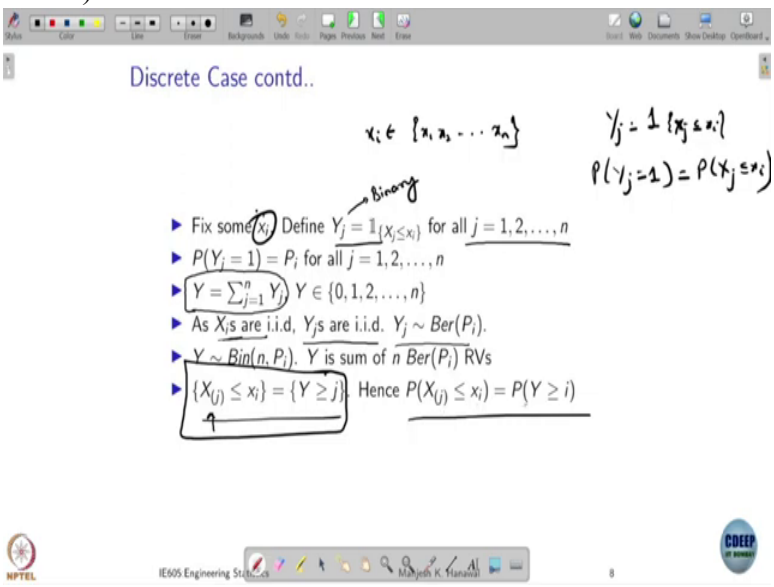
So, X, all my X1 X2 X3, they are taking values realization X1 X2 X3, and they it taking that probability that X is taking value Xi, I am going to denote it as Pi. Now, by definition, X of j is jth smallest value. Now I am interested in finding this jth smallest value is less than or equals to some particular value Xi. This Xi is one of these elements in this. Now what I am basically trying to find is I am trying to find cdf of jth order statistics.

So, notice that here, this is called first order statistics, second order statistics. And this is the nth order statistics. And now I am interested in finding the distribution of jth order statistics, how to do that. To do that, I am going to start taking P0 equals to 0, I am start doing iteration. And now, probability that X is going to be less than or equal to X1 is P1, everybody agree with this? So, notice that probability that X 1 is P1?

So, maybe I will just write this. So, these are ordered. And I am going to take this value with probability P1, this value with probability P1, P2, and this value equals to probability P2. So, it is like X1 X2, like X1, this is like P1 and this is like P2 like this. Just try to follow. So, probability X is less than or equal to X1, nothing is there below. So, I takes 1 it is going to get a jump of P1. And this is like a discrete case we have.

A probability let us tell X, X is less than or X2 this point and this point here, this is P1 plus P2, and probability that X is going to be less than equal to to till Xi is like if you are going to take this as Xi. So, this is going to be all the way P1 plus P2 all the way up to Pi.

(Refer Slide Time: 24:39)



So, now, take some Xi. So, Xi is from one of my outcomes X1 X2 up to Xn. And now Yj is indicator function that my jth random variable is less than or equal to Xi, I hope all of you understand this indicator. So, what are the possible values of Yj?

Student: 0, 1.

Professor: 0, 1 and I am doing it for my all end random variables which I have X2 j equals to 1 to n. Now, what is the probability that Yj equals to 1. So, now, since Yj is indicator that Xj is less than or equals to Xi, now probability that y j is equals to 1 what is this probability? Probability that X j is less than or equal to Xi agree? So, this is going to happen when only when this event is happening X j is less than or equal to Xi.

Now, let us focus on this Y, which is now defined as summation of this Yjs. Yj is a binary random variable. It is going to take either 1 or 0. Now, then what is capital Y is going to take values? It is sum of n?

Student: Bernoulli random variables.

Professor: Bernoulli random variables, so that is going to take either 0 or 1 all the way up to n. Now, if all these Xi are i i d you can if X i is iid you can say notice that Yjs are also iid. And I just argued that Yj is nothing but Bernoulli with value Pi. Now, this Y here is sum of Bernoulli each with the same parameter Pi, because I have fixed this Xi. So, now, because of this since y is

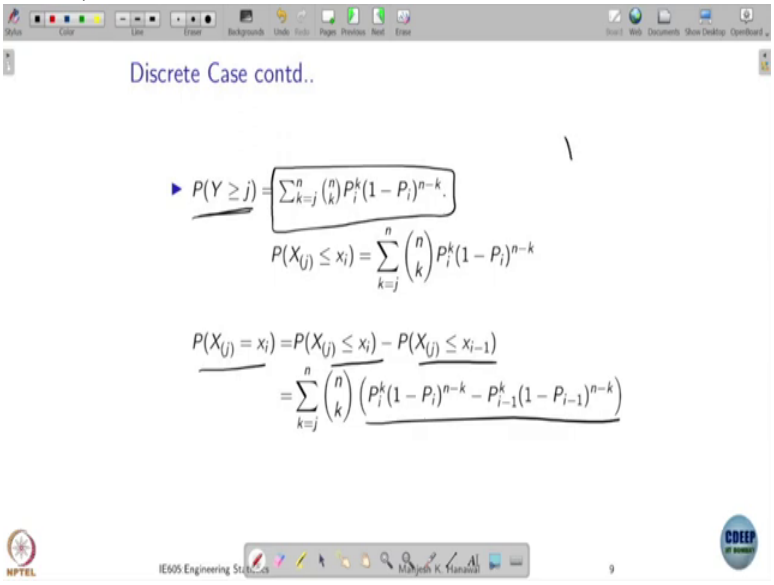
sum of n Bernoulli random variables with parameter Pi, I know that y is binomial with parameter?

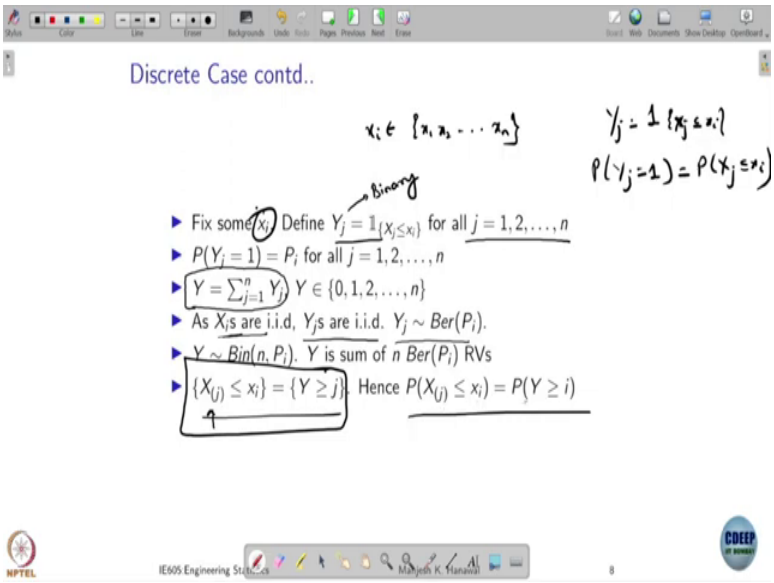
Student: n n pi.

Professor: n n pi. Now, this is true. Now, this is the last one crucial step. Suppose, X j that is my Jth order statistics is less than or equal to Xi, the claim is y has to be greater than or equal to jth that means, X j, jth order statistics is less than or equals to Xi. That means, at least in this sum Y there has to be j once that is why Y is going to be greater than or equals to j.

And then, if Y has to be greater or equals to j, you can verify that that actually means that X of j has to be less than or equal to X, you just verify that. I mean you need to think about this, these two events are same, if that is the case, probability that X j jth order statistics is being less than or equal to X is nothing but probability that my Bernoulli random variable is greater than or equal to i.

(Refer Slide Time: 28:27)





Now, you know, what is why is Bernoulli is greater than or equal to j you know how to compute its probability. And now, X j is a discrete now, I am trying to find what is the probability that Xj is equals to i. I know this is nothing but X j is less than or equal to Xi minus Xj is going to be less than Xi minus 1, this is again because the discrete one.

And now, using this relation, like I know that this is nothing but in terms of probability Y greater than I, I use this property and now I can compute this, this and plugin to get what is the probability that my jth order statistics is Xi? Verify this I know that you need to peacefully sit and read this and you will see this all this calculations come out. Fine, let us stop it.