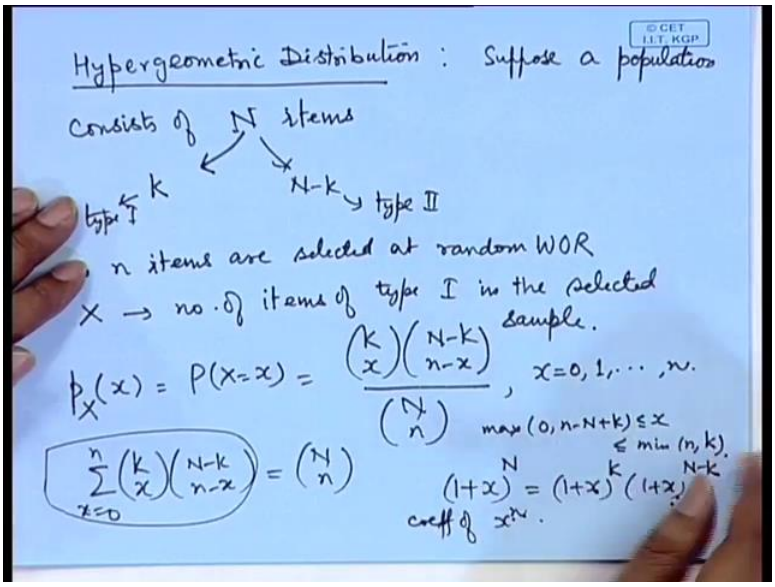
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 21
Special Discrete Distributions – III

In the Bernoullian trials that we considered, we are assuming that each trial has been conducted in a identical conditions. That means the probability of success or failure remains unchanged in each trial. However, in the finite sampling situations this type of probability being constant is not applicable. Suppose there are 5 people out of which 3 are say males and 2 are females, the probability that in the first trials we select a person then the probability that it is a male is 3 by 5.

Now, suppose in the first trial a male has been selected then now we are left with 2 males and 2 females, and therefore the probability that a male is selected in the second trial, it will be 2 by 4; that is half, so it has changed, it is not the same. Now to describe the distribution of number of success trails are probability of selecting of one type of persons or one type of items in finite population models is described by Hypergeometric Distribution.

(Refer Slide Time: 01:34)



So, let us consider this hypergeometric distribution, suppose a population consists of N items, N items, N persons or N characteristics. Now out of this k things are of say type

one and remaining N minus k things are of type two. So, you can think of like in a population, how many males and how many females are there, how many are smokers, how many are non smokers, how many are infected with the certain disease, how many are not infected with the certain disease, how many are of a particular educational level, how many are below that educational level etcetera.

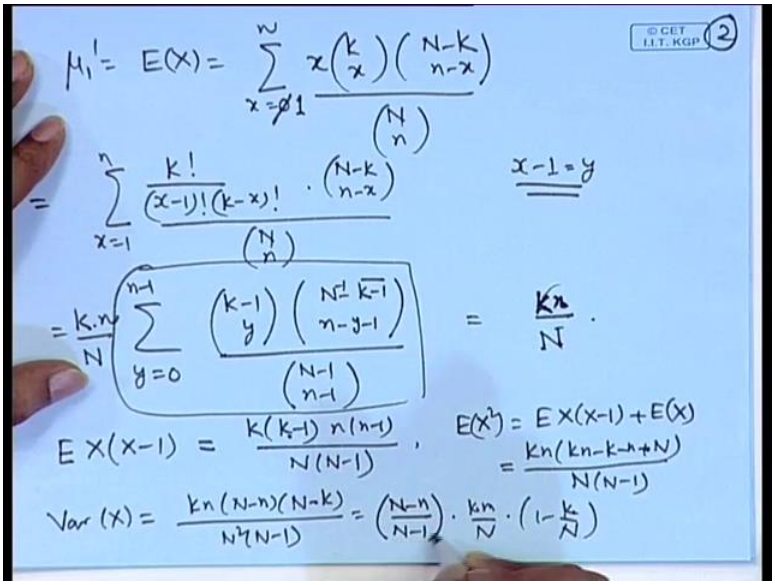
So, like the Bernoullian trials where we are looking at everything with only a two sided vision that success or failure. Here also the same thing is there, but the model is that we are considering finite population model. Now out of this n items are selected at random. So here Bernoullian model will be applicable, if we select one item, keep it back that is with replacement then select another one then the probability of success or failure in each trial will be same.

However here we do without replacement; that means, all of them are taken and that means, one is taken and then it is not kept back. So the probability of selecting an item of type one or type two changes at the next stage. So, X is the number of items of type one in the selected sample, then what is probability of X is equal to say a small x that is; so it is equal to k c x N minus k c n minus x divided by n c n; obviously, x will take values from 0, 1 to n. However, in order to write down this expression, these binomial coefficients must be valid; that means, x must lie between 0 and k, n minus x must lie between 0 to N minus k etcetera. Therefore, the actual restriction for x will be from maximum of 0 to n minus n plus k to minimum of n and k.

So, x takes integral values 0, 1 to n where x lies between these two, so this is called hypergeometric distribution. See the fact that sigma k c x, N minus k c n minus x x equal to 0 to n under this restrictions it is equal to n c n, this can be proved by considering the expansion of 1 plus x to the power n into this, we can express it as 1 plus x to the power k and 1 plus x to the power n minus k. So, if we look at the coefficient of x to the power a small n in this expansion then in this one, you will get n c n and on this side since it is product of two terms x to the power N can be obtained as x to the power 0 into x to the power n x to the power 1 into x to the power n minus 1 and so on. So, if we collect all the coefficients, we will get these terms, therefore the sum of this is equal to 1 and it is a valid probability distribution.

Let us look at the moments structure of this distribution, since the factorials are involved like in the binomial distribution; if we want to calculate say expectation x, now here the x in two something we will come therefore, this factorial as to be adjusted. Therefore, it suggests that like in the binomial distribution; the moments of the hypergeometric distribution can also be calculated easily using the factorial moment structure.

(Refer Slide Time: 06:45)



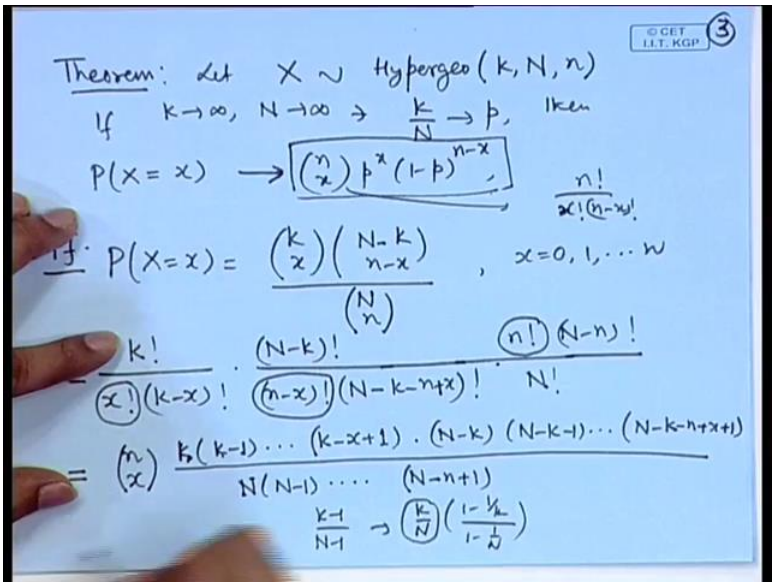
So, we may consider mu 1 prime that is expectation X that is sigma x; k c x N minus k c, n minus x divided by n c n x from 0 to n and as you can see the corresponding to x equal to 0 this term vanishes. So, this will be from 1 to n and then if we expand in the factorials; this can be written as k factorial divided by x minus 1 factorial, k minus x factorial; N minus k c n minus x divided by n c n, now in order to adjust the terms we may put x minus 1 is equal to say y. So, this becomes sigma y is equal to 0 to n minus 1, now accordingly all other terms have to be adjusted, therefore we can write it as k minus 1 c y N minus k c n minus.

So, we can write it as n minus 1 minus k minus 1 divided by n minus y minus 1, divided by n minus 1 c n minus 1. So, the terms that have been adjusted will give us k n by N. So, this sum is 1 therefore, you get k n by N as the expected value of x. Now if we explain this there are N objects, out of which k objects are of type one, so the proportion of the type one objects is k by N. So, in n draws what is the expected number of type one that will be n into the proportion.

So, in a similar way if we want to calculate say expectation x square; we will need to calculate the second factorial moment, doing the calculations in a similar way this 1 will become x equal to 2 to n. So, we adjust the terms in the form k minus 2 c y, where y will be from 0 to n minus 2; the final expansion will give us k in to k minus 1, n into n minus 1 divided by n into n minus 1. This gives us expectation x square which is equal to expectation of x into x minus 1 plus expectation x as k n into k N minus k minus n plus n divided by N into N minus 1.

Therefore, variance of x will become equal to k n; N minus n; N minus k; N square into N minus 1 is equal to N minus n by N minus 1; k n by N into 1 minus k by N. Notice here that if we consider large population where n becomes large, k becomes large. So, such that k by N goes to a fix proportion say p, then this converges to n p. Similarly here this will converge to n p 1 minus p; this will go to 1 that is going to n p q. So, the mean and variance of the hypergeometric distribution will converge to the mean and variance of a binomial distribution. This suggests wider phenomena that are in a hypergeometric distribution if the population size is considered to be infinitely large, such that the proportion of type 1 items is p then if we are considering a sample of size n, then the number of type 1 items must follow a binomial distribution.

(Refer Slide Time: 11:10)



So, this fact can be theoretically proved; let x follow a hypergeometric distribution with parameters k, N and n. If k tends to infinity, N tends to infinity such that k by N tends to

p then probability of x is equal to say x converges to n c x p to the power x 1 minus p to the power n minus x which is the binomial probability distribution, so let us look at the proof. So, in hypergeometric distribution probability that x equal to x is written as k c x; N minus k c n minus x divided by n c n.

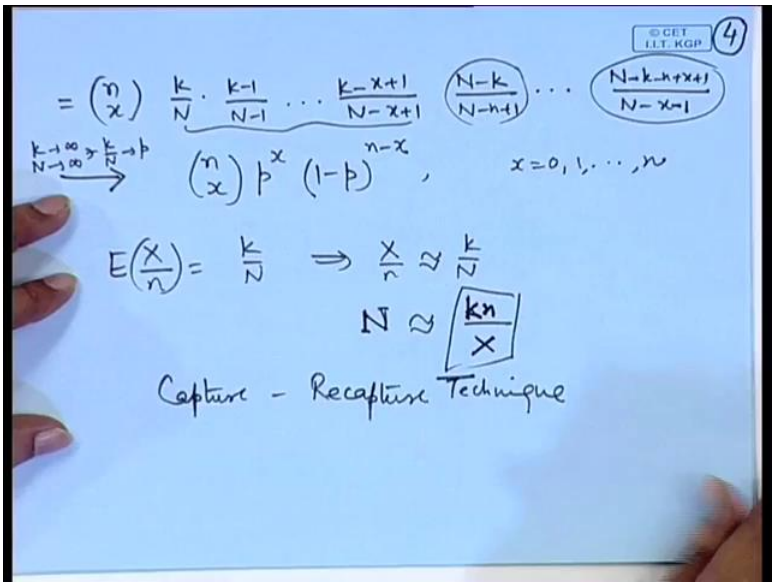
Here the values of x are from 0, 1 to n subject to a restriction that x is between maximum of 0 n minus n plus k to minimum of n k. Now here if you are assuming that n and k are infinitely large where k by n is a fix proportion p, then naturally the term k minus n is going to be infinitely large and negative therefore, n minus that will be negative; therefore, here maximum will be 0. Similarly, as k tends to infinity n will become the minimum value; therefore effectively speaking the range of x is from 0 to n.

Now, if we want to take the limit of this as k tends to infinity and N tends to infinity such that k by N tends to a fix number, you have to look at the expansion of these binomial coefficients because here we cannot take the limit. So, let us look at this it is equal to k factorials divided by x factorial; k minus x factorial N minus k factorial divided by n minus x factorial, N minus k minus n minus, so this becomes plus factorial divided by n c n. So, this is n factorial, n factorial; n minus n factorials. If we want to take the limit as capital N tends to infinity, k tends to infinity etcetera; we have to further simplify these coefficients.

Now, here if we look at the term in the binomial probability mass function n c x is coming which is having n factorial x factorial and n minus x factorial. So, naturally these terms can be seen here n factorial, x factorial and n minus x factorial. So, this term is readily available here, now this k c x, so this term we can write as k into k minus 1 up to k minus x plus 1 into k minus x factorial. So, this becomes k into k minus 1 up to k minus x plus 1; the second term is N minus k; N minus k minus 1 and so on up to N minus k minus n plus x plus 1 and in the denominator; N factorial can be written as n into n minus 1 and so on up to n minus n plus 1.

Notice here k k minus 1 up to k minus x plus 1 these are x terms and here if we look at the first term here k in the numerator and the denominator it is k by N; which converges to p. So, likewise if you look at k minus 1 by N minus 1; this can be written as k by N; 1 minus 1 by k divided by 1 minus 1 by n as k tends to infinity n tends to infinity this goes to 1 and k by n tends to p.

(Refer Slide Time: 16:11)



So, in a similar way we can consider up to k minus x plus 1, so after adjustment of the terms we can see that this is equal to n c x k by n k minus 1 by n minus 1 and so on up to k minus x plus 1 divided by n minus x plus 1. Now, x is up to n therefore, this term is coming N minus x plus 1 is coming somewhere here, N minus n plus 1 will come afterwards. So, there are n minus x terms is still left over, which we need to adjust with other terms. So, for example, here next term will become k minus; N minus x, N minus x minus 1 and so on up to n minus n plus 1. Now once again we need to adjust these terms. So, how will we adjust; if we look at N minus k divided by N minus n plus 1, then this is equal to if we take common n here this is 1 minus k by n divided by 1 minus N minus 1 by n.

So, as n tends to infinity this goes to 1 minus p, n minus 1 by capital n goes to 1 this goes to 0, so 1 minus this goes to 1. Therefore, we can adjust the terms like N minus k divided by N minus n plus 1 and so on up to N minus k minus n plus x plus 1 divided by N minus x minus 1. So, now if I take limit as k tends to infinity N tends to infinity such that k by N tends to p then this is converging to n c x these term as we explain each of them converge to p these are x terms. So, this goes to p to the power x and these are n minus x terms each of them are going to 1 minus p. So, this goes to 1 minus p to the power n minus x which is nothing, but the probability mass function of a binomial distribution.

So, this finite population sampling in case the population size is large is same as a binomial sampling. Some peculiar application of the hypergeometric distribution are as follows, suppose you want to estimate the number of tigers in a reserved forest, suppose you want to estimate the number of fish in a fresh water lake for example, a fishing company wants to estimate the number of fish which are available in a particular area of the lake. Now it is impossible to take out the water or go inside the lake and check how many fish are there; one can calculate this by certain sampling procedure.

So, we may take a sample of k fish from the total number of fish are capital N. So, we take a sample of small k fish and tag them and flow them back in the lake. Now you take sample of size a small n and see how many of them are tag, which is we call X. Then expectation of X by n is equal to k by n as we proved just now. So, this means that X by n is approximated by k by N, so N can be approximated by k n by X. Since we have consider sample of capital X by initial k type of fish were tagged and a small n is also known to us this number is known to us.

So, we can estimate the number of fish in the pond or in the lake of a particular type; this is known as Capture-Recapture Technique.

Thank you.