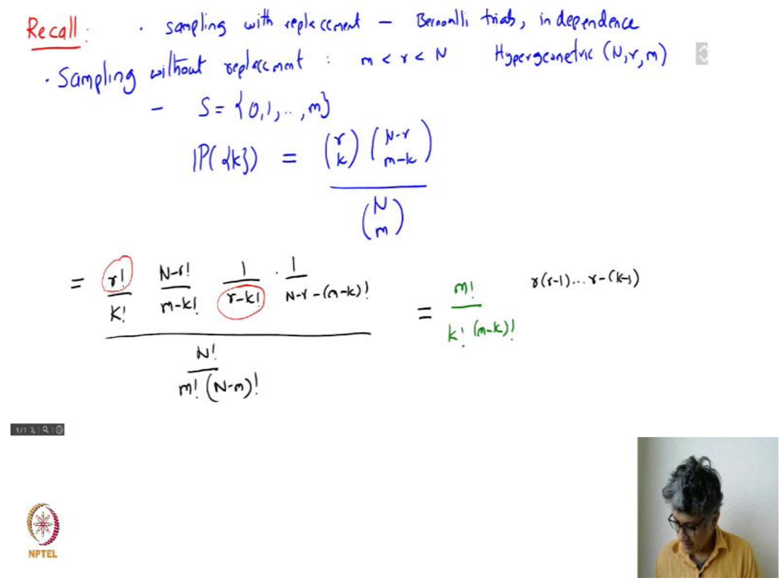
Introduction to Probability With examples using R
Professor Siva Athreya
Theoretical Statistics and Mathematical Division
Indian Statistical Institute, Bangalore
Lecture - 13
Hypergeometric Distribution and Discrete Random Variables

(Refer Slide Time: 00:44)



So, what I will do now is, I will just, just recall from last time. So what did we do.We were just, we had this concept of sampling without replacement, that would be the first time we discussed ourselves from dependency across trials. And then in this in this context, we saw the hyper geometric distribution. That was a let us assume that we had m less than r less than N. Its simple.

So, this will be hypergeometric distribution of r, I mean I do (())(1:20) reverse we did last time. N, N people are there, r of them are characteristic, let us say and you are choosing little n people and you get this hypergeometric distribution, which is supposedly in the following form that is S is from 0, 1 upto m. You could choose upto m people of characteristic r and chance of choosing the k people of characteristic r is going to be equal to r choose k, N minus r, choose m minus k and the whole thing divided by the number of ways you could choose m people from capital N , m choose N. That is where we were at last time.

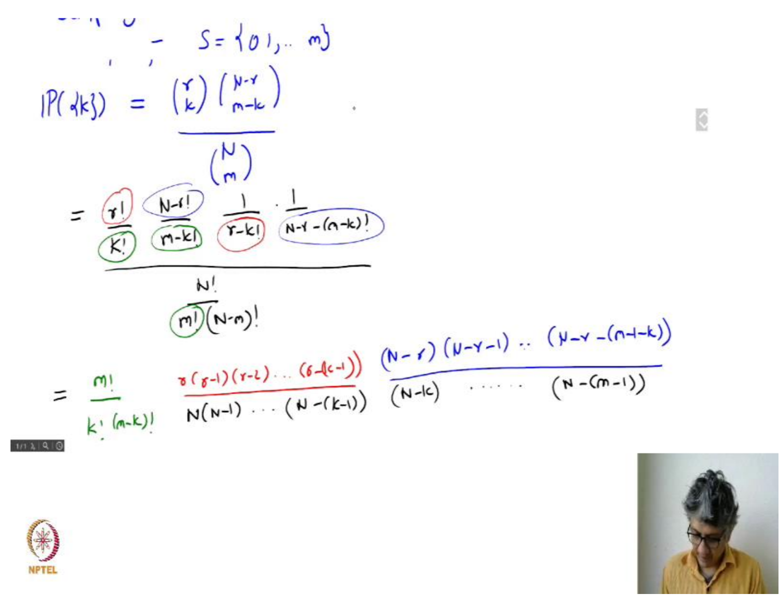
And… so is called (())(2:13) is working at contrast replacement sampling without replacement. So, sampling with replacement I started before itself there was that we need to put the person back in the population and this what we have done before and this corresponded to Bernoulli trials, independent. It had a crucial property of independence. And that is where the two things differed.

So, now let me just play with this formula a little bit. Let us start for today. So, this formula what happens is that you can rewrite this formula, in the following form, this r factorial by k factorial. Let me write it. This formula is equal to r factorial by k factorial. Then you have N minus r factorial by m minus k factorial and then you have two formulas.

You have an r minus k factorial for the first one and then you have a 1 over N minus r minus of m minus k factorial. And the whole thing divided by N factorial and m factorial and N minus m factorial. So, at the bottom, so if you rewrite this a little bit, so what will you get. You stare a little bit, so you can look at, let us look at, so there is an m factor at the bottom. So I will write that in blue, green, lets say.

So, I will just take this m factorial divided by, I will leave the k factorial as it is and I will leave the m minus k factorial like this. So, these 3 terms I will take as it is and the rest of the terms what I do is I will look at it. So I have r minus k factorial and r factorial. So, from these two people let us, let me take a, let me call it as a... Let us take this and this, I take these two guys. What do I get. I get r into r minus 1, all the way up to r minus k minus 1.

(Refer Slide Time: 04:50)



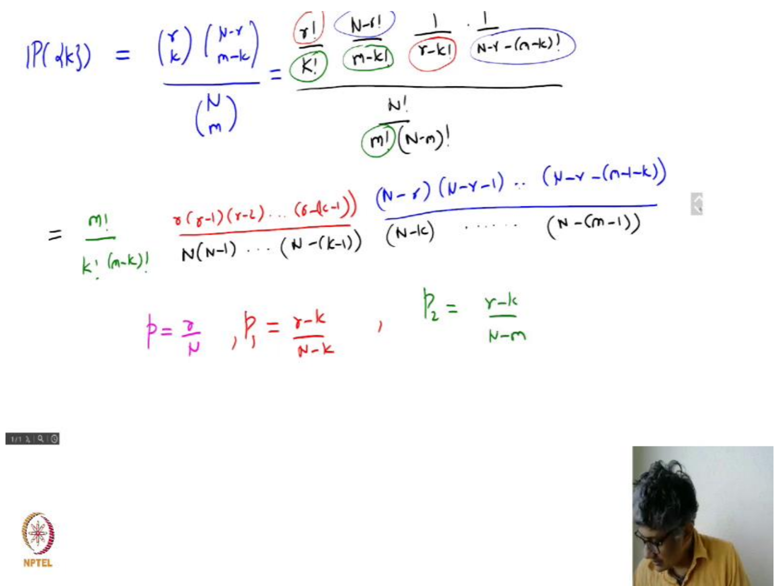
So, maybe I will go to the next slides, so its easier. I will select that. I will put this here. I will take this here. (())(05:06) So, if I do that, let me write this in red, so its easier. I will erase this. So, that is the red part, so that is equal to. Let me write it in red. Red part is r by r into r minus 1 into r minus 2, all the way up to r minus k minus 1. And the whole thing divided by you will have and I will take, I will only take this N minus m. I will remove this n minus k. So, I will get that, let me write that in green may be.

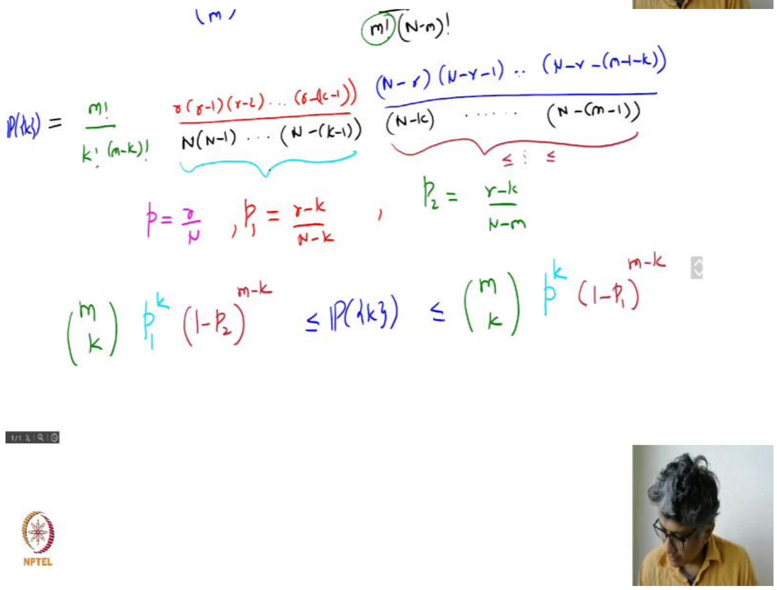
So, if I remove these two guys, (())(05:46) I will use blue may be. If I take these two people, this person and this person. What I get is, I get is N minus r into n minus r minus 1 all the way to r, N minus r plus 1, all the way to N minus r. N minus r minus, where is the (())(06:13) Oh sorry. Its n minus r plus 1, minus 1. Let us say and all the way to the N minus r minus of m minus 1 minus k.

And the whole thing I will divide by whatever is left. So, left or right in black. So, what is left is the rest the terms, so I have taken care of k factorial, m minus k factorial, n factorial. All I left is N by m factorial that is just the important part. So, that this gives you so that is split up into two parts. So I start with N into N minus 1, I will go all the way till N minus k minus 1. I will leave this here and the rest I will write here.

So N minus, sorry N minus k and I will go all the way till N minus of m minus 1. That is just this bottom part. So, this is the green part. The green part I took here this one, this one and this one, I wrote it as a green part. The red part I wrote like this and the blue part over this and the black part I just wrote it like this. So, now you are, in business so that means you re-identified p choose k in this format.

(Refer Slide Time: 07:56)





So, let me readjust this for you, so you can see the screen properly. So, P choose k is equal to r choose k by N minus m. And this gave you this formula. I will adjust this a little bit , so that you can the whole thing properly. Equal to this and then I, identify the red and the pink roperly to get the answer. Green is also here. (())(08:18)

So, now you are in business now. See now you know P choose k is this whole formula. So now once you know this whole formula, like this, like I discussed last time, so you take these two things. Let us say you take three numbers.

Let us use a different color. Let me take this colour this with purple. So, two different colours. So you take p to be equal to let us say r by N p equal to r by n. Very good. So, now you look at p1. Let me look at p1, p1. Let me look at p1, two different colours. Let me go with, let me use same colour (())(09:17). p1 sorry p2, p2 is equal to, sorry (())(09:24). So, lets say p2 is equal to the other one, r minus k by N minus k.

And then say so this is p1. First P and p1 (())(9:43) and then let us say I take p2 in green, p2 to be equal to the upper bound so r minus k by N minus m. So, now you are in this state. So this whole thing is P of k the expression, that I have at the top. So, now you can easily see it is quite easy. You take the you take the P of k. See this expression here this is P of k.

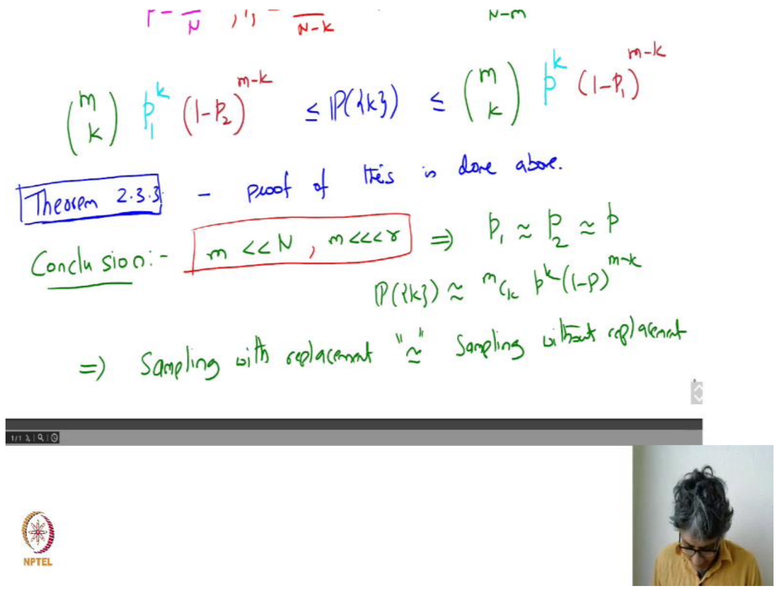
Now, you can put two bounds on either side. So, you leave the m choose k as it is. So, leave the m choose k in this, that is on both sides. So, m choose k is on both sides, m choose k. That is what this term is m choose k is on both sides. There is no problem. And then you write the two formulas down. So, you look at each of these red, red over black and blue over black. So, red over black, the red over black is going to be, has k terms in it. The red over black is going to be less than at least p1 to the power k.

Let me write that in black now (())(10:51) Let me add the colour. May be light blue let say. So red over black. This particular term right here right, this is always going to be bigger than it should share a little bit. So, it is always going to be bigger than each of the terms is bigger than p1. So, this p1 is r minus k at the top and N minus k at the bottom.

So, N minus k is much larger and r minus k is much smaller than (())(11:19) terms. So, each of the blue terms is going to be bigger than p1 to the power k. And it is going to be smaller than on the other side, it is always going smaller than P, because everybody is like r by n. So, everybody is smaller than P. So, it is smaller than p1 to the power k (())(11:33). So, it is the blue term is like this. You can add another colour. Let me see if I can add another colour. (())(11:44).

So, if you take this one. Let me take this one. If I go take this person right here. This person right here is always going bigger than if you stare a little bit, it is 1 minus p2, the whole to the power m minus k. The m minus k terms and each of them is bigger than 1 minus... And each of them is less than on this side look stare a little bit, is this 1 minus p1 to the power m minus k. So, that is how the, you just bound the upper bound and lower bound these term and here also the same thing. So the bounds you get are very interesting.

(Refer Slide Time: 12:20)



So, now that is what I stated that theorem last time. This is the theorem I stated last time, so let me write the theorem down. So, it is clear, so the theorem I wrote down last time was the following I said that. So here what I do, I state theorem I called it theorem 2.3.3. And I said that, this exact description I wrote down that it is a proof of this theorem. So, it provides the Theorem 2.3.3 from last time.

So, what does it give you. So, it tells you that p of k the hyper geometric is kind of close to binomial, but of course with different different parameters on both sides. So, binomial is not quite close by. So the idea is that, so the key conclusion is the following. So we made a conclusion. Conclusion is the following. So, if m is is much much smaller than capital N relatively small.

And also m is relatively small to r, so the number of people with characteristic r is also much much larger than m. This would imply that both p1 and p are kind of close by, so there is no problem; because m really does not matter. So, p1 and p2 are kind of close by. So m really does not matter. Because k is also, k is also something between 0 and m and it does not really matter. So, once this is, this and let us say they are both approximately equal to p.

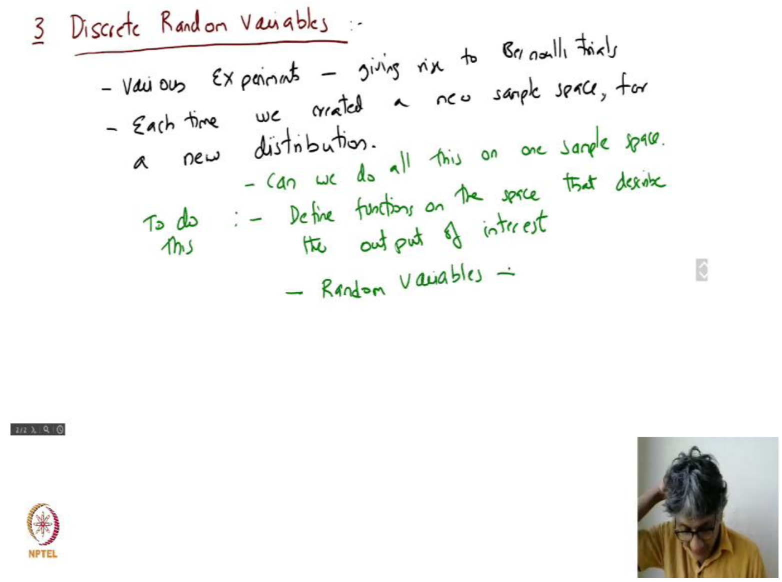
So, that means what you would get is p of k, you would get is p of k, is like m choose k, p power k, 1 minus p power m minus k. That is what you would get. And this is implies that in this situation when you have this idea that this sample is much smaller than m and the sample is much smaller than the amount of people with this characteristic of kind of sample from, then what you would end up getting is that, is the key observation is that; sampling with replacement is kind of the same as, it is kind of same as sampling, sampling without replacement.

So, that is why in opinion polls and other things you do not really bother. You just go and pick, pick people up with replace, without replacement, but you assume it is without it is, with replacement. So, you do without replacement sampling, but if the characteristic holds that m is much smaller than N and little m is much much smaller than r as well you just go and do; you assume that the whole thing is independent. (())(15:23) But this is the clear correct analysis for (())(15:27).

And I will make it let us go back again, quickly recall so without, with replacement means your independent trials without means of dependence. If you do that without formula, you get probability of p of k is equal to this. And that would imply p of k. You will do the algebra properly, it will all come out to be this. And then once you do that, you are on your way. So you just identify the bounds properly and then you get this answer but then you look at it does not look quite binomial. But if you look binomial in the sense if p1 is same as p2 same as P. Very nice.

So, that sort of concludes that section of independence and dependent trials. So, now I will quickly move on to the next topic where I will sort of formalize notation. So, here I want to use a different color, let us see if I can get the color.

(Refer Slide Time: 16:42)



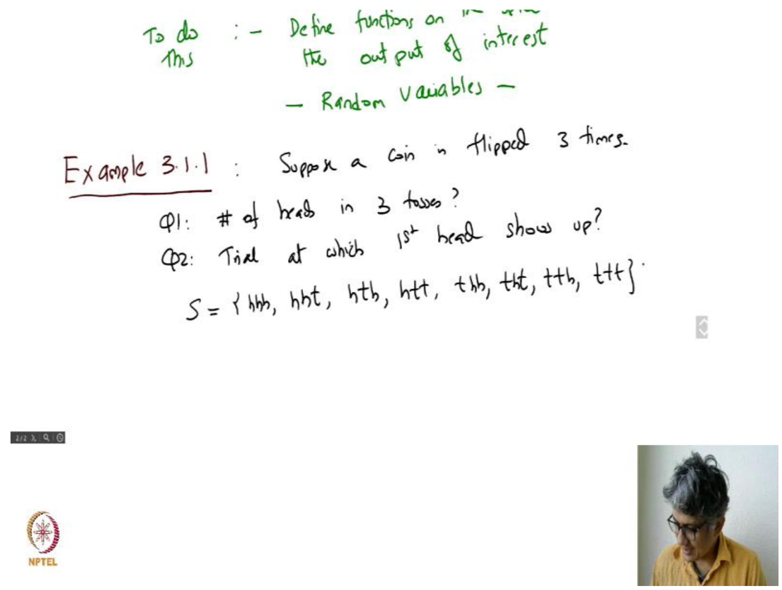
I want maroon (())(16:32). So, in this section what I will do is, I will now start on chapter 3, which is a discrete random variables. So, here I would like to do the following. So, here the whole idea was before was that, we had these Bernoulli trials and each time we created new sample spaces. So, if you noticed we had various experiments giving rise to Bernoulli trials and also dependent ones that we check.

But each time, each time we created a new sample space. So, this is kind of a sample space for a new distribution. So, which means that every time I gave you a, a I have a geometric abundance. I have sample is this poisson and I give another one and so on. So now the, this is something that little and convenient. Can we can we do all this in one sample space, because you do not want to change this.The question I want to ask is that can we do all this on one sample space.

Now, rather then keep on going and changing different different things and sort of. So, the jump is from this is to do, this to do this, so one needs to do one needs to construct a certain concept called random variables. So, what you do is, you define functions on the space. So idea, idea how to do this is you define functions on the space. So functions on the space, that describe the output; whose outputs relate to the questions under discussion (())(19.29) That describes the output. Describe the output of interest and once you do this you are on your way. So, these are called random variables, random variables.

That is the formal idea that is your random variables are brought into the picture because you want to work with only one sample space. You do not want to work with many many samples spaces. That is how you start. Let me just do a quick example and then I will try and I will try and illustrate the idea of random variables (())(20:20)

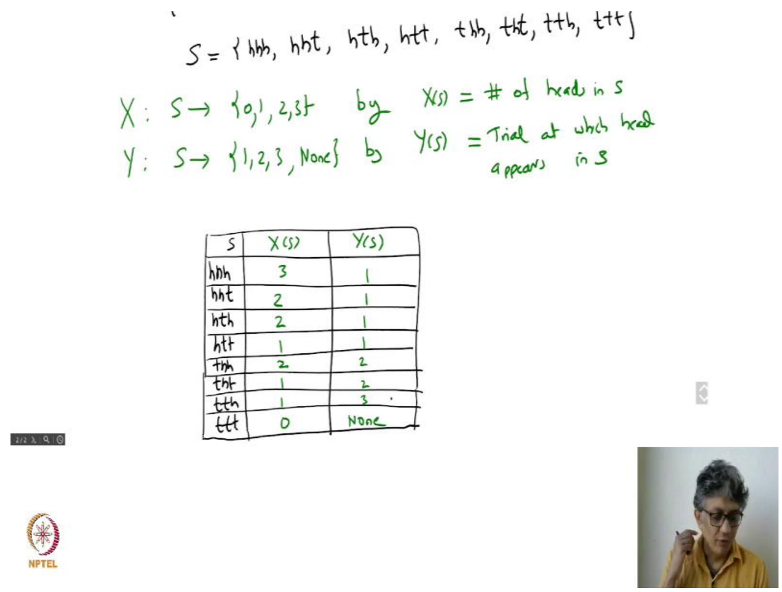
(Refer Slide Time: 20:24)



So, let me do example 3.1.1. So, here what I want to do is I want to I want to just set up this framework properly (())(20:41). So, how do I do this let me go back to black. And then here what I say is that suppose a coin is flipped. Let us say maybe 3 times. I flip it 3 times and I want to know the number of heads in 3 tosses or I want to know the first tosses are. So see there are 2 questions, 2 questions. Question one is the number of heads in three tosses, that is one question I have in mind. Question 2 is I want to know is that, the the trial at which the first head shows up. So the trial if it exists, at which, at which first head shows up.

So, now this is something we have to think about. So, you know how to do this. So, you just what what do you do before. You would write the sample space S as hhh, hht, hth and then htt and then write the whole thing down. I will write the other thing down, thh, tht, tth and then ttt. And then your these three samples. Take sample points and then from this you try to understand the question. So, then what you do is you try and you try and pick the number of outcomes that have 3 heads or number of heads you just write them. If you want to say 2 heads means you go pick the outcome that satisfies two heads and compute the whole thing.

(Refer Slide Time: 22:47)



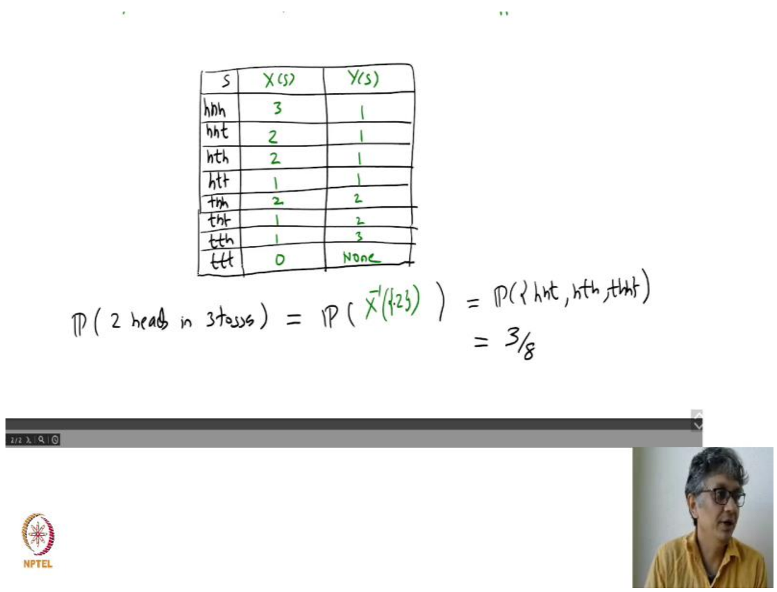
But let us just do a simple calculation. So now suppose I say let us define the function X. If X is a function from sample space f 2 0 1 let us say 0, 1, 2, 3. I have defined X in the following form. I define X to be, what I do to is take up of. I take a point X and I just count the number of X. So what I do is I re-add this a little bit. So X of S is this (())(23:16).

So, select if I define it like this. So I will define let us say Y as the S from 0 to let us say 0, 1 can occur the first trial so Y, I want for second question. So, if I say Y is let us say 1, 2, 3 or I say I use the the null notation called none. The alphabet notation none by Y of S is the trial at which head appears in S. There are two functions that are defined. So, once I do this let us do let us see if the table comes out nicely. So, let me do a table. So, the table is quite clear. So, I do this and then I do this and I do this.

So, I first write down the outcomes of S in this column. So, let us call that a little s (())(24.36). I have hhh, hht, hth, htt, thh, tht. We will erase the bottom line. I think I have to put thh, tth and ttt there are 3 columns I have. I go down all the way, fix this,very good. Then on this one I plot the values of X, X of S, So I go here, X over S (())(25.10) then what I do is I go here and I plot the value of X. This is in green like I said before and maybe X of s and this is Y of S. So, in this case X of S is 3 right. X of s is 3 right.

In this case it is 2 . In this case it is it is again 2, 1, 2, 1, 1 and 0. And here the number of trials at which the first time it appears is 1, 1, 1, 1 and then here it is 2, there is 2, here it is 3 and here it is none. That is how I would fill the table.

(Refer Slide Time: 26.09)

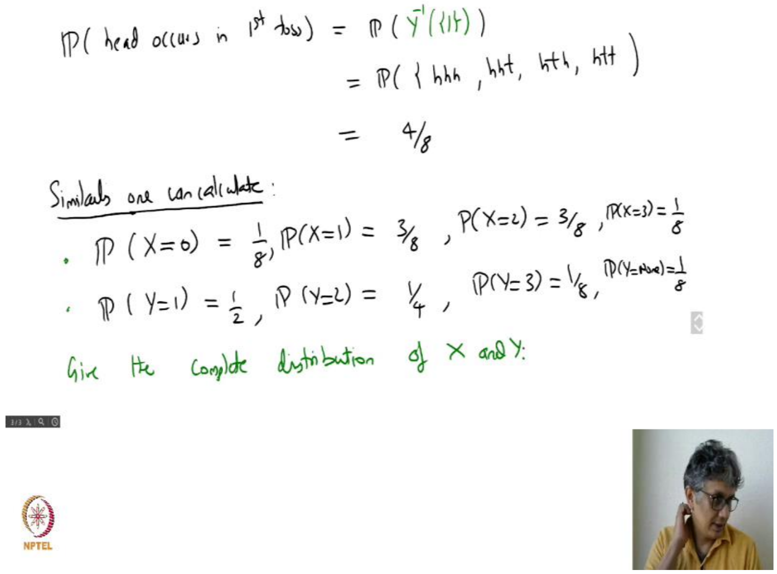


So, now if you want to know the probability of let us say from this question if you want to know the let us keep this page alive. From this if you want to know the probability of let us say 2 heads in 3 tosses. So, what you could do is you could just go and see the elements of S for which X of S is 2. So, that is the same as saying you look at the inverse image of X when X is 2.

And that if you look at this equation right if you look at the table, it is just the probability of the event let us say hht, hth and that is its, that's the two of them and then the one more thh. And if you do this properly, this calculation properly; it is just going to be again 3 by 8 and that is the answer.

So, I did not have to go and okay this is how I do the calculation, but I could use my X to get the answer, in one way in which I could try to erase this, let me write in this green. So, this is the same as X inverse of the set 2 and 2 and curly brackets and 3 are all the same things. Very good, so that is what it is. So, now you are on your way now.

(Refer Slide Time: 27:37)

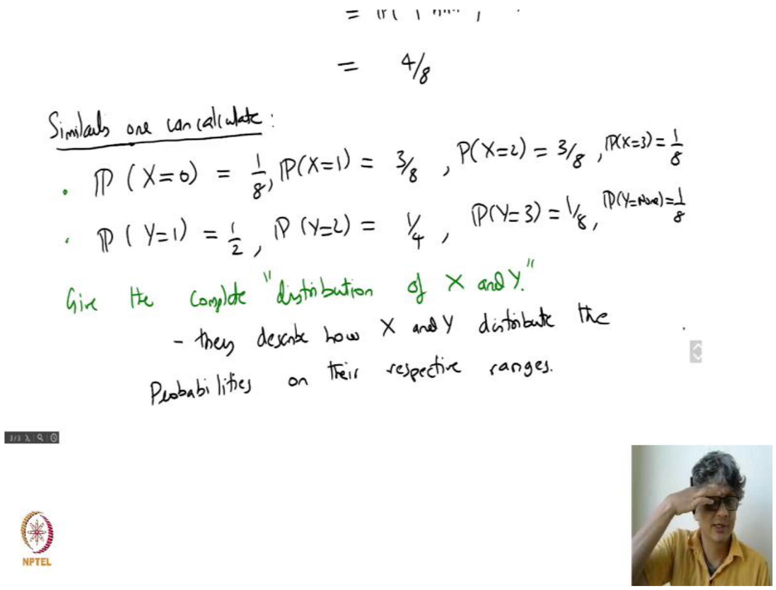


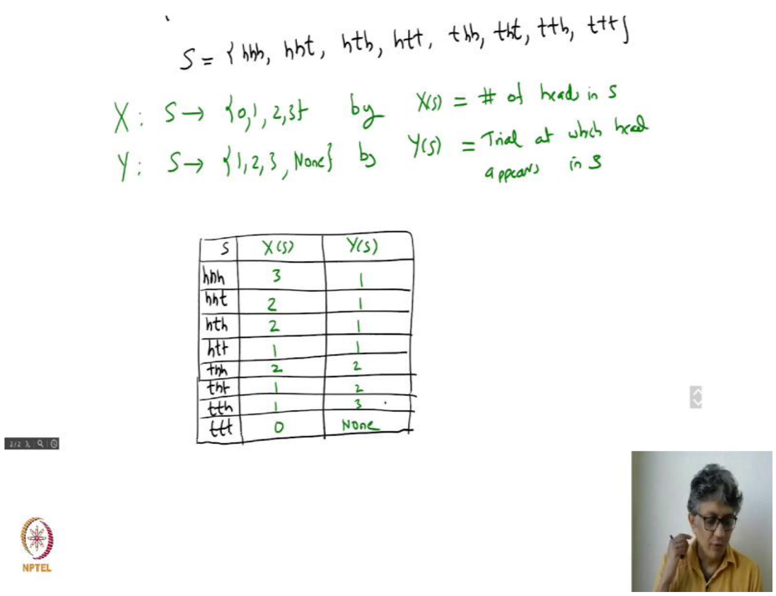
So, now you understand exactly how to do this similarly you can do that the thing of probability let us go next page. Let us do the next one, so let us say you want to know the probability that our head occurs in the first toss.That means that is the that is the same as the probability that Y inverse is, is, is one right you look at the first one the first one so that that you know again you go back and look at the, look at the sample space and that is the same as the probability that Y of your first asking your hhh, hht, hth and htt. These are the four points and the answer is just 4 by 8 and that is just half.

So, once you do this whole calculation this calculation. We have to do by hand. But let us say let us summarize, let us summarize. So, what you would get is the chance that,you the chance that, let us say I will write this as X equal to 0 that means there are no heads, is going to be equal to 1 by 8. You can do it chance that X equal to 1 we just calculated (())(28:46) is 3 by 8. X equal to 2 is 3 by 8 like we did before.

Sorry will rease this, equals 3 by 8. So, I will remove this, (())(29:03). And chance that X equal to 3 is just again 1 by 8. So, I would say similarly, one can proceed, one can try to summarize. I will say similarly one can calculate the other problem and similarly you can do the for the Y also you can do the same thing. You can say Y equal to 1 is which already done is for half, probability Y equal to 2 is the same as one quarter. Our probability Y equal to none Y equal to 3, is like one eighth and probability Y equal to none is again one eighth. That is how you finish off the problem. So, what you notice is that these two equations give the complete distribution of X and Y.

(Refer Slide Time: 30:34)





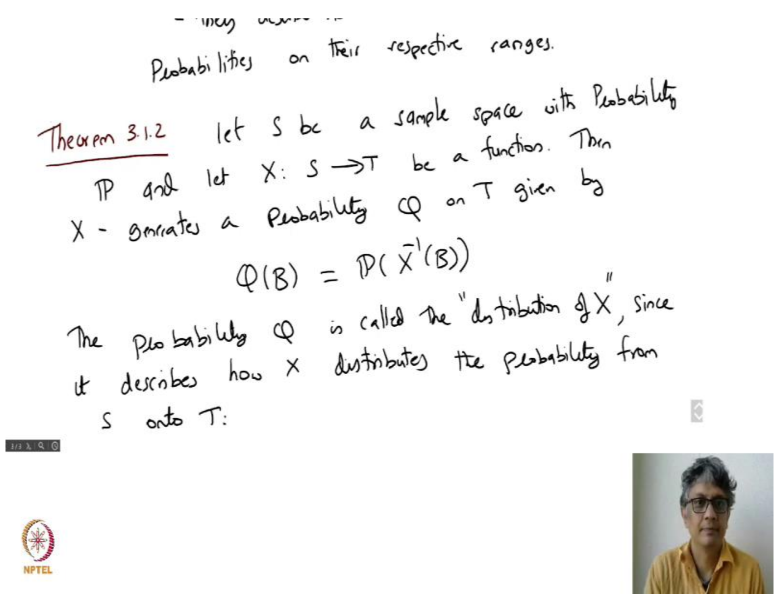
So, in some sense what is this, what is this, what does this line mean. What is line what I mean by this. So, it describes how X and Y distribute the probabilities on their respective ranges. So, in a sense, what I mean by this line S that, they describe how X and Y distribute their probabilities on their distributes is not there probably (())(31.20) the probabilities on their respective rates.

And that is what, that is what, this is (())(31.29). So, this is one, set up where you can use functions and we sort of forgot about the original samples. We looked at what the functions do and we looked at. You can just look at the range of the functions and understand their probabilities and once you know that you know exactly what X is doing what Y is doing. What the range is, what the probabilities are.

You can answer your questions. And essentially you have forgotten about the sample space. Sort of in the background, once you know these 2 equations. Of course computing these 2 is little hard but as a math. But you have to do it only once. So that is what I did, so I I did I had these two questions in mind. I constructed two functions. I did a table.Then I want to construct the question I have an interest is two heads and three tosses. That is the same as X inverse of 2. Then I go down I want another question let us say head occurs in the first toss. That is same as y inverse of one and so on.

So, maybe I will just say, head occur, I will say first head occur. So, now this is the structure we can formalize. So, how do you formalize the structure (())(32.45). So, here you can formalize. The formula is the following. So, let us do a theorem.

(Refer Slide Time: 32:56)



Theorem 3.1.1. Go in your definition before that. Maybe not its not there. Let us just do a simple 3.1.2. So, what is the the mod. So, let S be a sample space with probability P. And let X be a function from S to T, here. Then X generates a probability Q on T given by the problem the Q of any set B is given by the probability of X inverse. That is the set of all sample points for which X is lied inside.

This is a nice term so the probability Q, Q is called the distribution of X. So that is how Q how X distributes its value. Distribute probabilities onto a string, since it describes, describes the probability, probability of S, probability of (())(35.24) still describes how X that distributes the probability from S onto T. So, maybe I will take a quick 5 minute break and I will come back and try and show in the next class I will try and show this theorem 3.1.1.