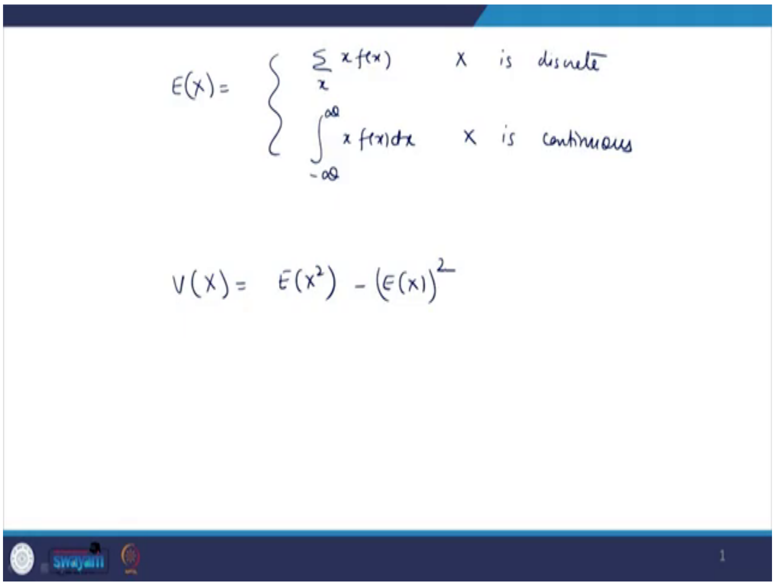
Essential Mathematics for Machine Learning
Prof. S. K. Gupta
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 34
Probability Distributions

Hello friends, welcome to lecture series on Essential Mathematics for Machine Learning. So, in the last lecture we have seen that if a random variable is given to you it is p.m.f or p d f is given to you that how can you find expectation or variance of that x.

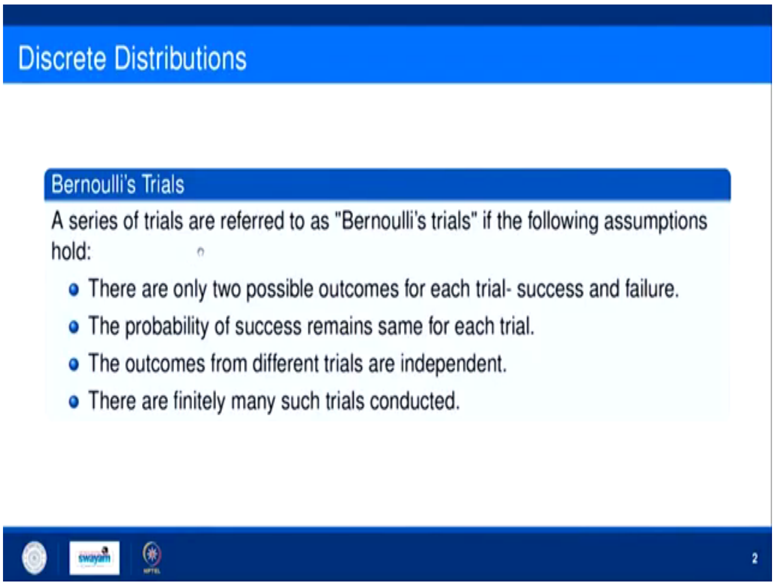
(Refer Slide Time: 00:47)



So, we have seen in that lecture that; if we want to find out expectation of x then this is nothing, but summation x, x of x when x is discrete. And, it is integral from minus infinity to plus infinity x f x d x when x is continuous. And variance of x is simply expectation of x square minus expectation of x whole square.

So, now in this lecture we will see that what are the important some of the important distribution functions; mainly Gaussian distribution or normal distribution which is used in machine learning and how can you find out mean or variance of those distribution functions. So, first distribution function is basically we will use that is binomial distribution function; discrete case. So, before going to binomial distribution what do you mean by Bernoulli trials?

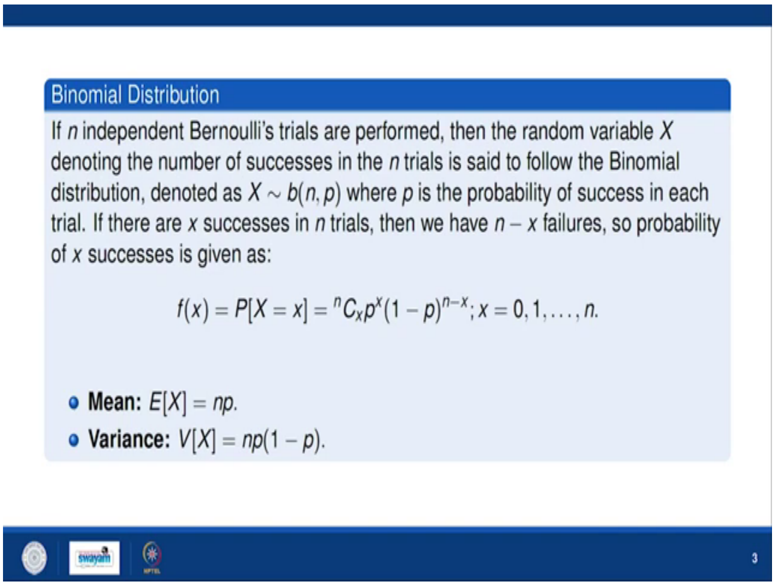
(Refer Slide Time: 01:54)



So, a series of trials are referred to as Bernoulli's trials; if the following assumptions hold. So, what are those assumptions? Number 1; there are only two possible outcomes for each trial either success or failure. So, if we are doing some we are doing some experiment and there are only two possibilities either pass or fail either success or failure; so that is the first assumption for Bernoulli's trial.

The second assumption is the probability of success remains same for each trial it will not change ok. Next is the outcomes from different trials are independent ok; the first trial will not affect the other trial, the third second trial will not affect the third trial and so on. And, there are there are finitely many such trials conducted it is not infinite this is finitely many. So, if these four assumptions hold; then the trials then the series of trials which we are conducting are called Bernoulli's trials.

(Refer Slide Time: 03:04)

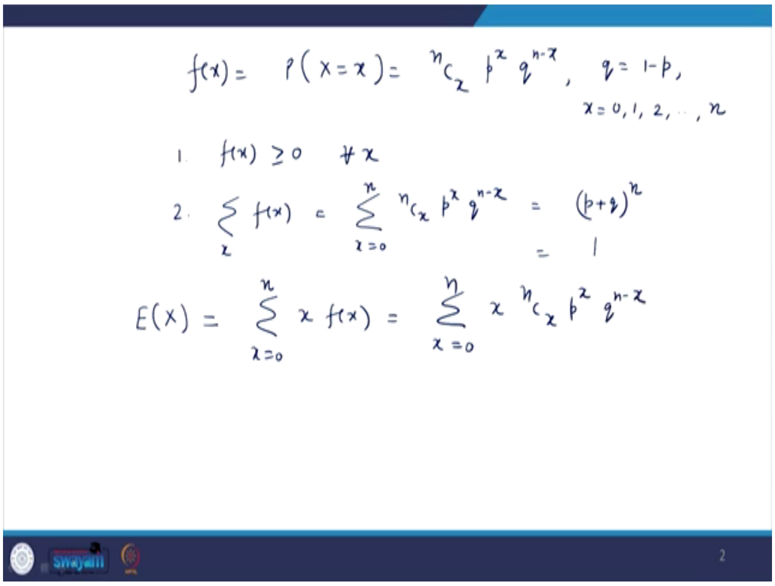


Now, let us come to binomial distribution what is what do you mean by binomial distribution. So, if n independent Bernoulli's trials are performed then the random variable X denoting the number of successes in the n trials is said to follow the binomial distribution denoted by X b n p this is a notation; where p is a probability of success in each trial here p is the probability so c is (Refer Time: 03:35) and denote the number of trials total number of trials.

If there are x successes in n trials then we have n minus x failure of course, because if out of n trials we are having x success; that means, remaining trials are failures and; that means, n minus x are failure. So, the probability of x success will be given by see the probability of x success; now out of n there are x success so it is n C x; n C x is simply factorial n upon factorial x into n minus x factorial n minus x; multiplied by p x which is the probability of success of getting x trials x success.

And this is 1 minus p is a failure and there are n minus x failures; so power is n minus x and of course, x is varying from 0 to n meaning when x is 0; that means, all are failures. When x equal to 1; that means, 1 success x equal to 2; that means, 2 success and when x equal to n similarly; that means, n success all success. So, this distribution is basically called binomial distribution.

(Refer Slide Time: 05:01)



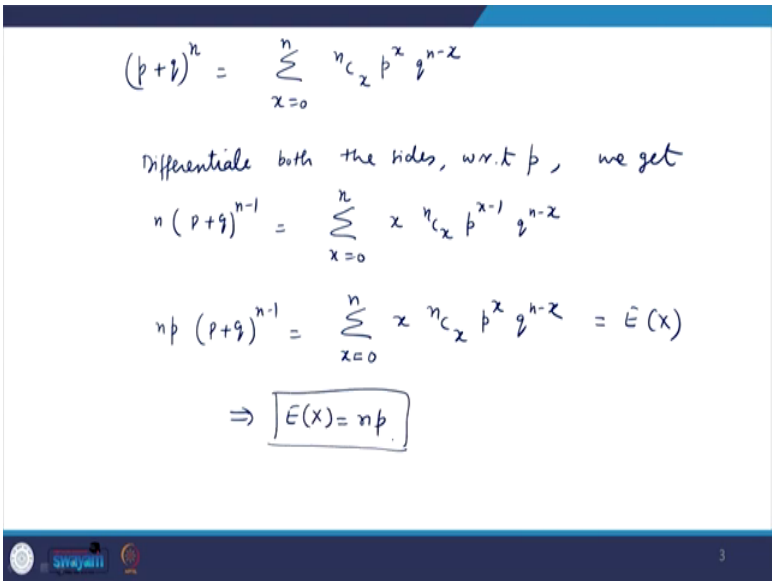
Now, of course, if you see this in this particular distribution which is which is f x equals to p x equal to x which is given by n C x p raise to power x or q raised to power n minus x where q is nothing, but 1 minus p and x is varying from 0 to n. So, if you see this distribution of course, for any x between if for any x from 0 to n it is always non negative. So, first condition is that f x is always greater than equal to 0 for all x that is of course, satisfied.

Now, if you see if you see a sum of all f x for x. So, if you calculate this so it should be 1 if it is a distribution function; so let us see. So, it is n C x p raise to power x q raise to power n minus x and x is varying from 0 to n.

So, it is nothing, but p plus q whole raised to power n as per the binomial theorem. And p plus q is 1; so 1 raised to power n is 1. So, how can we find out? So, for a discrete case it is a p.m.f how we can find out expectation of this random variable x and how we can find out variance ok.

So, if you want to find out expectation of x; expectation of random variable x; so that will be summation x f x x is varying from 0 to n. So, this is nothing, but summation x x from 0 to n. So, this is x this is n C x, p raised to power x q raised to power n minus x.

(Refer Slide Time: 06:45)



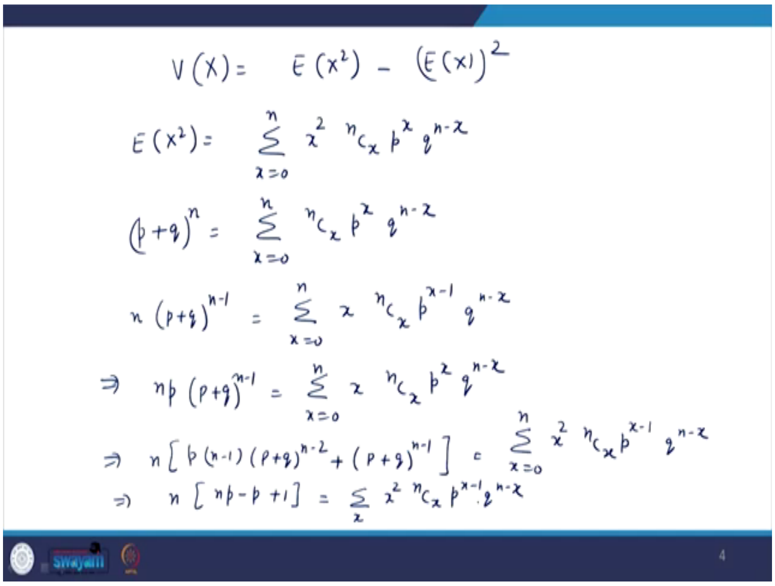
So, let us try to compute this value. So, let us let us expand this p plus q whole raised to power n; this is nothing, but summation x varying from 0 to n this is n C x, this is p raise to power x q raise to power n minus x ok. Now let us differentiate both side with respect to x. So, differentiate both sides both the sides with respect to x.

So, what we get? We get both side to p sorry with respect to p. So, it is n p plus q raised to power n minus 1 it is summation x varying from 0 to n. So, it is x and C x, p raise to power x minus 1, q raised to power n minus x ok.

Now, we want this value x and C x p raised to power x we need q raised to power n minus x. So, you multiply both sides by p if you multiply both sides by p. So, it will be n p into p plus q whole raised to power n minus 1 plus summation is equal to summation x from 0 to n, x and C x p raised to power x q raised to power n minus x.

So, this is nothing, but expectation of x and that this value is nothing, but because p plus q is 1; so it is nothing, but n p. So, this implies expectation of x is n p. We can say that mean of binomial distribution is n p where n are the total number of trials and p is a probability of one success ok; probability of succession in probability of success basically p ok. Now, if you want to compute variance of x. So, how can we compute variance of binomial distribution ok.

(Refer Slide Time: 08:55)



So, variance will be given by we know that is expectation of x square minus expectation of x Whole Square; so this we know that it is n p. So, we have to compute this so expectation of x square as per the definition it is summation x from 0 to n x square into n C x, p raised to power x q raised to power n minus x it is sorry it is x square ok.

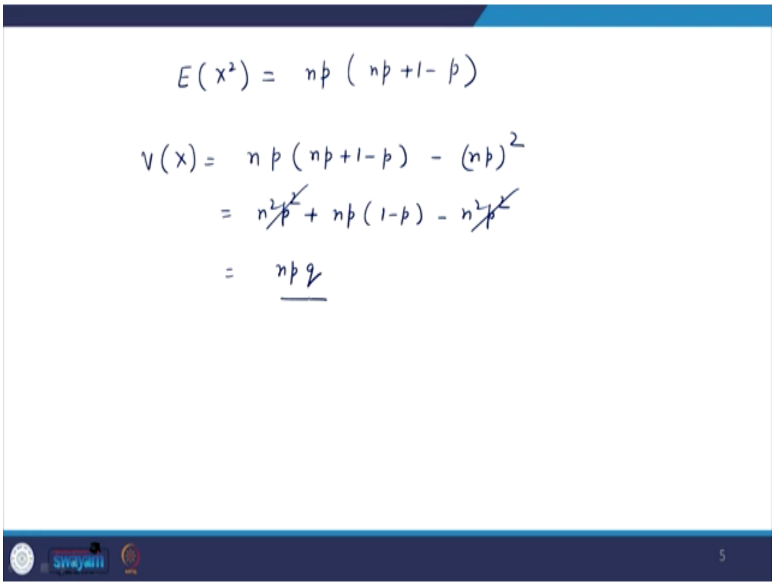
So, we want to find out this value. So, again what is what is p plus q raised to power n? As we have already seen it is summation x from 0 to n it is n C x, p raised to power x, q raised to power n minus x you differentiate both sides with respect to p. So, it is n p plus q whole raised to power n minus 1 summation x varying from 0 to n it is x and C x, p raised to power x minus 1 q raised to power n minus x.

Now you want x square here for this; so if you want x square here. So, first you have to make it p, p raised to power x. So, you multiply both sides by p if you multiply both sides by p; so it is p raised to power x q raised to power n minus x. Now again you differentiate with respect to x because you want x square. So, again you differentiate with respect to x; so that is nothing, but n times.

So, first as it is derivative of second will come here plus second as it is derivative of first will come here and that is nothing, but summation x varying from 0 to n it is x square n C x, p raised to power x minus 2 into q raised to power x minus 1 sorry it is n q raised to power n minus x. Now, as p plus q is 1. So, p plus q is 1 so it is n it is n p minus p plus 1 is equal to this value ok.

Now for this again you want to compute this; so for this you again multiply both sides by p so you multiply both sides by p.

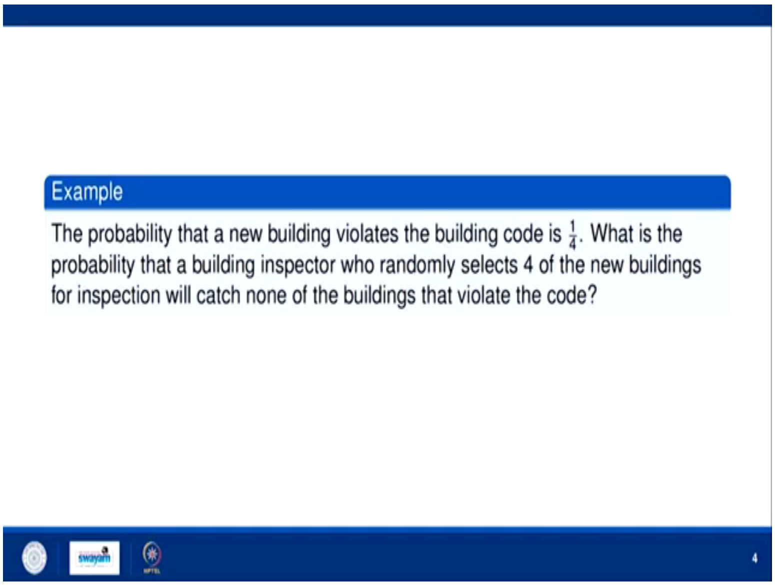
(Refer Slide Time: 11:41)



So, what is expectation of x square finally, you will get; this you get as n p into n p plus 1 minus p. Now what will be variance of x now finally; variance of x will be expectation of x square minus expectation of x whole square.

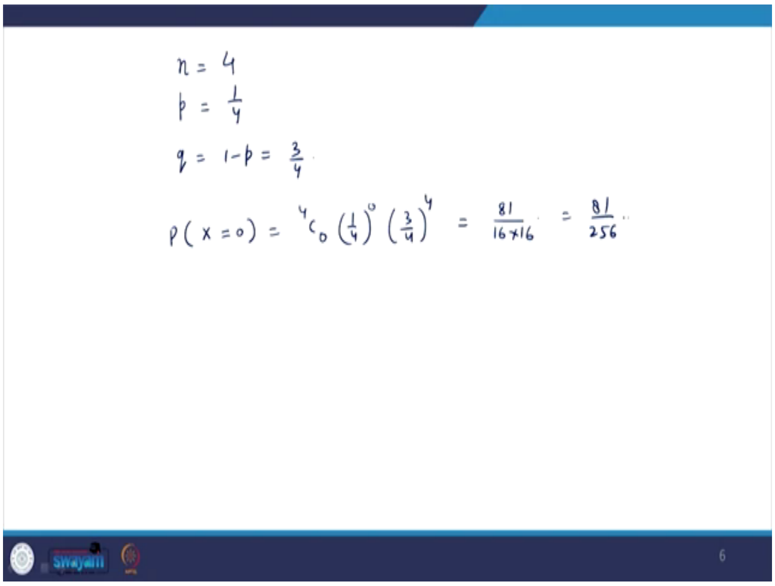
So, that will be n square p square plus n p into 1 minus p minus n square p square; so this cancels out. So, it is n p q or n p into 1 minus p; so that is how we can compute mean and variance of this random variable ok.

(Refer Slide Time: 12:27)



So, let us discuss this example; so this example is based on binomial distribution. The probability that a new building violates the building code is 1 by 4. What is the probability that a building inspector who randomly selects 4 of the new buildings for inspection will catch none of the buildings that violate the code?

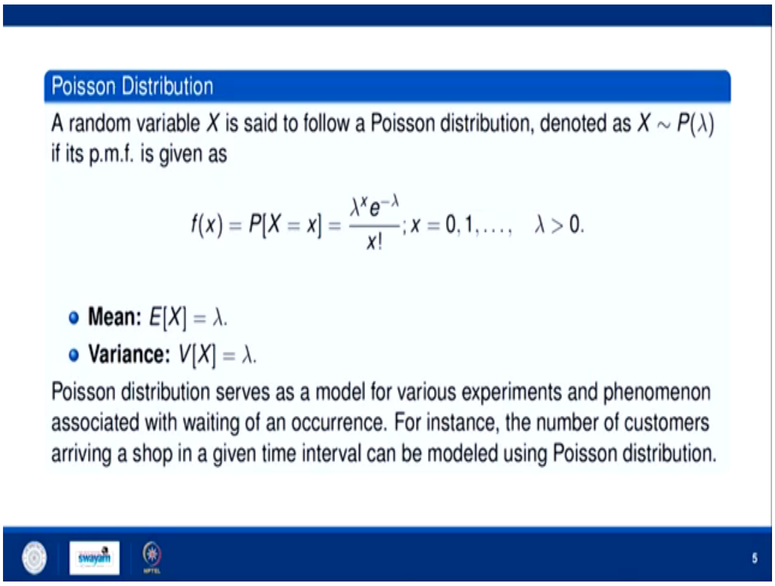
(Refer Slide Time: 12:50)



So, what is n? First identify n? What is n? Number of new buildings are 4. What are the probability of success? Here success is violating the code and that probability is 1 by 4 so; that means, 1 by 4. So, what is q? q is 1 minus p; that means, 3 by 4 not violating the code. We want that none of the building none of the that none that out of all this 4 the inspector will catch none of the building that violate the code; that means, for x equal to 0 we have to find.

So, so; that means, x equal to 0 and that is nothing, but 4 C 0 from here you can get it is 0 and; that means, 3 by 4 raised to power 4. So, that means, 81 upon 16 into 16 that is 256; so it is 18 upon 256 ok. So, in this way we can find out.

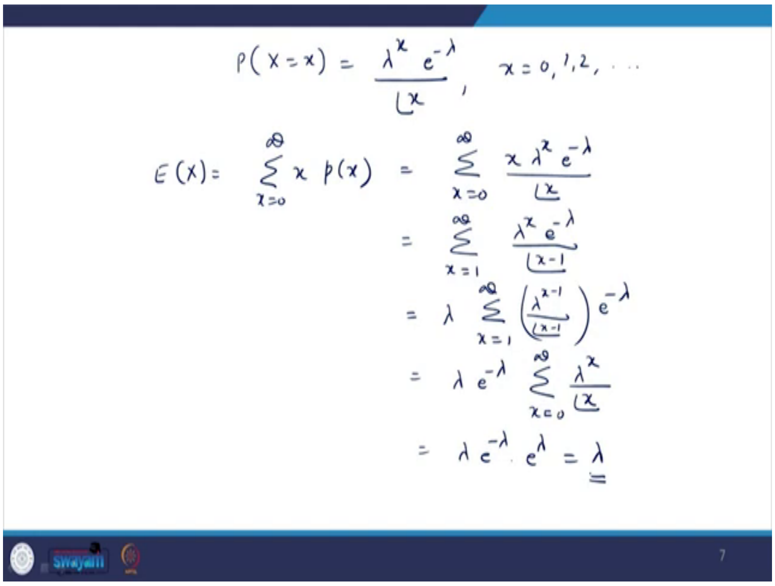
(Refer Slide Time: 13:53)



Now, the next distribution is Poisson distribution. So, let us suppose a random variable X is follow a Poisson distribution which is denoted by P lambda. This is here parameter is only 1 which is lambda; in the binomial distribution we are having 2 parameters n and P; here we are having only lambda. And it is p.m.f is given by lambda raised to power x E raised to minus lambda upon factorial x and x is varying from 0 to infinity lambda is greater than 0.

So, one can easily verify that it is a p.m.f because this is greater than equal to 0 for every x number one. And if you if you take the summation over entire X from X is varying from 0 to infinity then this sum comes out to be 1 that is easy to verify. Now if you want to find out the mean or variance of this random variable x which is following Poisson distribution. So, how can you find; so let us see.

(Refer Slide Time: 14:51)



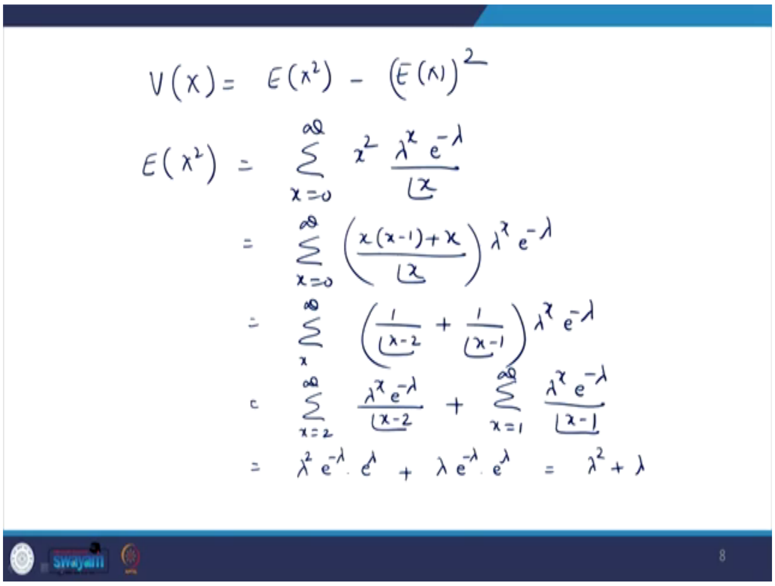
So, what is the p.m.f of this? This is lambda raised to power x e raised to power minus x upon factorial x; x is from 0, 1, 2, and so on. So, what is expectation of x? Here it is summation x p x ok. So, what is the p.m.f of this? It is lambda raised to power x e raised to power minus lambda fine. So, it is here it is e raised to power minus lambda ok. So, it is x from 0 to infinity.

So, it is summation x from 0 to infinity it is x lambda raised to power x E raised to power minus lambda upon factorial x; so that will be summation. So, it is lambda raised to power x E raised to power minus lambda upon factorial x minus 1; of course, when x is from 1 to infinity now because for 0 it is; obviously, 0. Now you can take 1 lambda common so it is x from 1 to infinity.

Again it is sorry it is e raised to power minus lambda. So, it is lambda raised to power x minus 1 upon factorial x minus 1 and it is e raised to power minus lambda. So, this e raised to power minus lambda will come out because it is free from x. So, it is lambda e raised to power minus lambda. And when x is 1 it is 0 upon 0 lambda raised to power 0 upon factorial 0.

So, we can write x from 0 to infinity lambda raised to power x upon factorial x and this value is nothing, but e raised to power lambda. So, it is lambda into e raised to power minus lambda into e raised to power lambda which is lambda. So, we can say that mean of random variable following Poisson distribution is lambda.

(Refer Slide Time: 16:43)



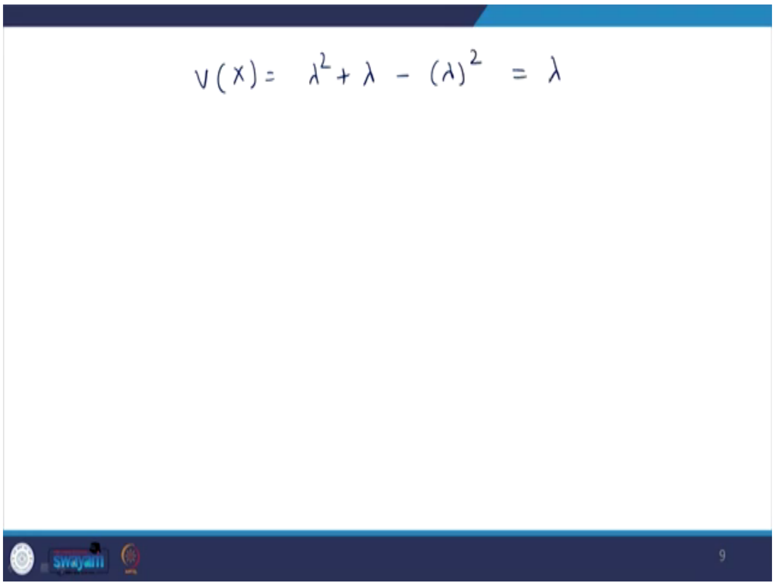
So, now if you want to find out it is variance. So, it is variance will be; expectation of x square minus expectation of x whole square as we already seen. Now what is expectation of x square for this is summation x from 0 to infinity x square into lambda raised to power x e raised to power minus lambda upon factorial x so this is x from 0 to infinity.

Now this can be written as x into x minus 1 plus x upon factorial x into e lambda raised to power x into e raised to power minus lambda which is x from 0 to infinity, 1 upon factorial x minus 2, plus 1 upon factorial x minus 1 into lambda raised to power x e raised to power minus lambda. So, here for this summation x starts from 2 and for this summation x starts from 1.

So, we can say that it is summation x from 2 to infinity it is lambda raised to power x e raised to power minus lambda upon this plus summation x from 1 to infinity lambda raised to power x e raised to power minus lambda upon factorial x minus 1.

Now you can take lambda square common from here to make the power and this factorial same. So, that it will be converge to e raised to power lambda. So, this will be you can say lambda square lambda square into e raise to power minus lambda into e raise to power lambda will come and plus it will be it will be lambda into e raised to power minus lambda into e raised to power lambda will come. So, this will be nothing, but lambda square plus lambda.

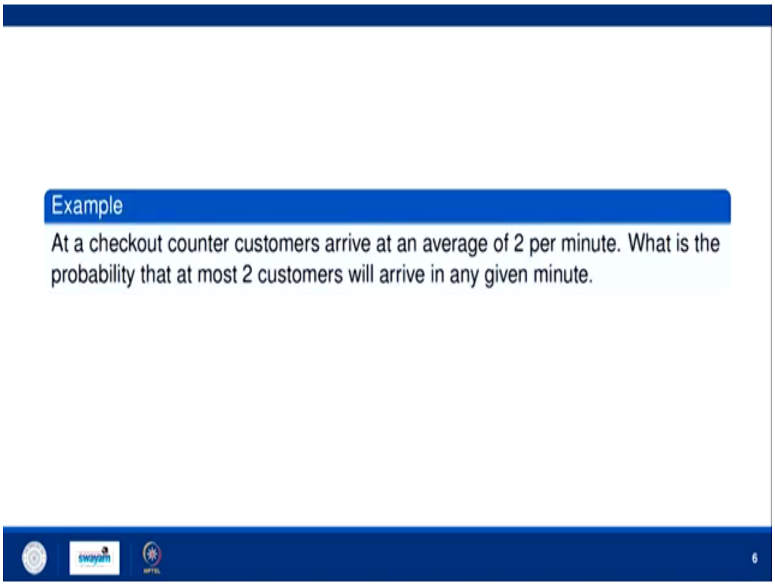
(Refer Slide Time: 18:36)



Now, you can find out variance using this expression. So, finally, variance will be lambda square plus lambda minus lambda whole square; so it is lambda. So, we can say that mean and variance of a random variable following Poisson distribution are same both are lambda.

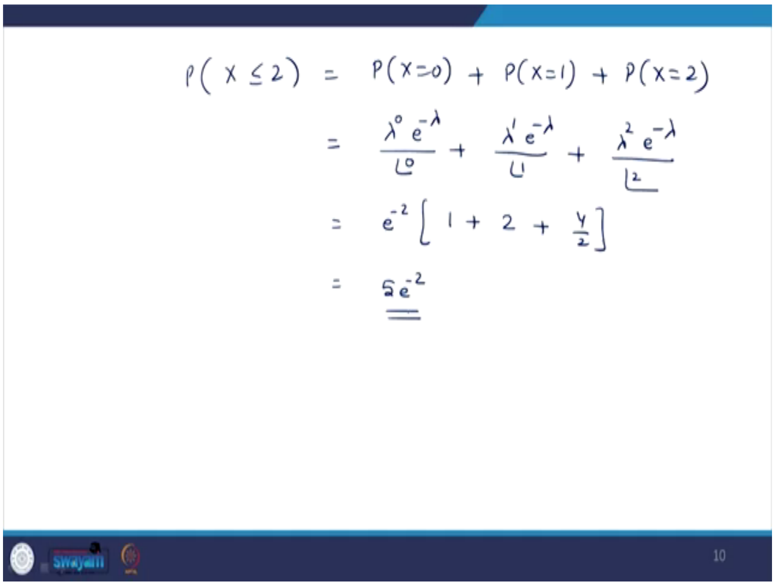
Now for which type of problems it is used; Poisson distribution serves as a model for various experiments and phenomena associated with waiting of an occurrence. In Quine theory we basically use such models; for instance the number of customers arriving a shop in a given time interval can be modeled using Poisson distribution.

(Refer Slide Time: 19:12)



So, let us calculate let us quickly do this example. Now at a check out counter customers arrive at an average of 2 per minute; so lambda is 2 ok. Then what is the probability that at most 2 customer will arrive in any given minute.

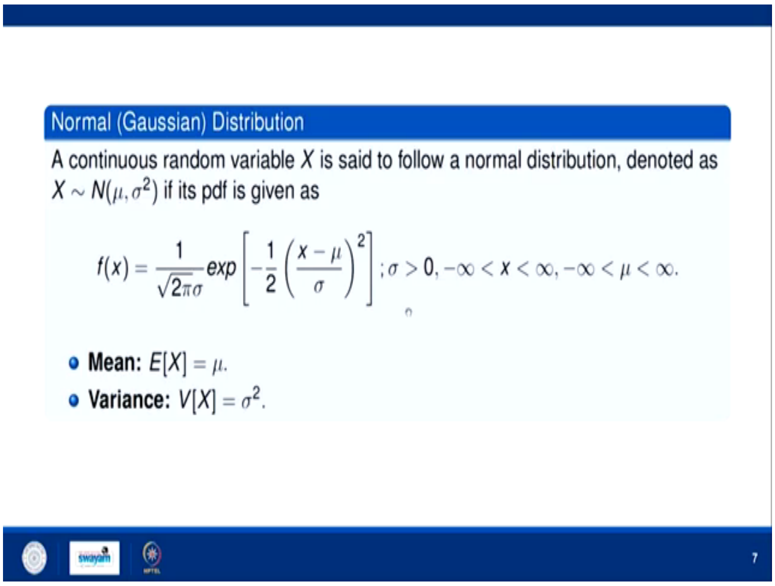
(Refer Slide Time: 19:36)



So, at most 2 customer means; P is less than P x is less than equal to 2 so; that means, P x equal to 0, P x equal to 1, plus P x equal to 2. So, we know that for Poisson distribution it is lambda raised to power x. So, x is 0 into e raised to into e raised to power minus lambda upon factorial x, plus lambda raised to power 1 e raised to power minus lambda upon factorial 1, plus lambda square e raised to power minus lambda upon factorial 2.

So, this is now lambda is 2 as given in the problem lambda is 2; so lambda is 2. So, it is e raised to power minus 2 will come here so it is 1 plus lambda is 2 plus it is 4 upon 2. So, this is nothing, but 5 e raised to power minus 2. So, that is the final answer that that is the final answer for this problem.

(Refer Slide Time: 20:38)

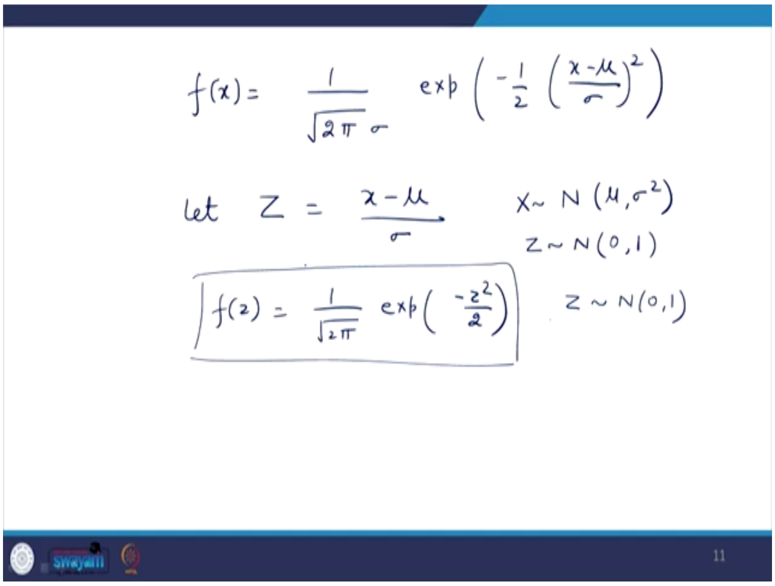


Now, let us come to Gaussian distribution which is a important distribution in terms of machine learning ok. Because you must have an idea for discrete case also that how we can find out mean, variance, etcetera for this discrete distribution then only I can come to normal distribution or Gaussian distribution which is a important aspect for machine learning.

So, if we follow what is the distribution function for this? So, let us see; a continuous random variable X is said to follow a normal distribution denoted as N mu sigma square; here mu is mean and sigma square is variance. If it is p d f is given by this expression; where sigma is greater than 0 and X is between minus infinity to plus infinity and mu is of course, from minus infinity to plus infinity; so it is mean is mu and variance is sigma square.

Basically here they are basically two parameters for this distribution. But they are not only the parameters here this mu is nothing, but mean and this sigma this is sigma is nothing, but standard deviation of this distribution. So, how can we say how can we prove it.

(Refer Slide Time: 21:58)



So, first for simplification let us see here the distribution is given by 1 upon under root 1 upon 2 under root the sigma ok. This e raised to power exponential raised to power, minus 1 upon 2, x minus mu upon sigma whole square. Now, let us let us take a small transformation let us take z which is x minus mu upon sigma ok.

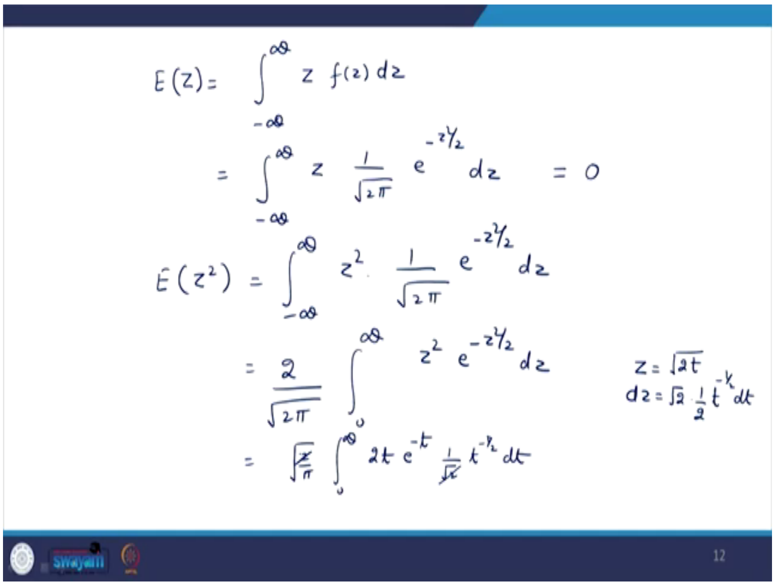
Now of course, if and if this is mu and sigma square if this N is mu and sigma square basically if this X is; so in terms of Z what will be N?. So, when x is when this is mu; so this is 0. And when you if you find the variance of this; so variance of Z will be 1 by sigma square variance of X and variance of X is sigma square. So, sigma square sigma square will cancel out; so it is 1. So, this is called standard normal variate basically we have transformed this into 0, 1.

Now so what will be f z? f z will be 1 upon under root 2 pi e raised to power minus z square upon 2. So, because of this transformation sigma will cancel out. So, now, now this is a very simple expression to see the Gaussian distribution and which is a which is a normal variate with mean 0 standard deviation 1.

Now, let us try to prove that this mean is 0 and standard deviation is 1; of course, if mean of this is 0 then mean of then if this is 0 then mean of this for this distribution will be mu. And if it is variance is 1 then variance of x will be sigma square that we can simply obtain by using the property of variance or expectation of x.

So, now, the only thing we have to show that the mean of this distribution is 0 and the variance is 1

(Refer Slide Time: 24:02)

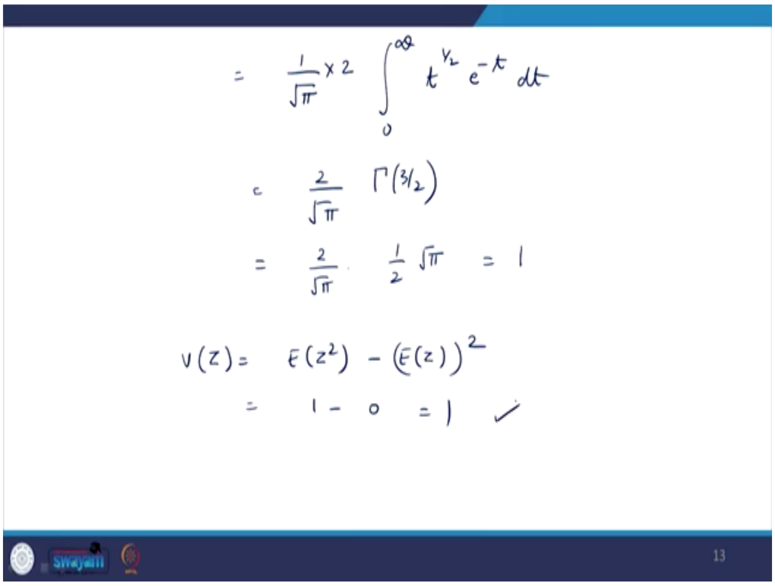


So, expectation of x we have to find. So, or z we have to find basically expectation of z here which is minus infinity to plus infinity this is z into f z, d z. This will be minus infinity to plus infinity z 1 upon under root 2 pi e raised to power minus z square upon 2 d z ok. It is an odd function and for odd function it is; obviously, 0. So, that is clearly one is one step answer you can obtain that it is 0 ok.

Now you have to compute for variance you have to compute expectation of z square that is integral this to this z square 1 upon under root 2 pi e raised to power minus z square upon 2 d z. And now it is an even function for this it will be two times 1 upon under root 2 pi 0 to infinity it is z square e raised to power minus z square upon 2 d z.

Now, you can take z as under root 2 times t suppose ok. So, d z will be what; under root 2 it is 1 by 2 t raised to power minus half into d t. So, that will be under root 2 upon pi integral 0 to infinity z square is 2 t, e raised to power now it is nothing, but minus t ok. And d z is what? d z is 1 upon under root 2 1 upon under root t raise to power minus half into d t. So, this will cancel with this two.

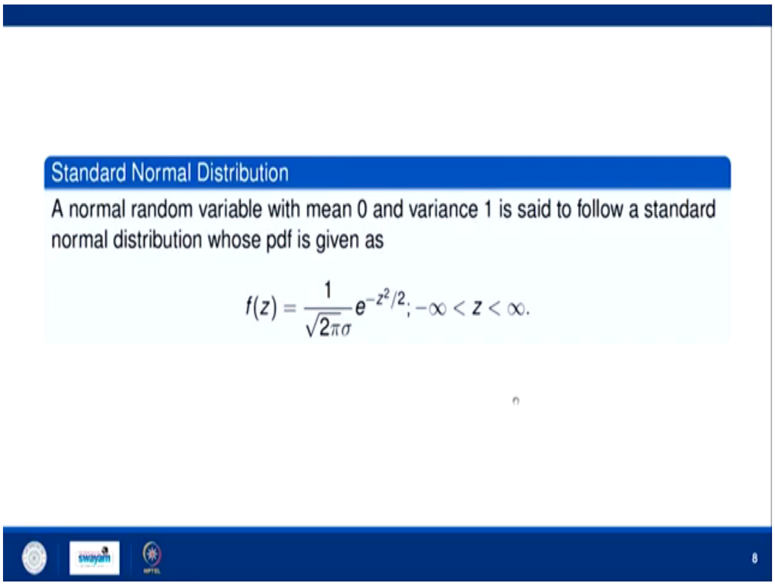
(Refer Slide Time: 25:55)



So, finally, we are having here is equal to 1 upon under root pi ok; 1 upon root pi will come from here 2 also will go out. So, that will be into 2 integral 0 to infinity that will be t, t raised to power half e raised to power minus t finally.

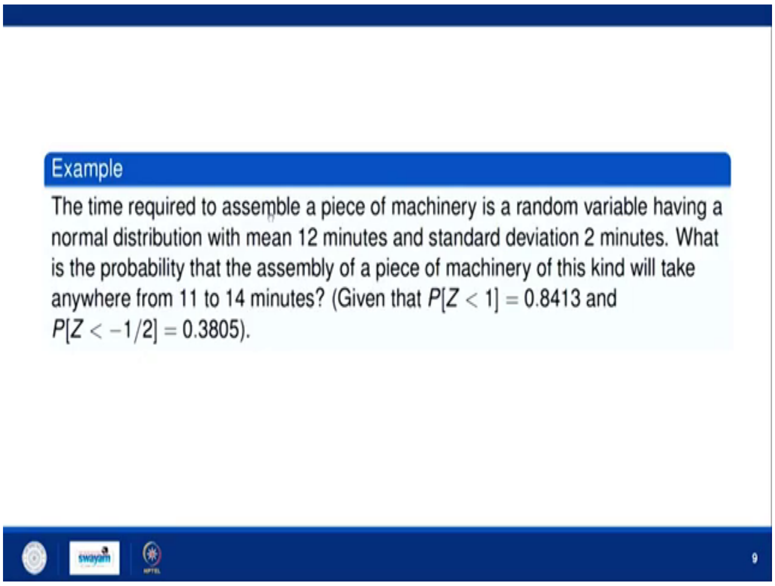
And this is nothing, but gamma 3 by 2 and that is further equal to 2 upon under root pi and this is gamma 3 by 2 is 1 by 2 under root pi. So, this is this is 1. So, variance of x is what expectation of x square minus expectation of x whole square so that will be 1 minus 0 that is 1; I mean for z this is z here ok.

(Refer Slide Time: 27:02)



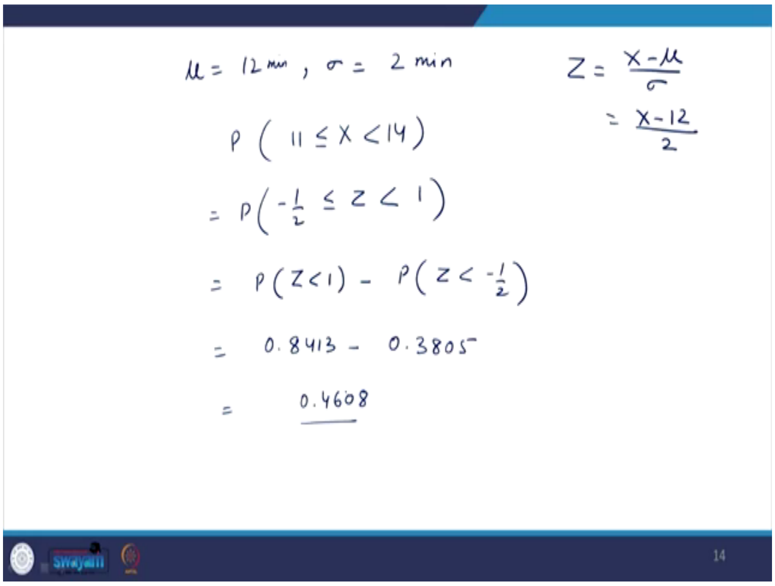
So, in this way we can say that for this normal for this distribution for this random variable which is following normal distribution for this mean is mu and variance is sigma square ok.

(Refer Slide Time: 27:17)



So, now let us come to one problem based on this simple problem based on this. The time required to assemble a piece of machinery is a random variable having a normal distribution with mean 12 minutes; that is mu equal to 12 and the standard deviation is 2 minutes; that means, variance is 4.

(Refer Slide Time: 27:37)



So, what is given to us; it is given to us that mu is 12 and sigma is sigma is 2 2 minutes this is also minutes. What is the probability that the assembly of piece of machinery of this kind will take anywhere from 11 to 14 minutes? So, what we have to find? We have to find that probability of x when x is between 14 and 11 ok.

So, this is nothing, but so for standard normal variate z is x minus mu upon sigma which is x minus mu is 12 upon sigma is 2. So, when you substitute x as 11. So, it is minus 1 by 2 less than equal to z less than when you when you put 14; so it is 1. And this is nothing, but P when x is less than 1 minus P when it is z

So, it is z so z less than 1 and it is z less than minus half which is given to us. It is given to us that this value is this and this value is this; so this is simply difference of these two. So, you can find out a difference; so it is 0.8413 minus it is 0.30, 0.3805. So, this difference is 0.4608; so this you can though this is the final answer for this.

So, sometimes this is simply for illustration that. If problem of this type is given to you that then how you can find out how you can convert that x into standard normal variate. And then how you can find out the respective probability based on that. So, this distribution which is a last distribution here Gaussian distribution or we sometimes called as RBF Kernel also in terms of machine learning.

So, in this lecture we have seen that if you are having discrete distributions, or continuous distribution, and we want to find out expectation or variance. And if you want to solve some problem based on simple problem based on that then how we can do.

In the next lecture we will see something about joint distribution functions their properties ok and of course; covariance matrix and other properties.

Thank you.