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

Lecture – 09
Hypergeometric, Poisson, Normal Distribution

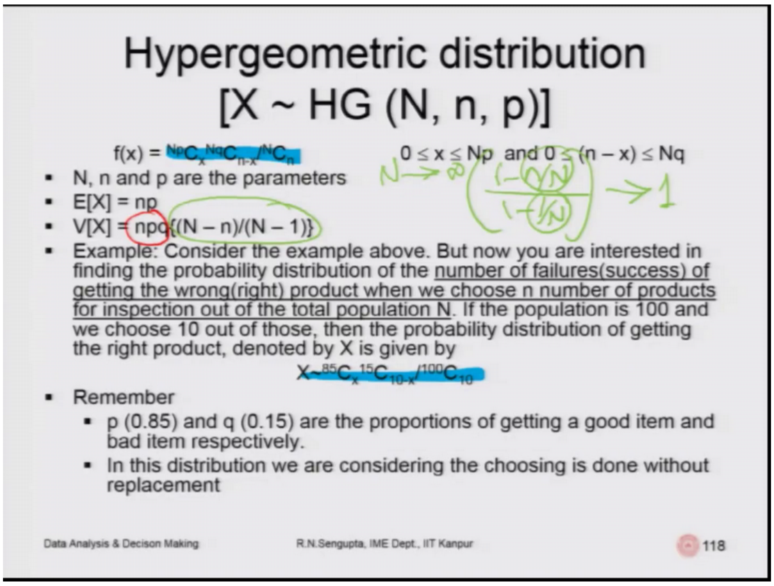
Welcome back my dear friends; a very good morning, good afternoon and good evening to all of you. This is the DADM 1 course under the NPTEL MOOC series. This is a 12 week course for 30 hours. Each week we have 5 classes and each class being or lecture being for half an hour and as you know my name is Raghu Nandan Sengupta from IME department, IIT Kanpur. So, we will be discussing different types of distributions for the discrete case and we were basically giving the formulas mentioning what were the parameters based on which the distribution could be characterized trying to give a simple equation and then, drawing there a graph in the excel sheet and I did mention time and again you try to basically use the excel in order to basically draw the graphs.

That because very easy you can understand I have a feel of how those distributions look like as you change the parameters. We also solved very simple problems and also after solving the problems gave assignments. Assignments I am again repeating, they are not to be confused with the assignments which will be given after each and every week. They would be related to the assignments as we go along with discussing the experiment and on the examples. So, in the next distribution to be considered would be hypergeometric distribution.

Now, we will pause here or I will take a pause and mention a very important thing about all the distributions we have considered till now; considering a hyper geometric distribution would be conceptually a little bit different than and again continue with other distributions. Now, if you remember something when we mentioned about the binomial distribution, the geometric distribution, the negative binomial distribution, the uniform discrete distribution whatever we considered that this is a huge population which is infinite and we basically take observations. So, n what basically the observations which we are taking based on which we could basically classify how the PMF which is the Probability Mass Function looks like, what is the distribution, what is x, what are the parameters and so on and so forth.

Now, in the hyper geometric case, we will consider that there is a huge so called population which is n and this is not infinite. We are picking absorptions from them, but we will also see that as this capital N which is the population tends to infinite, how this hyper geometric distribution basically picks up and actually tends to the characteristics of the other distributions we will see that.

(Refer Slide Time: 02:57)



So, X is the random variable and hypergeometric distribution is generated by HG and the parameters are capital N, small n and small p, where now you should remember where small x are the real s values and basically, the number of observations would be basically technical depending on what the characteristics we want to find out would be between those two limits limiting values. Now, consider the hypergeometric distribution is, let us consider setting an example. You are picking up light bulbs in the shop floor, testing them and saying whether they are right or wrong or you are basically trying to find out whether goods are good or bad or the color is red and black and whatever this is, that means there are two outcomes if you remember the boundary layer trial.

Now, we will also denote that p is the proportions of; technically this is the probability of the good items and 1 minus p which is q is the probability of the bad items. Now, if we denote x as the random variable capital X, obviously x, small x become the Rayleigh's value capital X as the random value you denotes the good items then obviously small x would always be between 0 and n into p. Now, why n into p, because p is the probability so, the maximum number of times the maximum value of the good item would be the total population which is n into p, similarly the value of q would also be between 0 to n in bracket 1 minus p or n into q.

Now, if x is the random variable denoting the good items, obviously Rayleigh's values for the bad items would be small x small n minus x because from the whole lot of capital N, you are picking up a small n trying to find out in that small n, how many are good and how many are bad. So, if say for example, there are 1 million products in the shop floor per day which are being produced and you pick up 500 of them in that one that this case 1 million would basically be capital N which technically should be infinite. If you are considering the actual population distribution and the characteristics small n was basically 500 and if say for example, probability of p or the probability good item is say for example 0.2, then the values of good items would always be between 0 to 500 into 0.2 and that would basically denote the number of items which are good.

Similarly, the number of bad items would be from 0 to 500 in bracket 1 minus 0.2 which is 0.8 because that is the total probability of the bad items. Now, in this case the expected value is n into p we can prove it. Now, the variance in this case for the hypergeometric case would be consisting of two terms and I am mentioning these two terms very specifically for a reason the two terms is n p q multiplied by n capital N minus small n divided by the whole thing divided by capital N minus 1. Now, look it added carefully as capital N tends to infinite. So, let me basically point it out till now I have not been pointing out I will just point out clearly. So, these are the two terms this is n into p I will use another color say for example, green. This is the second term.

Now, look at it carefully in case capital N tends to infinity, then the value I am taking capital N outside. So, it becomes 1 minus n by capital N divided by 1 by 1 by capital N. So, this tends to 1 because these value would basically tend to 0 and obviously, this value would also tend to 0. So, in this case the variance becomes in the long run when capital N basically tends to infinite becomes n p q now, pause here. We have seen for which case the expected value was n into p and we have also seen for that case the variance was n p q that was for the binomial distribution. So, the hyper geometric distribution in the long run as capital N tends to infinity becomes the binomial distribution.

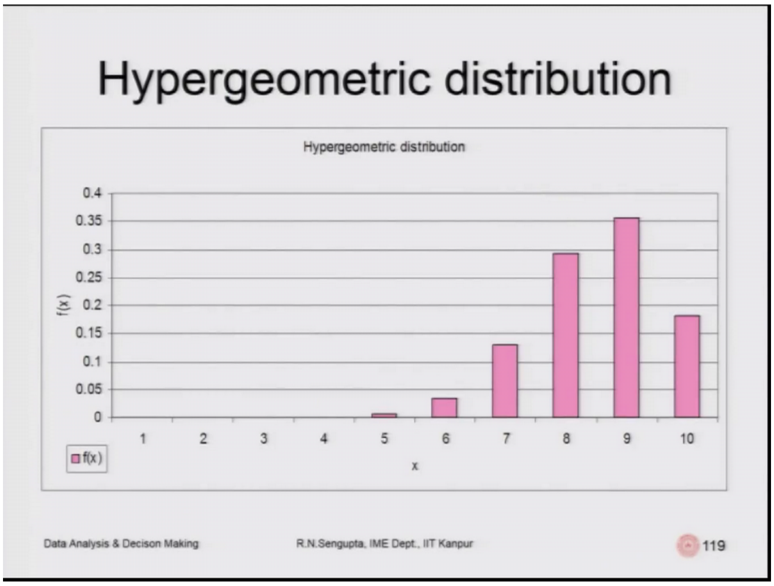
So, example consider the example of a, but now you are interested to finding the probability distribution of the number of failures or successes of getting the wrong or the right item whichever you look like. If you want to find out the right items obviously, x would denote that if you want to find out the bad items or the red items or the good items, obviously which is the complementary of p, that would be basically b q. Then, obviously you can denote the hyper, this hypergeometric distribution accordingly. So, let me continue reading the number of failures of success of getting the wrong or the right item, right product when we choose small n number of products for inspection out of a total population of capital N.

If the population is 100, we are considering capital N as 100. In my example, I took a very large number. If the population is 100 and we choose 10 which is n out of those, then the probability distribution of getting the right product where we denote x as the random variable would be denoted by 85 c x, ok. I did not go into the distribution as such so, if you concentrate here, if you concentrate here, so it becomes basically there are 3 terms; 2 in the numerator and 1 in the denominator. So, N c n capital N c n obviously means that you from the total lot, you are picking up how many ways you can pick up small n. So, that can be in any combinations. So, that would be combination of capital N c n.

Now, if you want to find out the bad items or the good items, it would be like this. There are small n number of chairs on a small n number of places to be fitted out of which x would be good small. Small n minus x would be bad or red, blue, green, white or whatever the complimentary experiment or the event is, so the number of good places can be filled up. Now, let us ask ourselves a question. What is the total number of good items? Total number of good items would be n into p capital N into p. So, you are going to pick up x out of them. Hence, the probability and the combinations would be n p c x. The rest number of items would be filled up the bad items. So, what is the total number of bad items which need to be picked up is small n minus x. So, the total combination would be n into q capital N into q c n minus x and that is basically the corresponding PMF or probability mass function.

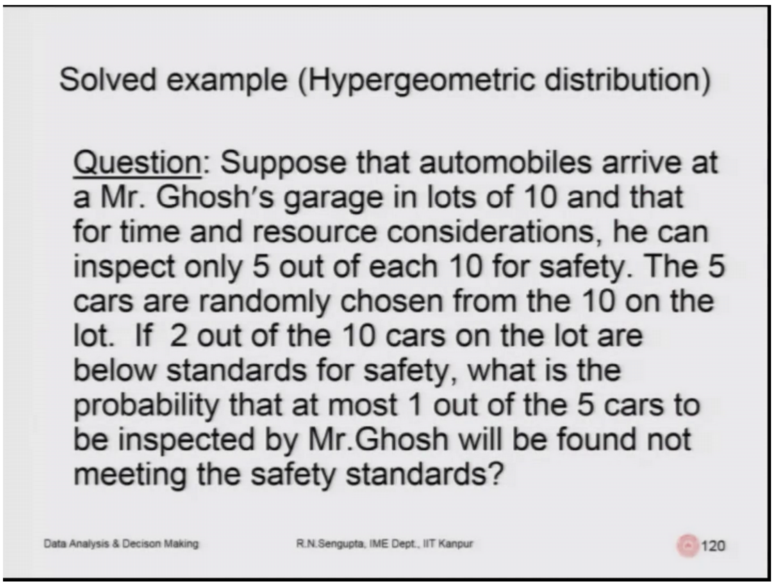
So, the probability mass function in this case if is 10, then you are picking up the observations and 10 is the total number. So, if the success and non-success are given, you can find out the corresponding PMF's as 85 c x which is 85 is n into p and 15 c 10 minus x, where 15 is n p n into q and this small n is 10. That would be divided by 100 c x, where capital N is 100 and small n instant. So, remember here probability p is given by 0.85 and probability q is given by 0.15. So, hence the sum is always 1. So, these are the proportions of getting a good item or a bad item. In this distribution, we are considering the choosing is done without replacement because it is done the corresponding case with replacement, then obviously the corresponding probabilities remain the same as you have seen in the case of the binomial and the other distributions initially.

(Refer Slide Time: 11:17)



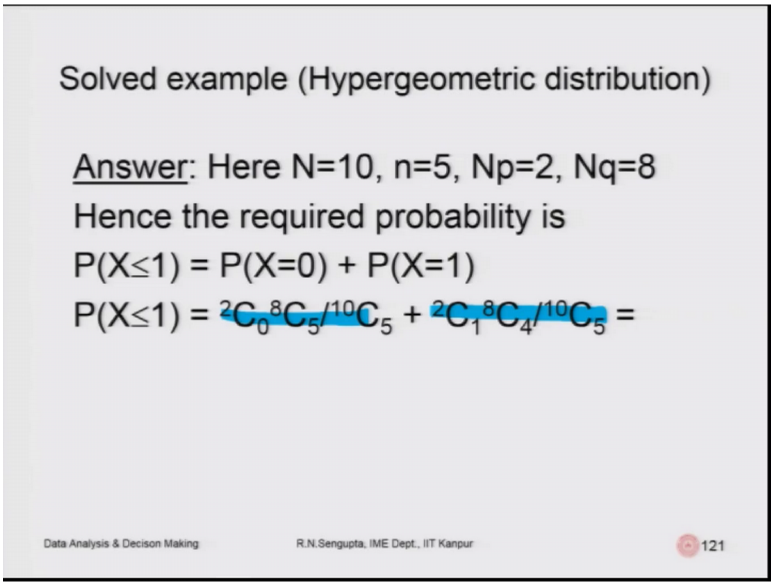
So, considering some values, this is the hyper geometric distribution though where you are measure, we are basically measuring PMF along the y axis and x axis are the values of x.

(Refer Slide Time: 11:31)



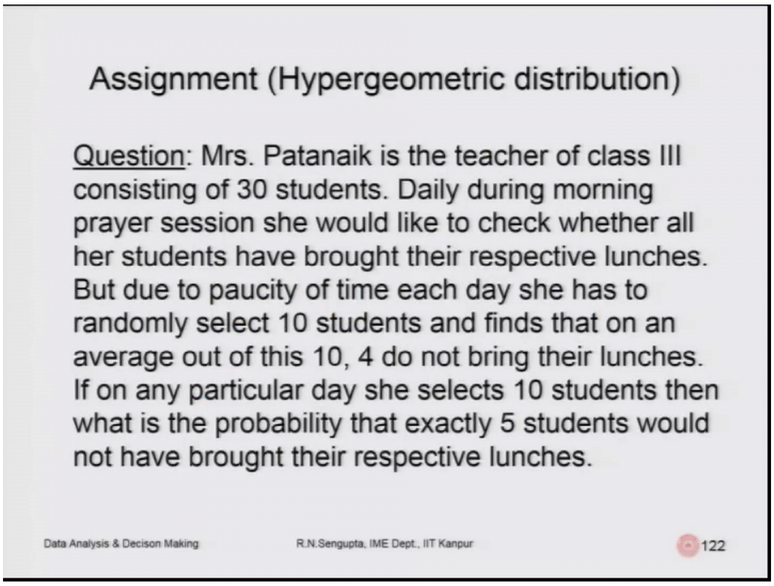
So, let us consider an example. Suppose that automobiles arrived at Mister Ghosh's garage and that for time and resource consideration, he can inspect only 5 out of each 10 of these for safety. The 5 cars are randomly chosen from 10 on the lot. If 2 out of the 10 cars on the lots are below standard for safety, what is the probability that at most 1 out of 5 cars to be inspected by Mister Ghosh's will be found not meeting the safety standards?

(Refer Slide Time: 12:04)



So, here as per the information given capital N is 10, small n is 5. Now, corresponding the probabilities you multiply this p into capital N to find out the total number of good items and total number bad items or defective, non-defective whatever we are denoting. So, in this example capital N into p is equal to 2, capital N into q is equal to 8. Hence, the required probability if you remember you want to find out that they should basically be x value should be less than equal to 1. So, obviously the Rayleigh's values of x would be 0 and 1. So, at x is equal to 0, you have the corresponding PMF value. At corresponding x equal to 1, you have the corresponding PM value. You add them and you get the corresponding value for the case to help Mr. Ghosh to find out the corresponding probability.

(Refer Slide Time: 12:59)

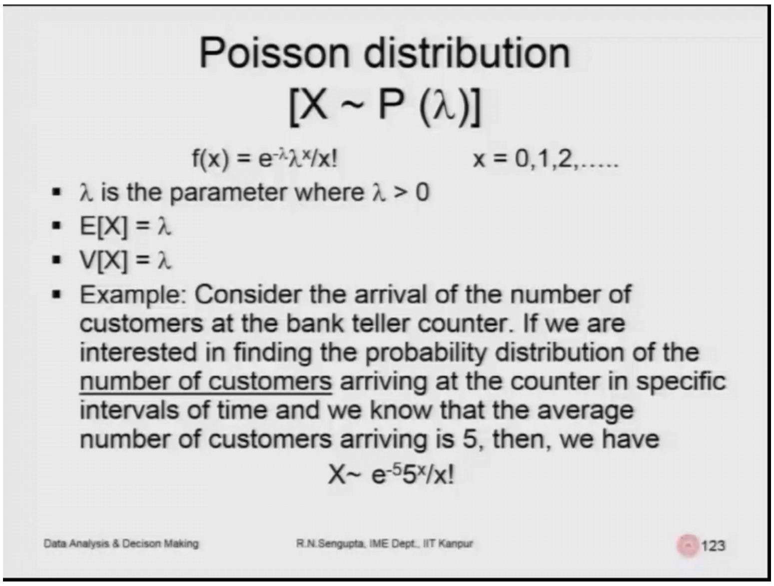


This is again an assignment again I am saying these assignments are to be solved by you just for your understanding.

Mrs. Patanaik is the teacher of class III consisting of 30 students. Daily during morning prayer session she would like to check whether all her students have brought their respective lunches or lunch boxes, but due to paucity of time each day, she has to randomly select 10 students and finds that on an average out of this 10, 4 do not bring their lunches. If on any particular day she selects 10 students, then what is the probability that the exactly 5 students would not have brought those respective boxes? Obviously, we will consider 10 as the capital N and she picks up the students according this and finds.

So, the total number of students is 30. So, capital N is 30, small n is 10 and on information which is given 4 do not bring their lunch boxes and 6 bring. So, obviously 4 would basically if you utilize to find out that the corresponding value actually p is 4 by 10 and q would be basically 6 by 10, but when you put it in the equation n into p or n into q, they are already given which is basically 6 and 4, 4 and 6 whichever you try to denote x or n minus x. So, using that you can find out the corresponding probability, but here it is given that exactly 5 bring their lunch boxes. So, you have to put the corresponding x value as 5, but remember I am saying that 4 and 6 and the corresponding values of capital N into p and capital N into q respectively.

(Refer Slide Time: 14:42)



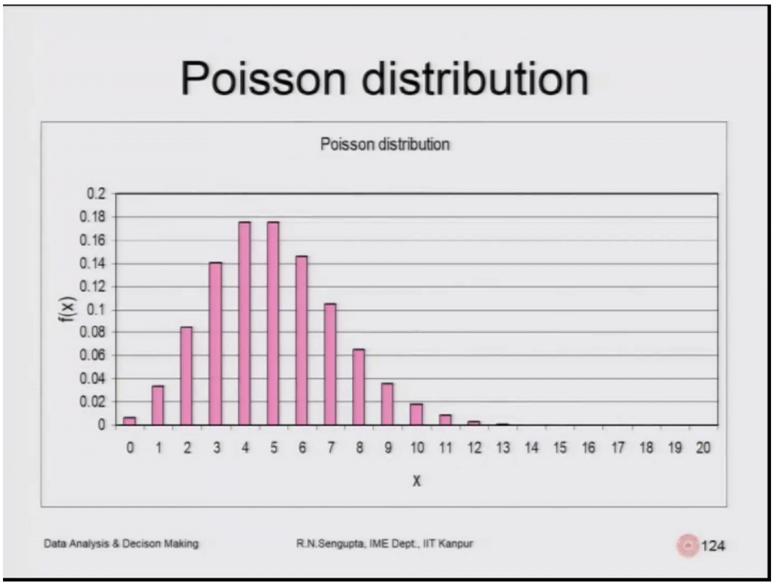
So, now we will consider another discrete distribution which is Poisson distribution. So, that Poisson distribution, the random variable X is given by the Poisson distribution means P and the corresponding parameter is lambda. Now, Poisson distribution is given if you and obviously, the x value which is for the random case is from 0 1 2 3 till infinite. The f of x which is the PMF value is given by e to the power minus lambda, lambda to the power x by x factorial. Here x is the parameter, where x is a lambda.

Lambda is the parameter, lambda is greater than 0 and the expected value and the variance corresponding to Poisson distribution, both are same and that is equal to lambda. So, consider this example to highlight what how, what we mean by Poisson distribution considering the arrival of the number of customers at the bank teller counter if you are interested in finding the probability distribution on the number of customers arriving at the counter in specific interval of time. So, now this point I want to highlight I will come to that later on when we do other concepts of statistics or DADM concepts.

So, we are interested to find out the number of persons who are arriving per unit time or some that unit of time maybe say for example, 5 seconds maybe 5 milliseconds, maybe 5 hours, maybe 10 minutes. What does not matter you would basically are measuring the number of arrivals standing at a point or basically standing at a particular instance, such that you are more interested to find out the inter arrival numbers per unit time. That unit time has to be specified. So, obviously in the long run if you want to find out the corresponding distribution that corresponding time difference or interval of time in the long run should be 10 to 0, hence you will be able to find out what the exact distribution.

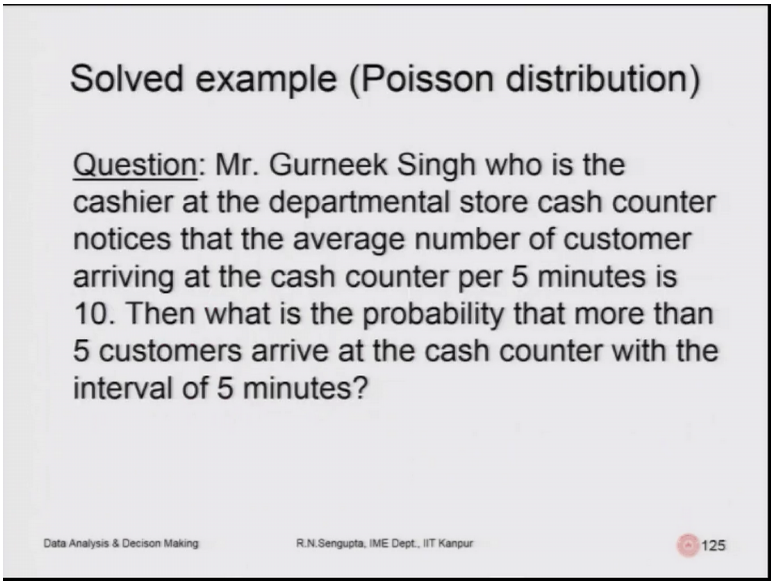
So, when we mentioned the Poisson distribution, later on we will see when we mentioned the exponential distribution on these things, they would mean that per unit time what is basically the numbers which is arriving. So, it finds the probability could distribution corresponding to the number of customers arriving at the counter in specific intervals of time and we want to know we know that the average number of customers arriving is 5. So, hence we put it in the formula and then, find out the corresponding PMF is e to the power minus 5, 5 to the bar x divided by x factorial. You put the values of x and basically solve it for the Poisson distribution.

(Refer Slide Time: 17:18)



So, the Poisson distribution corresponding to some values which you are taken, these are all discrete distribution. Hence, they are just concentrated at those corresponding x values. So, correspond this f of x which is all along the y axis has the PMF value and the corresponding x values are the Rayleigh's values based on which we are trying to find out what f of x is and then, try to plot PMF which is Poisson distribution.

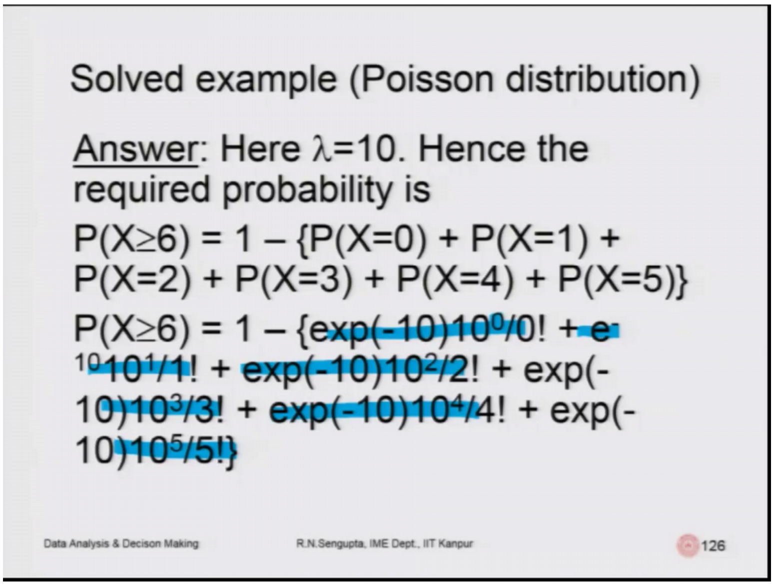
(Refer Slide Time: 17:46)



We will consider here an example. Mr. Gurneek Singh who is the cashier in the department store cash counter notices that the average number of customer arriving at the cash counter per 5 minutes. Now, this case of 5 minutes is basically the inter arrival timings. So, some time difference you are taking based on which you are trying to find out what is the total number of customers you are arriving and so on and so forth. So, this 5 minutes can be made 4 minutes, can be made 1 second, can be made 1 nanosecond, but technically in the long run we will try to basically take the time difference as tending towards 0 and based on that we will try to basically find out what the exact distribution is which we are just stating is the Poisson distribution.

And obviously, it will also mean there are other assumptions also that inter arrival timings on between two intervals are independent and so on and so forth. So, if the cash (Refer Time: 18:40) let me continue reading it arriving at the cash counter per 5 minutes is given by 10, then what is the probability that more than 5 customers arrive at the cash counter with the interval of 5 minutes? So, here lambda would be given as 10 and we will solve the problem. So, let me give the solution.

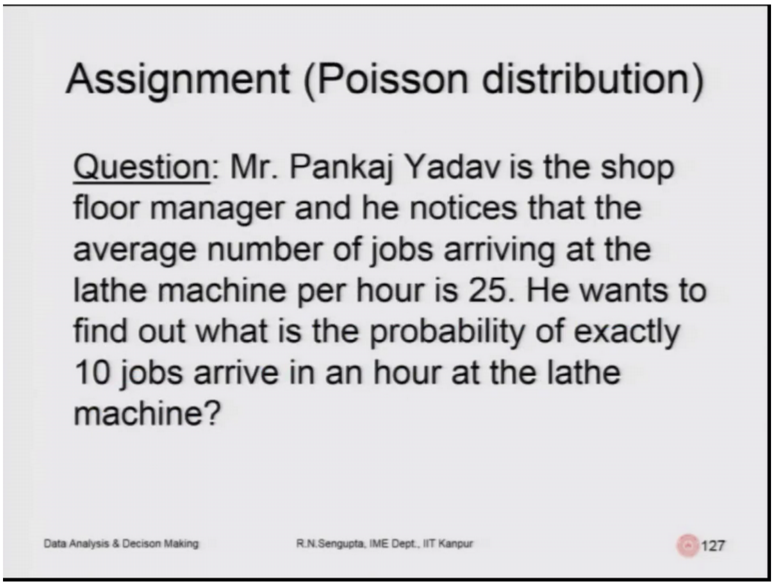
(Refer Slide Time: 18:58)



Here lambda is 10 and we want to find out the number of customers who would be arriving is greater than 5. So, obviously it is 6 7 8 9 10 11 till infinity. So, best would be we find out the probability corresponding to the number of customers arriving in that interval of time of 5 minutes as in 1 is 0 added up with probability corresponding the case when the number of customers arriving is 1, add up when the probability is 2, add up the probabilities 3, add up with the probability 4 and obviously add up the probability 5 because it is greater than 5. So, obviously all the values corresponding to 6 7 8 9 till infinity will be utilized and 1 minus that value will be utilized so, this is what we are doing.

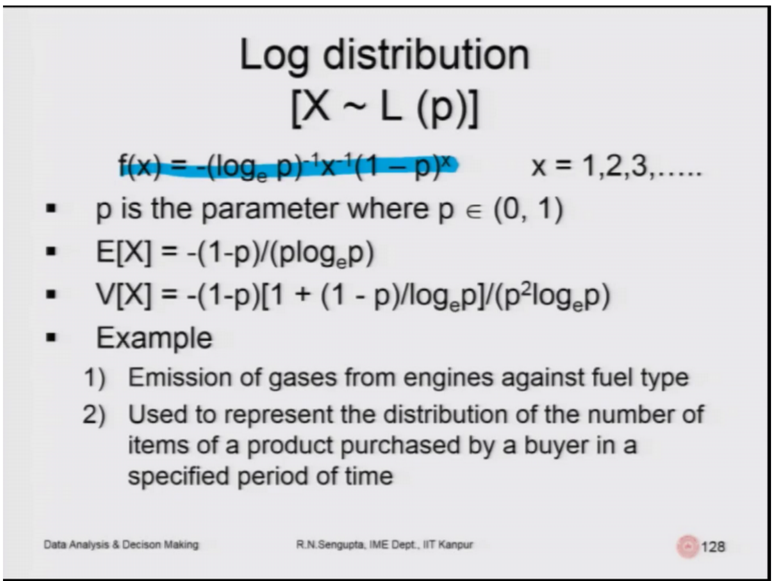
So, it will be 1 minus probability of x is 0, x equal to 1, x equal to 2, x equal to 3, x equal to 4, x equal to 5. So, the corresponding values of x 0 or 1 or 2 or 3 or 4 or 5 are given here. This is for the case of 0 we mark this. So, that will basically give you the hint this is the probability for 1, this is the probability for 2, this is for 3, this is 4, this is for 5 added up 1 minus that value would give you the answer and assignment.

(Refer Slide Time: 20:14)



Mr. Pankaj Yadav is the shop floor manager and he notices that the average number of jobs arriving at the lathe machine per hour is 25. So, here per hour is the inter difference between the time intervals based on which you are trying to note down. So, 25 would be basically lambda. He wants to find out exactly that what is the probability of exactly 10 jobs arrive in an hour and the lathe machine. So, obviously the realized values is given as 10 and the inter arrival time given as 1 hour. So, that is basically what I am being trying to say that you will basically find out the corresponding distribution based on the fact that that interval time is important. So, lambda is 25, plug it in the values and solve this problem accordingly.

(Refer Slide Time: 20:59)



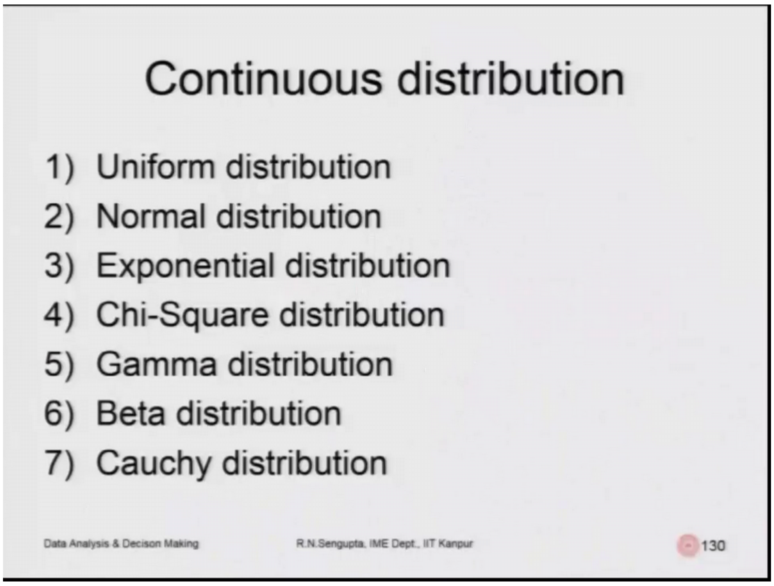
We have the log distribution. So, the log distribution, the x is another variable log, l is the log destitution, the parameters is p and the values of log distributions of x basically tend from 1 2 3 4 till infinity. So, if x is equal to 0 is not possible you know that so, f of x which is the log distribution PMF, f is given as the function as written out here. Here p is the parameter value which basically is between 0 and 1. Expected value is given variance is given.

So, what are the examples where the log distributions are given? This is; do not confuse you with a log normal distribution which you are going to go do later on it is log distribution only. So, here are examples being an emission of gases from engines against fuel type. You want to find out the emissions and you can use the log distribution. They used to represent the distribution on the number of items of product purchased by a buyer in a specified period of time and based on that you can basically use log distribution for other case.

(Refer Slide Time: 22:06)

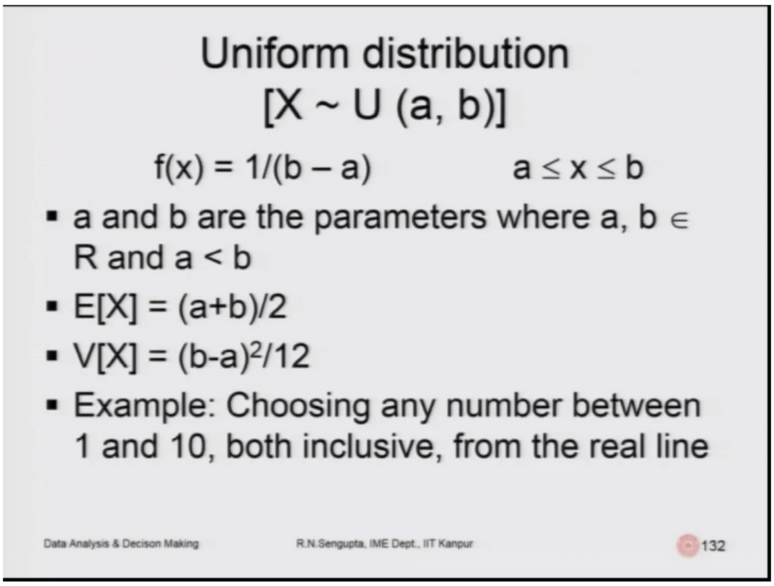
So, the log distribution for some specific values, it looks like this where in x axis you of the Rayleigh's values and along the y axis your basically the PMF. Now, we will consider the few continuous distributions so, in the continuous distribution remember that this x value which we are denoting at the random variable is in this continuous case is continuous and in the discrete case till we have considered now is discrete. Hence, they are PMF or probability mass function depending on the concepts the probabilities are concentrated on masses at different x points.

(Refer Slide Time: 22:47)



So, continuous distributions are uniform, normal, exponential, chi-square, gamma, beta distribution, Cauchy distribution and so on so forth. We have t-distribution, F-distribution, log-normal distribution. If you remember I did mention few minutes back we have the Weibull distribution, double exponential distribution, Pareto distribution, logistic distribution and on these values.

(Refer Slide Time: 23:13)



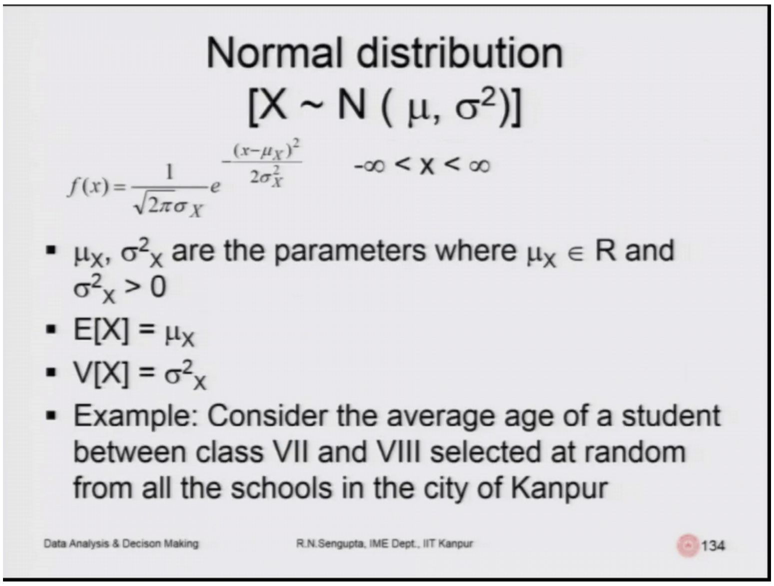
This uniform distribution what we are going to consider is exactly almost similar concept with respect to the uniform discrete case. In this case, x is the random variable, u is basically the symbol with capital U denotes the uniform distribution and the parameters are e and b, both inclusive, so f of x or PDF now it is not PMF. Now, it is Probability Density Function is given by 1 by b minus p a. So, you have basically trying to find out that how the total probability will be divided between these ranges in the uniform case. The x values are inclusive a to b between a and b; a and b are the parameters, where a and b are real numbers and b is greater than a. The expected value is given by a; a plus b by 2 obviously which is the center of gravity. You can understand that if you are basically adding up a plus b and dividing by 2, you basically have a distribution which is uniform and based on the center of gravity where you can balance the weight that would be the expected value.

The variance is given by b minus a whole square divided by 12 examples being choosing any number between 1 and 10, both the inclusive from the real line. So, you will basically find out so real and can be any values you can pick up. It is not only discrete picking up of Jett's marked by either 1 2 3 4 5 6 or 1 3 5 7, whatever it is.

(Refer Slide Time: 24:34)

The uniform distribution would look like this where in again in the x values along the x axis, we have the Rayleigh's values and along the y axis, you have the PMF values.

(Refer Slide Time: 24:46)

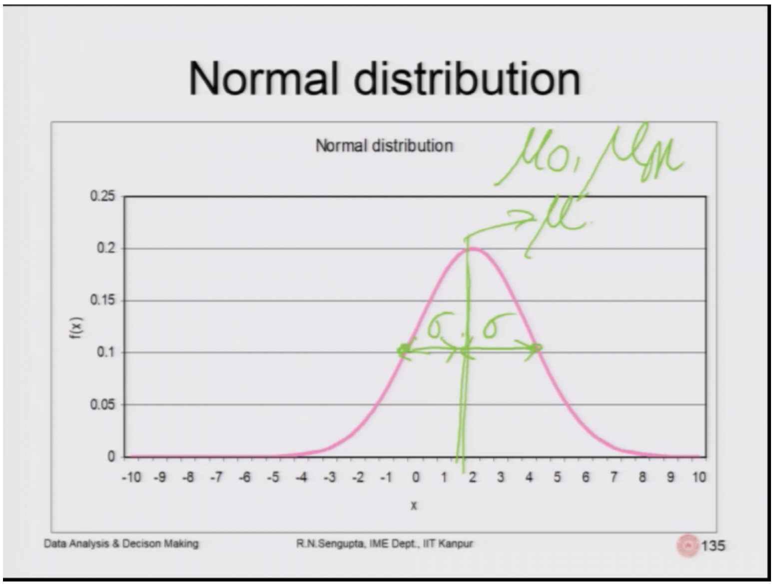


Now, we will consider the normal distribution and we will spend some time and come back to the normal distribution time and again. So, now normal distribution is given for the case, where x is the random variable, n is the symbol which denotes the normal distribution and the parameters are mu and sigma square. We will see later that mu is the expected value and sigma square is the variance or square root of the sigma square is basically standard division which means is sigma or the standard deviation.

So, here the PDF is given by as you can see from the formula 1 by square root of 2 pi and then, denominator we have sigma suffix x multiplied by in the numerator you have it exponential or e to the power minus in the bracket x minus mu whole square divided by 2 sigma square. Here mu suffix x and sigma square suffix x are the parameters, such that new value can be basically the real lines from minus infinity to plus infinity and obviously, sigma square would be greater than 0.

The expected value is given by mu suffix x or mu. So, I have given a suffix x it denote the random variable is x. So, if the random variable is y, it will be denoted by mu suffix y variance is given by sigma square example being considered the average age of a student between class 7 to 8 and selected at random from all the schools in the city of Kanpur and you want to find out what is the average age of the students. You could basically have very simple exponential distribution. It can be utilized and normal distribution is utilized correspondingly in many of the cases.

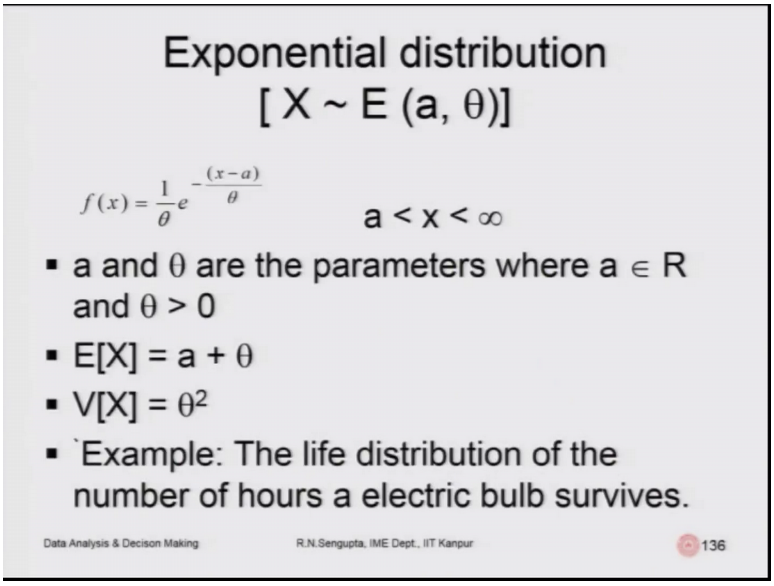
(Refer Slide Time: 26:32)



The normal distribution looks like this, where along the x axis we have the x values and technically, the x values are from minus infinity to plus infinity. So, along f of x you have basically the PDF values and remember here the normal distribution would be such that the overall concentration of the mean, the median and the mode would be the same point which is mu suffix x and for this case, we have a value and I will just denote this is the mean value. So, this will be mu, this is also the mode and also, the median;' mode is basically the highest value. The mu suffix 0 and median is basically which it divides the total distribution in two parts 0.5, 0.5. So, this is the media and you can find out the corresponding values for the standard deviations or for degree. So, standard deviations would be the shift.

So, I just denoted randomly and come to this; these concepts that were that the standard deviations are basically these are from left to the medium value to which point? So, this is what is important. So, we will have some scientific reasons also to give the discussion based on those.

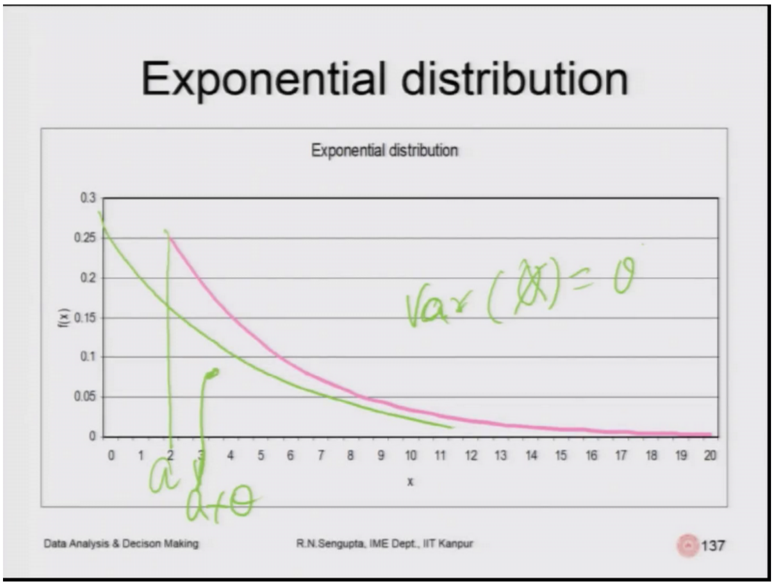
(Refer Slide Time: 28:04)



Next we will consider the exponential distribution. Exponential distribution is given by x is the random variable; the exponential distribution is given by E and the parameters are a and theta. Now, remember a, and theta in some cases, it can be made to 0 depending on what are the assumptions. So, f of x which is the PMF is given by 1 by theta e to the bar or exponential to the power minus in bracket x minus a divided by theta and x values basically are all from a to infinity. If a 0, if it is a is sorry, a some positive value and if a is 0, then obviously x would be from 0 to infinity.

So, n theta the parameters where a is in the real line and theta is basically greater than 0 and the expected value is given by a plus theta, if a is 0, the expected value is theta and the variance is given by theta square whether a is present or not. Example being the live distribution of the number of hours and electric bulb survives. So, now let me come to the diagram and then, I will explain the concept of a you will understand.

(Refer Slide Time: 29:15)



So, now in this exponential distribution looks like as given in this figure, where along the x axis, you have the x values and along the y axis, you have basically the PMF. Remember one thing the value what we are drawn here is a, and if say for example, the case of a 0, then the distribution would be going from here. So, when I am basically drawing the value of and marking the value of a plus theta is basically the center of gravity a plus theta. So, the value would be given by theta so, this is the variance we will consider accordingly.

So, with this case and this example and end the lecture, 9th lecture and continuing the discussion of more of the distributions and give examples and later on come into the details about the normal distribution, what is the significance of normal distribution. If you remember faintly, I did mention that we will consider the t-distribution, the F-distribution in the chi square. So, they would be coming out later on that why in the normal distribution is important and what is the significance of these distributions. I will end this class with this node.

Thank you very much and have a nice day.