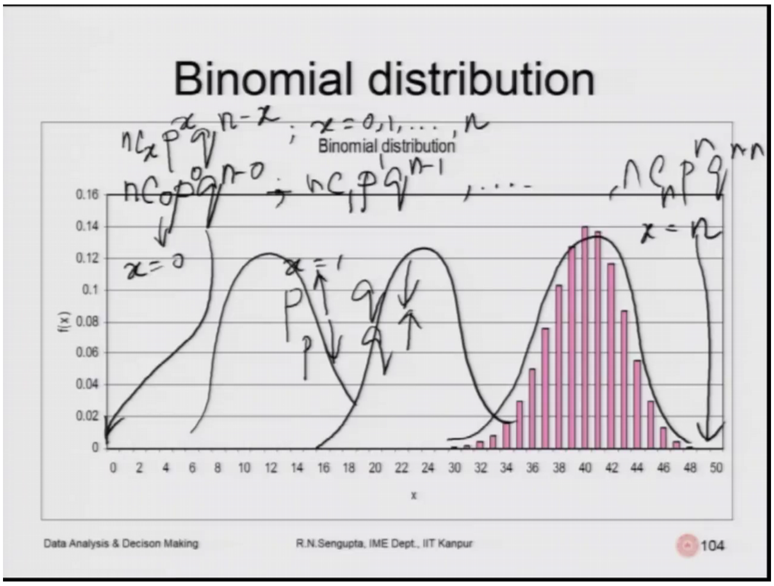
Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 08
Binomial Distribution

Welcome back my dear friends a very good morning, good afternoon, good evening to all of you; I am Raghu Nandan Sengupta from IME Department IIT Kanpur. And this is the DA DM lecture 1 let; that means, the DADM 1; lecture number 8.

And as you know this is a course for 12 weeks which is 30 hours and we have just started discussing about the different type of distributions discrete case distributions. And we will give simple examples as well as solve them and give small examples for an exercise which is as I told are not part and parcel of the assignments, but they would be more to inculcate the interest in all the people who are taking this course.

(Refer Slide Time: 01:02)



So, if you remember we considered the binomial distribution just the last slide for the seventh lecture. So, binomial distribution is basically as I told you is an expansion coming from the concept of binomial expansion. Now, binomial distribution is in some way related; obviously, for all the distributions when we are saying it is the Bernoulli trial.

So, in the binomial distribution there are 2 outcomes head, tail, good, bad, red, yellow, yes, no whatever it is. And the distribution is given by the binomial distribution which is n cx p to the power x q to the power n minus x. Now remember that what I told repeatedly in the last class is that p is the probability of a success or a failure whichever you look at, q is the probability of this of the failure of the success because if you p is success; then q would be the failure and if p is the failure q would be the success or p can be the number of yes, q can be number of no's, q can be p can be number of no's q can be number of yes.

Whatever is there are complimentary to each other; x would be the total number which is coming out as true or false or red or white or black or white, whatever it is depending on what the value of p is. So, if p is yes then x would be the number of yes's.

If p is basically white then x would be for the number of whites and n is the total number of throws which you are going to do. And you remember that I did mention that I will only give the formulas for the expected value, but technically the derivation on these things are very easy, they can be derived accordingly.

So, in the case when you have the distribution as say for example, binomial the expected value is n into p and the variance is n p q. And the corresponding the way how you denote is that x is binomial which is b with parameters n and p. So, if you draw it I am using very simple excel sheet the binomial distribution would look like this.

So, you have basically 50 as n and n would basically be depending on how where you are basically consider. So, if you are considering the expansion let me write it. So, it was n C x p to the power x q to the power n minus x and x can take values of 0, 1 till n.

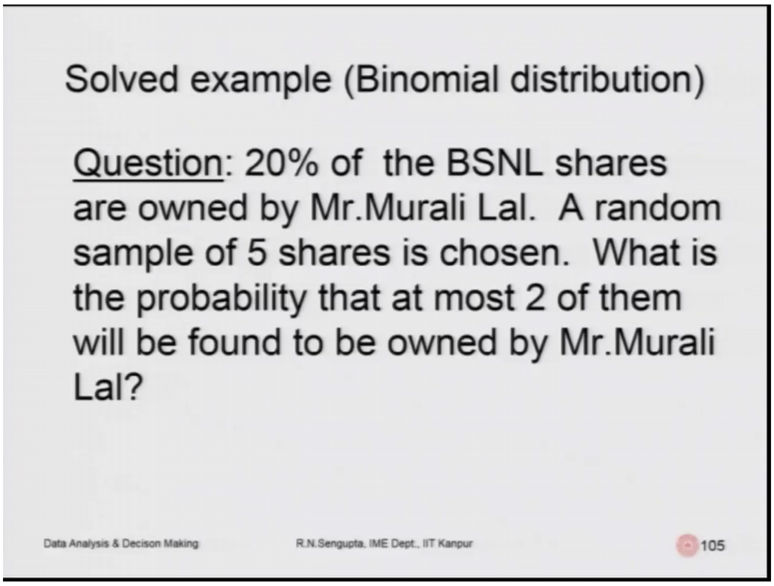
So, the first term is n C 0 p to the power 0 q to the power n minus 0, then next term would be basically when you go to the next term this is x is 0. When you have x as 1, you will have n C 1 p to the power 1 q to the power n minus 1. And it goes to the last term which is n C n p to the power n q to the power n minus n x n.

So, this value would basically come in here this value would basically here and so, this would be the corresponding radius you are going to plot. Now remember, if you increase p correspondingly q would decrease or decrease p q would increase. So, in that case the distribution which you see here would basically shift here or here depending on the value. So, it will just a shift; so how you plot it would basically depend on the values keeping n as fixed, it will depend on the value of p and q.

So, p is very small q would be very large; that means, large means almost equal to 1; p would be almost equal to 0 or p is 0.5, q is 0.5 or p is very large almost equal to 1, q would be almost equal to 0; you can basically get the values accordingly; it is very easy if we draw it and basically understand and accordingly.

Now, consider us a very simple example, 20 percent on the BSNL shares are owned by Mister Murali Lal. A random sample of 5 shares is chosen, so I want to find out or we want to find out what is the probability the that at most 2 of them would be found to be owned by Mister Murali Lal. So, if 5 are chosen; so n would be 5, so as we are discussing the manimal distribution. Let us also discuss a very simple example for this.

(Refer Slide Time: 06:03)

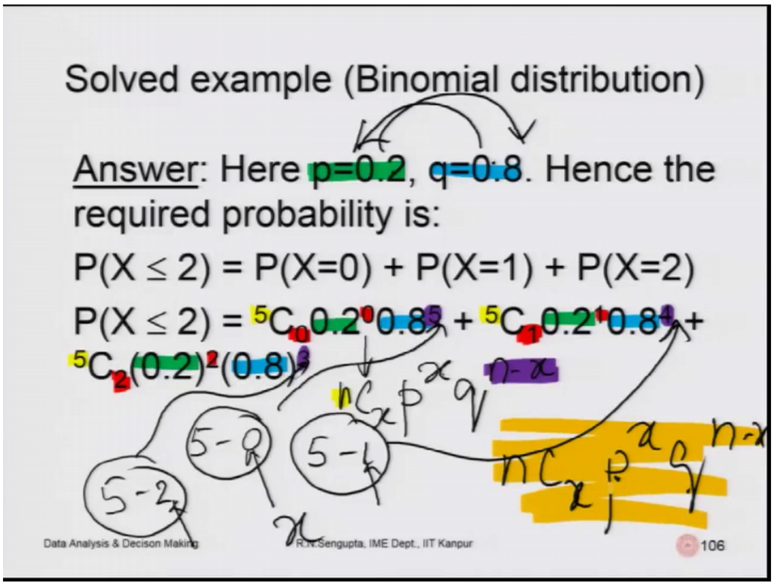


So, the problem states that 20 percent of the BSNL shares are owned by Mister Murali Lal or random sample of 5 share is chosen. So, which means n is 5; so, what is the probability? That at most 2 of them would be found to be owned by Mister Murali Lal.

So, at most 2 means the value of x; now what is x we are going to define that. The value of x can be 0, can be 1, can be 2; in other sense if you are looking from the other side at most 2 would be the compliment would be 3, 4, 5. So, how you place that problem and what to consider p and q would be important. In one case if you take p as 20, q would be

0.8, in another case can p can be 0.8 and q can be 0.2. So, we will state the problem we will understand and then I will explain.

(Refer Slide Time: 06:54)



Consider here a p as 0.2 so; that means a 20 percent of them as our owned 80 percent are not owned; so, q would be 0.8. Hence the required probability would be in the case at most 2; so, we will find out x values being 0 1 2. So, when it is X is 0 the corresponding probability would be ns I am writing on the formula and then going to say that what are the values corresponding to n p x and so on and so forth.

So, we know n is 5; so, this is 5 I will also use a different color to highlight it. So, this is 5 which is n; so, this is n ok; now we will do consider then I use a different highlighter red. So, x value would be corresponding to the case when X is 0, X is 1, X is 2; so X is 0, X is 1, X is 2.

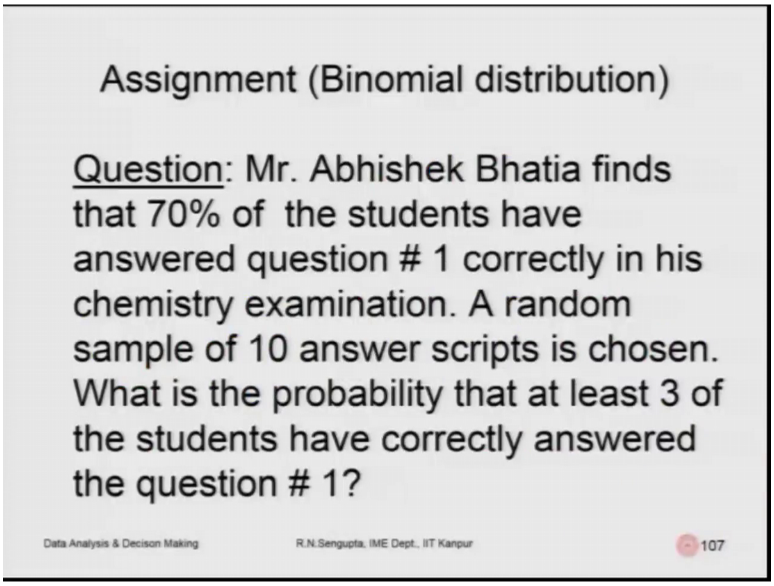
So, now we you need to and these values are also 0, 1, 2; now I want to find out and highlight not find out highlight that value of p we you have considered as 0.2. So, p would be 0.2 here, 0.2 here, 0.2 here; we have considered q; so, q is 0.8 here, 0.8 here, 0.8 here and the radius which you have I use let me use a magenta color.

So, the value which you see here technically the; this one 5 here, 4 here, 3 here that is basically n minus x. So, in the first case it will be 5 minus 0 I should be using. So, it is 5 minus 0 for the first case which is here 5 minus 1, which is 4 which is here, 5 minus 2 which is 3 is here and the corresponding values of x is are changing. So, these values which I am writing 0, 1 2 are basically the x's depending on which reading you are taking.

Now, as I mentioned there just in the last slide that if you take the complementary part; that means, you take the formula this as 0.8 as p and 0.2 as q in that case I will write the formula in the same way, there is no change. So, p becomes 0.8. q becomes 0.2, but here the x what I am denoting here let me highlight using light blue let me use orange.

The x which I am using here that would give me the probability of the complementary part corresponding the case when you are considering x as the number of shares which you are holding. So, in that case it will be the number of shares we are not holding. So, in that case when you want to find out x is maximum value of 2; in the other case it would be the values complementary would be 3, 4 and 5 corresponding to this is the overall in universal set is 5.

(Refer Slide Time: 10:49)

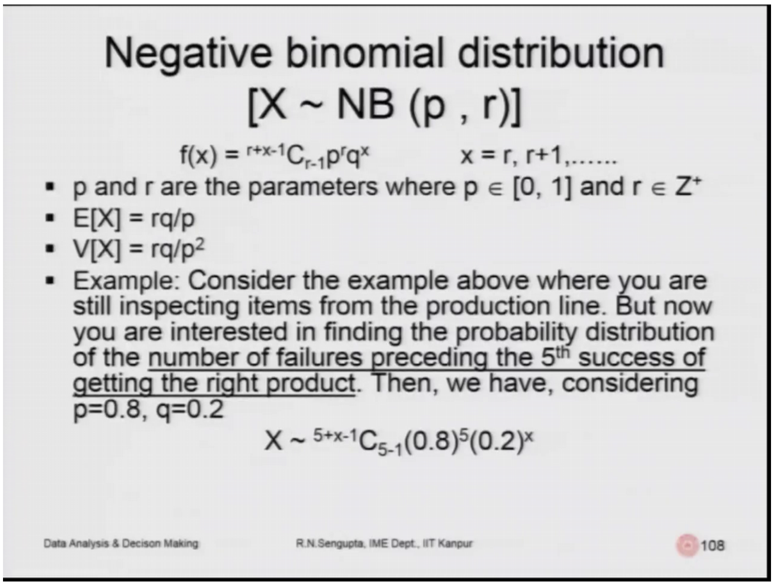


So, Mister Abhishek Bhatia finds that 70 percent of the students have answered question number 1, this is an assignment which you can solve when you understand the concept, but they would be always not impartment part and parcel of the other assignments which you are going to solve each and every week.

Mister Abhishek Bhatia finds the 70 percent of the students have answered question number 1 correctly in his chemistry exam. So; obviously, if you consider p as 0.7, q would be 0.3; a random sample of 10 answer script. So, n is 10 is chosen; so you want to find out what is the probability be there at least 3 of the students have answered question number correctly.

So, at least 3 would be 3 4 5 6 7 8 9 10; if that is the case then you will consider p who have answered the question as 0.7, q as 0.3. Now, if you reverse it in the sense you want to find out those who have not answered. So, and on; obviously, in that case if it is at least 3 so; obviously, it will be the company part would be 0 1 2. In that case p would be 0.3m q would be 0.7 and you would solve the problems accordingly.

(Refer Slide Time: 12:03)



Now, consider the next distribution for the discrete case or a probability mass function for the negative binomial distribution. Negative binomial distribution here again x is the unique, the random variable, NB is the negative binomial the short form for that p and r are basically the parameters which are corresponding to that distribution.

So, f of x is given by now this is not complicated, but it is very straightforward if you understand it. So, the distribution will be r plus x minus 1 C r minus 1 p to the power r q to the power x and p values should mark it here the suffix for p is r which is a fixed value while q is x. And x values are r plus 1 r plus 2 still infinity. So, p and r are the parameters p is; obviously, between 0 and 1 correspondingly q would also be between 0 and 1.

Because p plus q is 1 and the r values would be integer values like 1 2 3 4 it can be 0 also 1 2 3 4 5 6 to infinity. Expected value is given by r q by p and like you to note down this; this formula r q by p and why I will come to that later. Variance is given by r q by p square and an example; consider the example above where you are still inspecting items from production line, but now we are interested in finding the probability distribution of the numbers of failures; preceding the fifth success of getting the right product.

Then we have corresponding to p is equal to 0.8 q is equal to 0.2, the actual distribution would be given by. So, fifth success; so here r is 5; so it will be r plus x which is 5 plus x minus 1 C 5 minus 1 which is the r minus 1, p is 0.8; 0.8 to the power r which is 5; r is 5; so, 0.8 to the power 5 and q which is basically 0.2 to the power x.

Now, listen to the actual framework or the background of the problem; consider that you have, you want to basically find out the distribution till you get the third head or third bad items or third good items whatever it is 3 is basically what we consider an r. So, it can change also.

Now, consider the problem like this the first 2 and so and consider that x is basically the number of such trials you are going to do. Now consider this; if the third one comes as head you stop the experiment. So, fix the third one before that on to the left you would have what are the actual combinations you can have? So; obviously, there are 2 heads already in some place on to the left of the third one; that means you are starting from the left side.

Now, consider there are x plus 2 number of seats or places to be filled up. So, 2 corresponding to the initial first and the second head and x places are corresponding to the tails. Now, if I ask you a question that if there are x plus 3 minus 1 with this 2 x plus 2 number of seeds and you want to fill up with them with either x places being tail and the 2 other places being head.

So, the now total number of combinations would be very simply r plus 2 C 2 or r plus 2 C r minus 1 in this case because you are considering the x minus 1 that is why you are we are considering. So, these places would be filled up; so consider the first is tail next is head next is tail, tail, tail, tail then head then tail and the last one is basically head the third head.

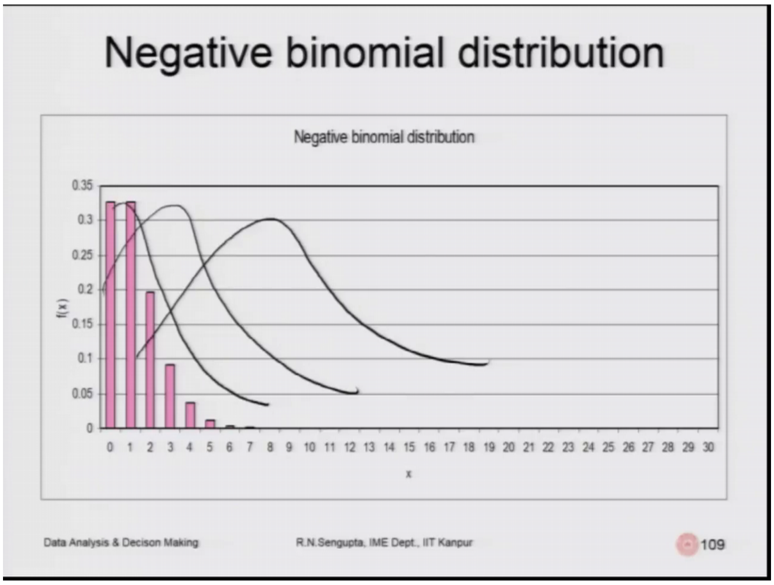
So; obviously, you will you close the experiment. So, the total number of combinations which you can have before the third head comes is basically given r place r minus 1; which is the heads x are the tails they can be combined in the values of r plus x minus 1 C r minus 1 or r plus x minus 1 C x. So, whichever you know the combinations will come out to be 2. Now, if that is done now the next question is what are the probabilities?

So, the probabilities of picking of the tail is q because we are considering that in an unbiased way and; obviously, it will be half enough the probability of picking up the tail is q. So, how many such cases has they will being picked up. So, they are x in numbers; so, the corresponding probability would be q into q into q such x number of times hence it is q to the power x.

Now, the next question is that how many number of p's are there when that mean heads, remember that the third one has already been picked that is why he ended the experiment. So, technically we had r minus 1 heads; so, this is p to the power n minus 1 already picked up, the last place is also p which is head. So, the total number of such combinations of the heads would be p to the power r minus 1 into p; that means it is p to the power r.

So, the total probability would be the total combinations would be r plus x minus 1 C r minus 1 p to the power r q to the power x or else it can be r plus x minus 1 C x p to the power r q to the power x; so, whichever you look if the problem answer is same.

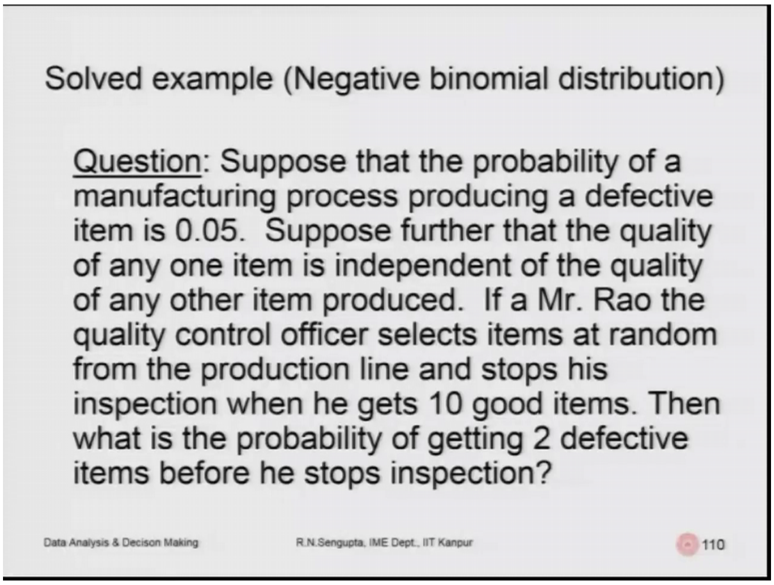
(Refer Slide Time: 18:11)



So, let us consider a problem accordingly. So, we will lay before that as usual this is the diagram. So, the negative binomial distribution depending on what the value of p is what the value of q is you can basically have the distribution like this, sorry I am using a highlighter, I should change it if you change the value of p and q it can be go like this and so on and so forth.

So, you can draw it and understand in accordingly just take here I have taken 30, just take 20; plot them to considered p value as 0.5 q value as 0.5 and fix r; r take as 3. And then you keep changing the values in p and q; such that sum is always 1 and you can get the this nice distribution of that.

(Refer Slide Time: 19:00)

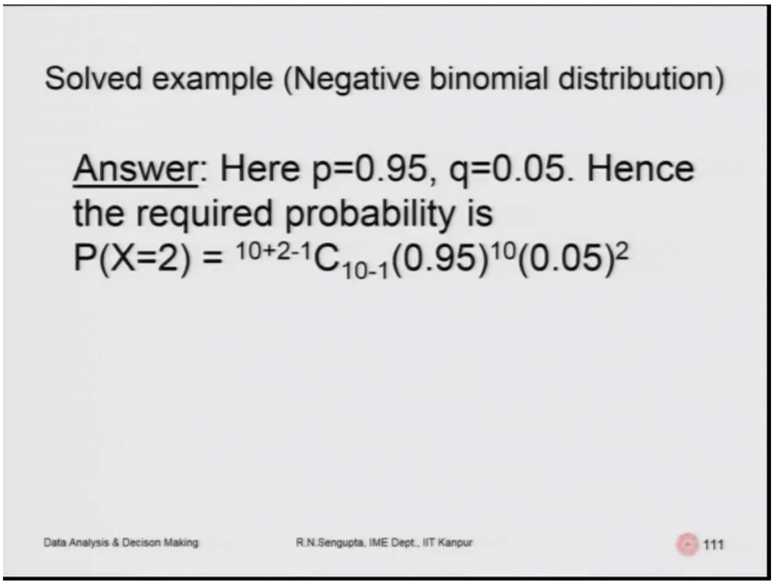


Now, suppose that the probability of a manufacturing process producing and defecting item is 0.05. And also suppose further that the quality of any one atom is independent of the quality of the other item which is produce.

So, they are independent; that means, you are going to the concept of Bernoulli trials; if Mister Rao is the quality control officers who selects items and random from the production line and he stops inspection when he gets 10 good items. So; that means, the tenth one he gets good he stops it, then what is the probability of getting 2 defective items before he stops? Which means that he has to have 9 good items and consider the case; so, these 2 are the defective items.

So, the total combinations we can fill up the places before the tenth one is picked up is good is there are 9 items which are good, 2 items which you are mad they have to picked up accordingly, the defective items probabilities are 0.05. So, 2 items being bad is 0.05 square and 9 items being good, it will be 9 to the power 1 minus 0.05 which is 0.95.

(Refer Slide Time: 20:16)



So, with this background I will basically come to the actual values to illustrate. Here p is point 0.95 which is good, q is basically bad which is 0.05; hence the required probability would be for the case there are 2 bad 9 good beforehand so; obviously, the formula would be. So, if you remember it is r plus x minus 1; so r is what? You have basically picked up r as 10. So, 10 because you will stop it when you get the 10 th item is good.

So, 10 plus 2 minus 1 C x or C r minus 1; so, r is basically 10. So, it will becomes 10 minus 1 or X is basically here given as 2; p to the power r, q to the power x p is 0.95 to the power r is 10, q is basically 0.05 to the power X which is 2; solve that and get the value. So, and this is the just a simple assignment which I will just discussed as I am discussing that for the other distributions also.

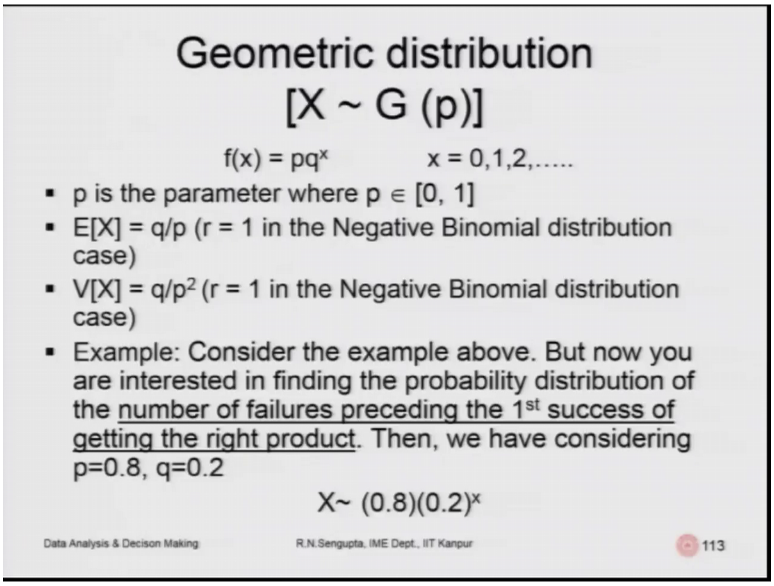
Mister Agrawal is the immigration official of Indira Gandhi International Airport and he knows that the probability of an Indian origin person coming in any flight from abroad is 10 percent; so obviously coming 10 percent foreigners being complimentary of that which is 0.9. He notes down the country of origin of the person and stops when he knows he has noted on exactly 3 people of Indian origin. Then what is the probability of Mister Agrawal noting down 2 non Indian origins before he stops he is checking.

So, if we exactly 3 of Indian origin which means 2 Indian and 2 non Indian because 2 plus 2 was already completed his 4 then the fifth one he gets an Indian he stops. So, this

2 and 2 combinations can be done in the way; that means 2 places as Indian 2 places and as non; non Indians and he basically find out the corresponding combinations.

The values of p and q are given Indian origin coming is 0.1 in non Indian origin people come his 0.9; here the value of x which is non Indian is 2 here also, Indian is 2 plus 1 because there is the third one when he comes he meets the Indian when he stops; so, if I put those values and solve it accordingly.

(Refer Slide Time: 22:27)



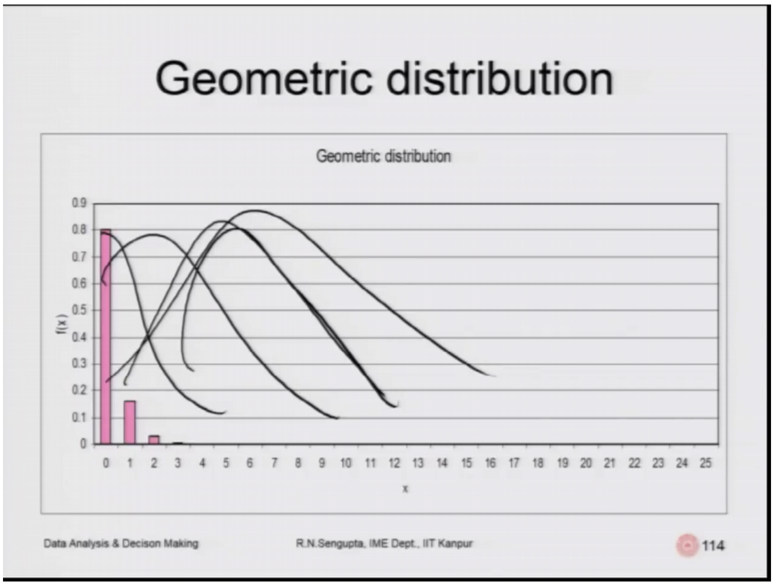
Now, we will consider the geometric distribution. So, if you have understood should the last distribution; negative binomial geometric distribution just simply and extension of that. And, if you remember I told you and please note down the expected value and the variance from the negative binomial which were basically r q by p and r q by p square. So, k q and p can be interchange depending on how you have been basically do you know trying to denote p and q good, bad, red, yellow whatever it is.

Now, this distribution which is random variable X is given by the geometric distribution G and the parameter is p. The underlying fact is here r is one; that means, the moment you get the first bad item or a first good item whichever you are looking at you stop. P is the parameter values are 0 and 1, q is the parameter value which is complimentary 0 and 1. Expected value is given if you put r as 1; so it becomes q by p variance again r as 1 radiance becomes q by b square.

So, consider the example below, but now you are interested in finding the probability distribution on the number of failures preceding the first success of getting the right product. Then we will have basically p as 0.8, q as 0.2; put them in the equation which is because if is r is 1 so; obviously, p to the power r where is 1; it all will be fixed that is it will be p. And you want to find out the number of such bad items; that means, all of them are bad before you pick up the good item.

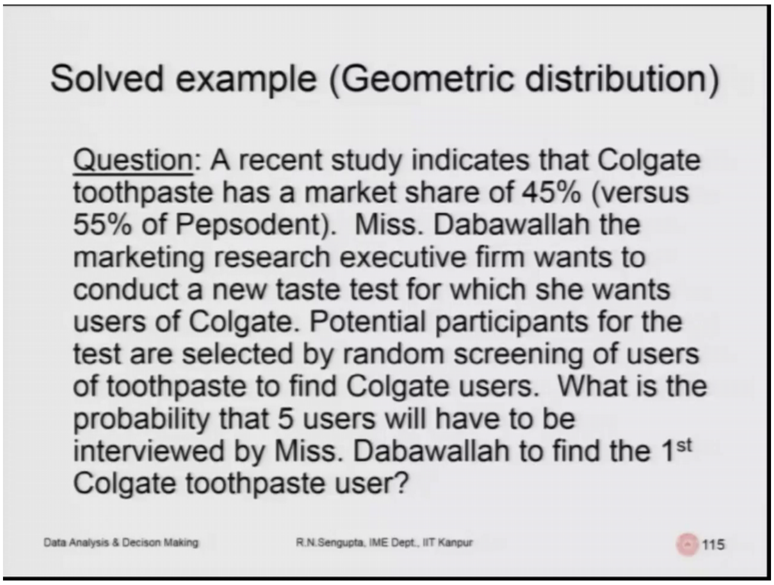
And; obviously, if the places x are all being filled up by bad. So, now; obviously, the number of combinations would be 1 Pico traces you are picking up n from the n or you are technically you are picking up x from the x because r is 1. So, in that case the formula becomes, the initial variable in front of p would basically be 1.

(Refer Slide Time: 24:28)



The geometric distribution would exactly look like the negative binomial; obviously, as the parameter of p and q again will change, but here remember the r value is 1 and you will basically have the distribution. So, it can be like this, this depending on how you have being trying to pick up the values of p and q. So, p is very large, q is very small large means almost standing to 1, q being almost standing to 0 or q being very large means equal to 1; q being and p being basically very small tending to 0; so, you can basically do it according.

(Refer Slide Time: 24:55)

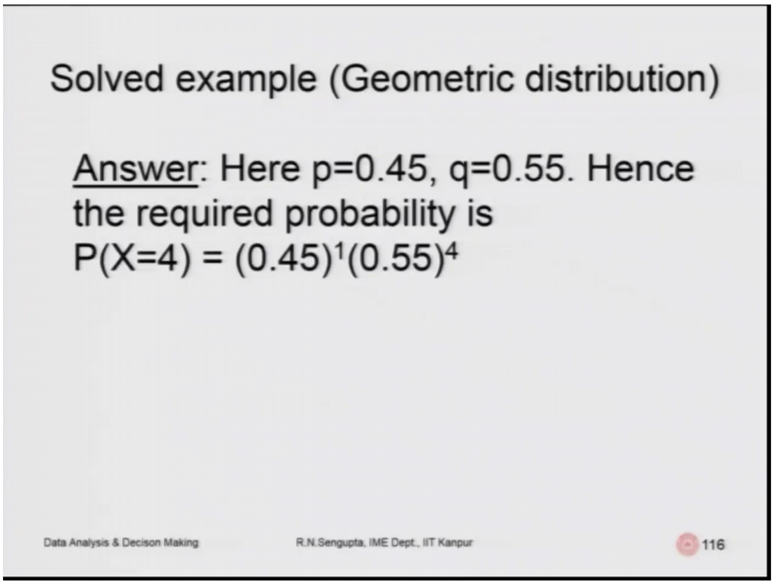


So, geometrical distribution problem a recent study indicates that Colgate toothpaste has a market share of 45 percent versus 55 percent of Pepsodent. Miss Dabawallah; the market marketing research exhibiting firm wants to conduct a new taste test for which she wants users of Colgate. Potential participants for the test are selected by random screening of users of toothpaste to find out Colgate users.

What is the probability that 5 users will have to be interviewed by Miss Dabawallah to find out the first Colgate tooth paste users? So, here 55 percent of Pepsodent, 45 percent are Colgate and she wants to pro find out the probability that 5 users will have to be interviewed by Miss Dabawallah to find on the first Colgate paste user.

So, the all the all the other 4; 5 remember here we listen to the words very carefully, a 5 users have to be tested which means the fifth one is Colgate other fours are have to be the other way round.

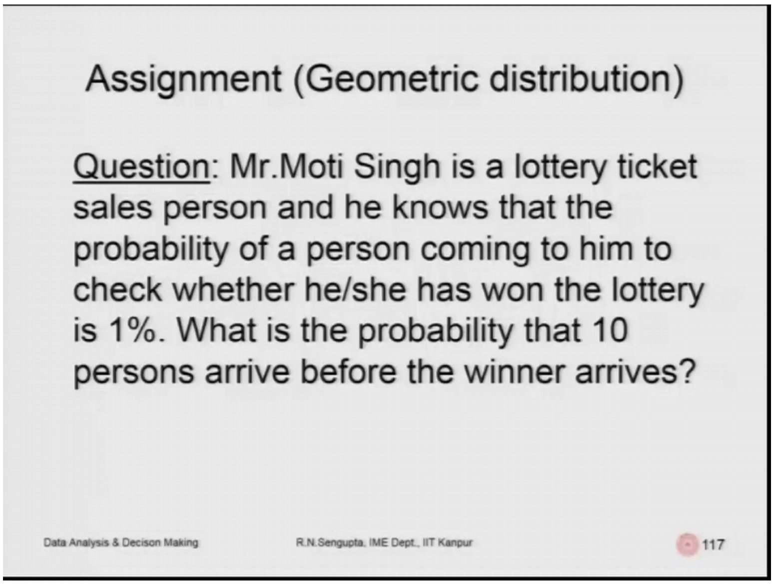
(Refer Slide Time: 25:57)



So, x would be marked accordingly. So, here p is 0.45, q is 0.55 and the probability would be X is 4 because 4 of them are non Colgate. So, here it will be; obviously, the variable in front of p; that means, what you have been using for negative binomial distribution as r plus X minus 1 C r minus 1 or r plus X minus 1 C x would; obviously, be 1.

So, you will have 0.45; 0.45 to the power 1 and the value of the; for the Colgate and or the Pepsodent the variables are such that it will be 0.55 to the power 4 depending on as X as 4. So, if it was in the expect sample it was given that you have to pick up 4 non call; 5 non Colgate in that case it would be 0.55 to the power 5.

(Refer Slide Time: 26:45)



Assignment for the geometric distribution; so, the question is like this Mr. Moti Singh is a lottery ticket sales person and he knows that the probability of a person coming to him whether he or she has won the lottery is 1 percent; what is the probability that 10 percents arrived before the winner arrives?

So, in this case consider whether he or she has won the lottery is 1 percent is 0.01; non winning is 0.99 and what is the probability that the 10 percents arrive before the winner arrive? So; that means, 10 percent have already arrived who have not won; eleventh one is basically a person who has won. So, the total corresponding probability would be; so if it is not winning and if with the probability is 0.99 to the power 10 and p is basically for winning his 0.01 to the power 1.

Because; obviously, that is as we have considered as r as 1, you put the values and solve it accordingly. So, with this I will close this eighth lecture and continue the discussions further on for the other discrete distribution and then come slowly into the continuous distribution have a nice day and.

Thank you very much.