Applied Statistics and Econometrics

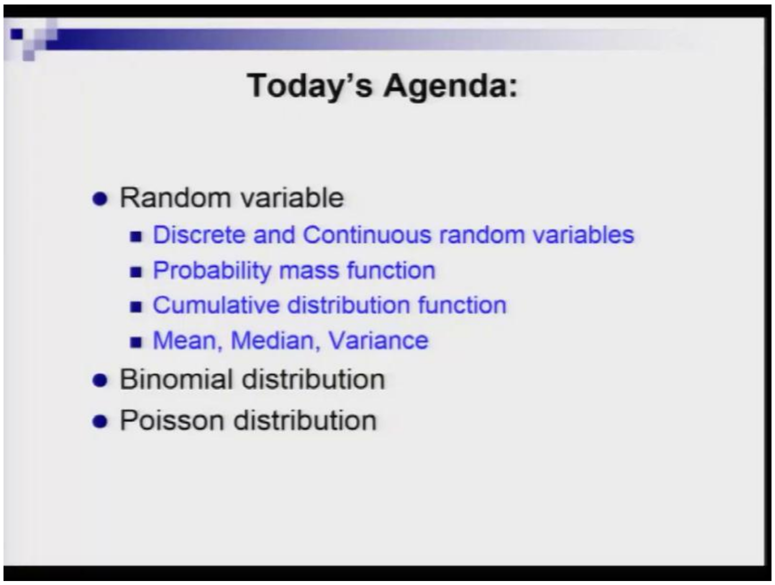
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur

Lecture 05

Discrete Random Variables and Probability Distributions

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. Today we are going to start our discussion on random variables. And before we go to the theoretical discussion let us have a look at the today's agenda items as usual.

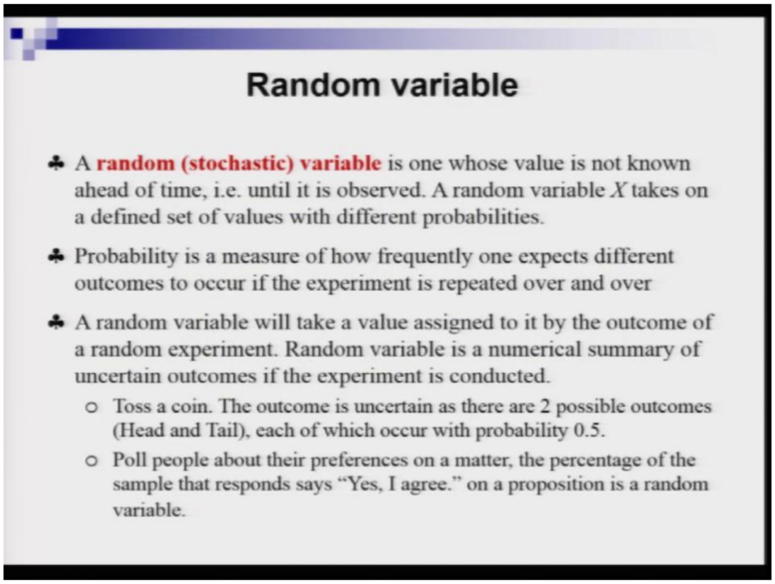
(Refer Slide Time: 00:31)



So today we are going to start with the definition of random variable. What is a random variable we will first discuss. Then we will discuss two types of random variables, discrete and continuous random variables. And then we are going to discuss two important concepts which are related to random variables namely probability mass function and cumulative distribution function.

Then we are going to talk about the summary measures of random variables and the probability functions, which are mean, median and variance. And finally we are going to show you two examples of discrete random variables which are mostly used in applied economic and social science research and they are Binomial distribution and Poisson distribution,

(Refer Slide Time: 01:24)



So let us now have a look at the random variable concept. So we have gone through the concept of a variable at the beginning of the course. But note that, then we were in a deterministic world. Everything was very certain. Now I am going to introduce you to the world of uncertainty where there is uncertainty around outcome of a variable or the values of a variable that is being materialized.

So a random variable, sometimes it is also called a stochastic variable, is one whose values are not known ahead of time, until they are observed. So random variable is denoted by the capital letters, with italics sometimes, sometimes, the common notations are X, Y and Z. They denote random variables. And this type of random variable takes a defined set of values with different probabilities.

So what do you I mean by that? In a layman's language if X takes different values say, X1, X2 to Xn then each of these values has an associated probability of being observed or being materialized. Now at this stage, it is not a bad idea to stop for a moment and revise the concept of probability because probability plays a big role here in the case of random variable.

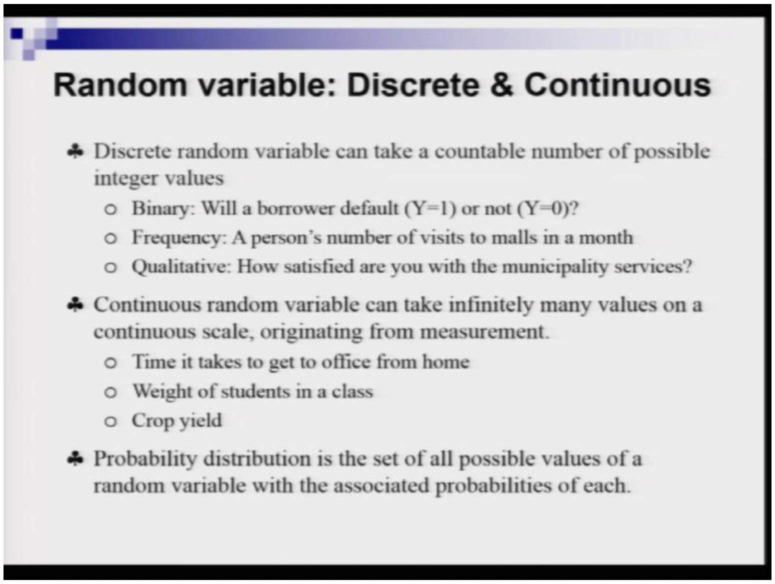
Now I assume that you have done probability in your school level mathematics but just, I would like to stop here for a couple of seconds and refresh your memory. So probability is a measure of how frequently one expects different outcomes to occur if the experiment is repeated over and over. So that is basically the concept of probability in a frequentist approach. So a random variable will take a value assigned to it by the outcome of a random experiment. So random variable is a numerical summary of uncertain outcomes if the experiment is conducted in practice.

So I will now give you two different examples. So think about a cricket match or a football match. And the match starts with coin tossing, right? So if you toss a coin then there are two different possible outcomes, head or tail. But note that these outcomes are uncertain in nature. A priori you do not know which one is going to happen. Each one of these possible outcomes have probability of 0.5 to occur.

Now let me move on to the second example. Suppose you want to conduct a survey and you want to know people's preferences or public opinion on a particular matter. And then you put a statement in front of the respondent. And then you ask, whether you agree or not with this statement. And the person either say, yes I agree or the person may say, no I disagree. So given this data set what is the percentage of sample or the number of, what is the percentage of respondents who says that, yes, I agree to this proposition or statement, that is basically a random variable because this percentage number is going to change from one sample to the other.

So probably from these two examples you can see that the random variables can take different sort of values. Sometimes it takes discrete values like the first coin tossing example where you have two outcomes and they are countable numbers. But in the second case when I talked about the percentage of sample saying yes then, percentages are basically on a continuous real line. So it can take any particular value. So, in the first case we have a discrete random variable. And in the second case, the case of public opinion sample we have a continuous random variable. So these are the concepts of random variables through examples. Now let us move on to the discrete random variable case in much more detail.

(Refer Slide Time: 06:12)



So now we are going to start our discussion on different types of discrete random variables based on the nature of the data that is being captured by a discrete random variable. So we will first start with the binary case where the discrete random variable can take only two values. So if you remember the coin tossing example it was of binary case because the possible outcomes were only two. So either it is a head or it is a tail if you toss a coin.

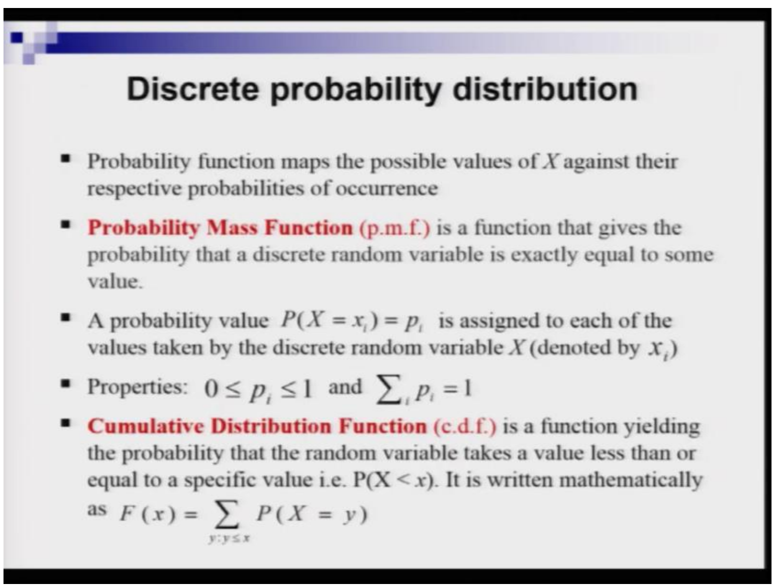
Now another example could be like the following, that suppose a person has taken loan from a bank. And now the bank manager is interested to know whether the borrower will default or not. So if he defaults then the random variable Y takes value 1 and if he does not then Y takes 0 value. So this is also binary discrete random variable. There could be a frequency type discrete random variable. So example could be a person's number of visits to malls in a month. And there could be second example like the number of students attending a particular class.

Now the last type of discrete random variables is of qualitative in nature. So one example could be, suppose you ask a group of respondents in a survey how satisfied are you with the municipality services. So it is a public opinion poll. And then, the responses could be of different categories like very happy, somewhat happy, not happy at all. And these are the qualitative discrete random variables.

Now briefly I will now introduce continuous random variable. And this is in stark opposition of the discrete random variable case. So here one can define a continuous random variable is one which takes infinitely many values on a continuous scale. And these values are originating from measurement. There could be examples, actually a lot of examples but I am showing only three here. So suppose we are interested to measure time that it takes to get to office from home. So, of course that could be a continuous variable. Then weight of students in a class or crop yield from a village or from a district, from a bunch of farmers.

So these are the continuous random variables, examples of continuous random variables. Now let me introduce this important related concept of probability distribution. So we have spoken before that the random variables are actually associated with probabilities. Different values of random variables are associated with different probabilities. So for each number there is a probability of occurrence. Now given that knowledge, what is the probability distribution? So it is the set of all possible values of a random variable with the associated probabilities for each value.

(Refer Slide Time: 9:13)



Now we are going to talk about discrete probability distribution because in this lecture we are going to focus on discrete random variable only. So again the important concept of probability function exists. And what is it? So it maps the possible values of Xs against their respective probabilities of occurrence. Now an important concept, probability mass function can be defined as a function that gives the probability that a discrete random variable is exactly equal to some value. So in mathematical terms, let us express this concept now.

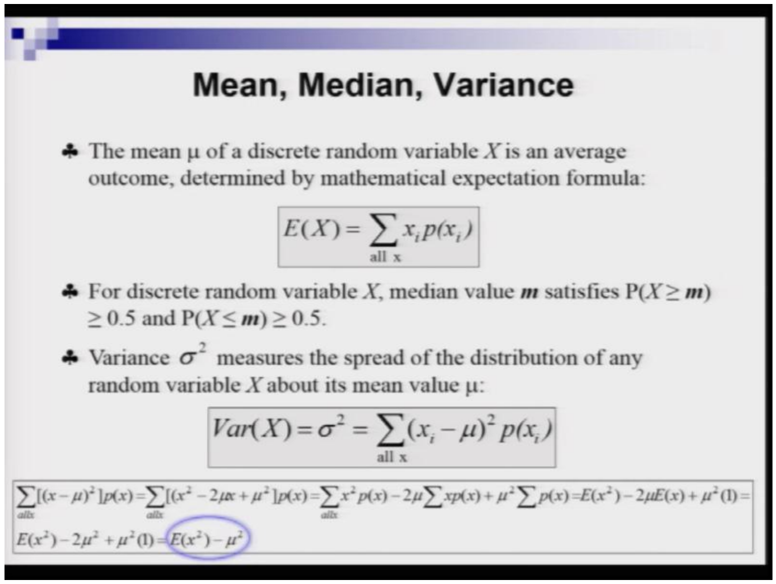
So suppose we denote probability value by pi. And pi can be elaborated as probability that capital X random variable takes value small xi. And then, that is assigned to each of the values taken by the discrete random variable X. So suppose, the values will be small x1, small x2 to small xn. And, all these n values, possible values are denoted by xi symbol or abbreviation or nomenclature, notation, whatever you want to call it.

Now note that when we are talking about this probability mass function and discrete probability distributions, there are two properties. So, first of all, each probability value pi is a bounded number. And there is a lower bound of 0 and an upper border of 1. So probability of an event cannot be higher than 1. And it cannot be negative also. And if you sum these probability values pis over the all possible values of the discrete random variable X then the sum will be equal to 1.

Now we will move on to the important concept of cumulative distribution function or CDF. What is it? It is a function giving us the probability that the random variable takes a value less than or equal to a specific value. So in mathematical terms we can write that we are interested to know what is the probability that capital X will be less than a particular value small x of the random variable. It is, written mathematically as capital F of x equal to summation of probability values p capital X equal to small y, where small y is basically less than or equal to of, value of the random variable which is small x.

Now this concept of cumulative distribution function will be more clear through an example that we are going to show later in this lecture. Now we are going to talk about mean, median and variance in the context of a random variable. We have seen how to compute these summary statistics from the deterministic variables. Now we are going to see how they can be computed in the case of random variables.

(Refer Slide Time: 12:33)



So the mean mu of a discrete random variable X is an average outcome determined by the mathematical expectation formula. Now note that mathematical expectation formula is given in the box below where I have written expectation of X equals to sum of products of xi and p of xi for all possible xs.

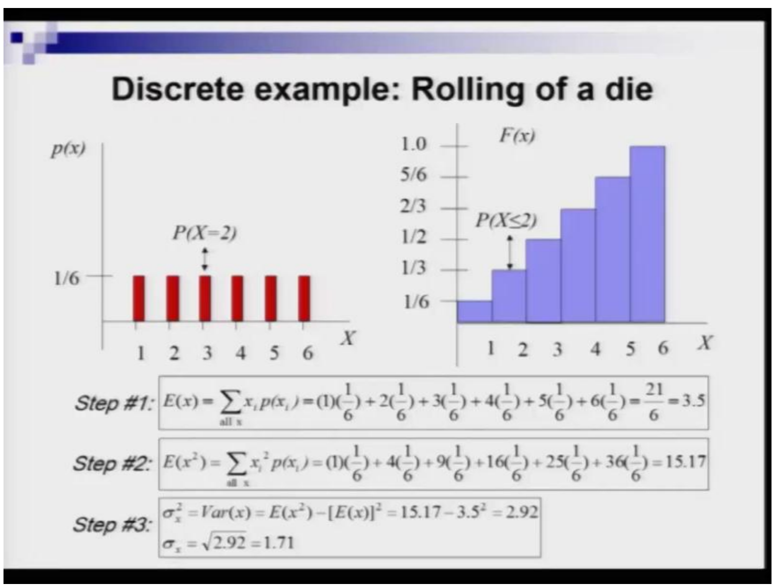
Now if we look at the median now, so then in that case for a discrete random variable X, the median value small m will satisfy two conditions. So first probability of X greater than equal to small m will be greater than or equal to 0.5, and the probability of X less than or equal to small m will also be greater than or equal to 0.5 So actually, the, these two probability measures will take value 0.5 only at the equilibrium or at the median value because that is what median does. It splits your data set or your frequency distribution in equal halves, or two equal parts.

Now we will move to the concept of variance. So variance sigma square measures the spread of the distribution of any random variable X about its mean value mu. So in mathematical terms, I am showing you the expression in the box given below. So here the sigma square is expressed as summation over small xi minus mu whole square times probability value of individual values xi and the sum should range over all possible x, say.

So what actually you have to do in a nutshell? You have to first compute the mean. Then you have to subtract that mean from the individual values of the discrete random variable. And these values are given by xi. So when the values in deviation form are obtained then can you take square of these numbers and then express them in all positive numbers. Then you multiply these numbers with the corresponding probability values that are observed for the individual values xi. And then you sum over all possible x values. Then you get the variance.

Now apparently it may look a bit complicated because you need to first compute mean and then take difference and all. But note that this particular expression can be simplified further to come with a simpler expression which is easy to remember, and for practical purposes it is also handy to compute the variance. So I am showing you the steps of calculation or derivation in the bottom box here that is the last box in the slide. And, you see I have put an oval shape around the final result. And the final result says that the variance is equal to expectation of x square minus the square of the mean which is mu. So if you follow this particular formula the variance computation will be easy.

(Refer Slide Time: 15:56)



Now let us look at these concepts of probability mass function, cumulative distribution function, mean and variance of discrete random variables through an example. So now we are going to talk about the computation of various features like PMF, CDF, mean and variance related to the discrete random variables through the example of rolling a die. So rolling a die is an experiment. Of course the outcome is uncertain because when you roll a die then you do not know which outcome you are going to observe.

So there is this set of possible outcomes which are basically 1, 2, 3 till 6. There are six possible outcomes. But we do not know a priori which one is going to appear when you roll a die. Now, of course, these different values 1 to 6 have got equal probabilities associated with them. Each one sixth, if the die is a fair one or it is an unbiased one. So if we take this assumption that the die that we are going to roll is an unbiased or fair one then how we can compute the statistical quantities? That we are going to see the next.

So first let us concentrate on the left hand side panel of the slide. And here you see I have plotted these six values of the discrete random variable X. And that is along the x axis. Then along the y axis I am plotting the probability of this discrete random variable X taking a particular value. That is denoted by small p x. And you note that these outcomes 1 to 6, these six possible outcomes are equi-probable. And the probability is 1 over 6 for all potential or possible values. So basically each bar in red color is actually showing you probability that X will be either 1, 2 or 6. So this is basically the PMF diagram.

Now let us concentrate on the diagram that you see on the right hand side of the slide. Now here you see along the x-axis I am plotting the same X but along the y axis now I am plotting the CDF values. So as I told you earlier, that the highest value of CDF could be 1. And that is basically the upper limit of F of x. Now note that here the diagram is a bit different from the previous one. Now here you see that there are different bars in blue colors, all in blue. And they actually have started from the number 0. And the first bar from 0 to 1 actually has a height of 1 over 6.

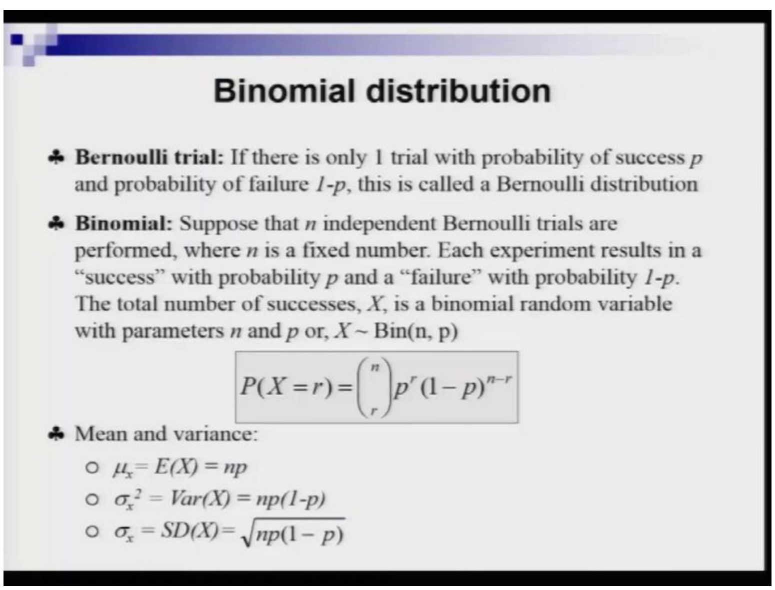
What does that mean? That means that, that the random variable X will take a value less than or equal to 1, that is 1 over 6. Then the second bar in the diagram that, all I also an arrow denoting what does that mean? So that height here is basically 1 over 3. Now from where did I get this number 1 over 3? So that, for that you have to go to that probability expression. So here I am saying that the height is actually showing you the probability of X less than equal to 2. So when I say that probability capital X less than or equal to 2 then that means that I am talking about sum of two different probabilities.

So one is basically the probability of observing X equal to 1 or probability of observing X equal to 2. So you need to sum them. So you need to sum them and then you get 2 over 6 because each probability quantities are coming from the left hand side panel and they are equal, 1 over 6. So if you add 1 over 6 two times then basically you get 2 over 6. And that is basically one third. And you see the height of the second bar in blue color actually gives me the same number 1 by 3. Similarly, the other bars can be explained. And you see the, the last bar that is basically for the interval 5 and 6 that basically gives you the highest value possible and that is 1.

We need to compute the mean. So the mean is basically mathematical expectation of the random variable X and it is given by the symbol E of x. And I have already shown you the formula. So if I apply the formula in this rolling of a die example then you see I am multiplying each possible values of the random experiment with their probability of occurrence. And, I get a number 3.5. That is basically my value of the mean.

Now to get the variance I will use the simplified formula that I have shown and explained in the last slide. So for that I have to compute expectation of x square. So here I am showing you how to compute that. So it is very simple. You have to just, take the square value of the possible values of the random experiment and then you need to multiply these squared terms with their probabilities and you get 15.17. And in the last step you apply that simplified formula of variance and you get value of 2.92. And if you take a positive square root of that you get 1.71. And that is the standard deviation of the discrete random variable X.

(Refer Slide Time: 22:21)



Now with this brief introduction on discrete random variables, we will now look at two different discrete random variables and associated distributions which are extremely useful and popular in applied social research. So first we will start with Binomial distribution. But before we discuss Binomial distribution we have to go through a very important concept which is basically the foundation behind Binomial distribution. And that is the concept of Bernoulli trial.

So if there is only one experiment or trial with a probability of success p and probability of failure 1 minus p, that is called Bernoulli distribution. So basically think of an experiment which can take place only one time. And the experiment has, experiment is random. It means the outcome is uncertain. And there could be only two outcomes, success or failure. And success has a probability of p. So, of course the probability of failure will be 1 minus p because sum of these two probabilities of success and failure will be equal to 1 only. So in this context this distribution is called a Bernoulli distribution.

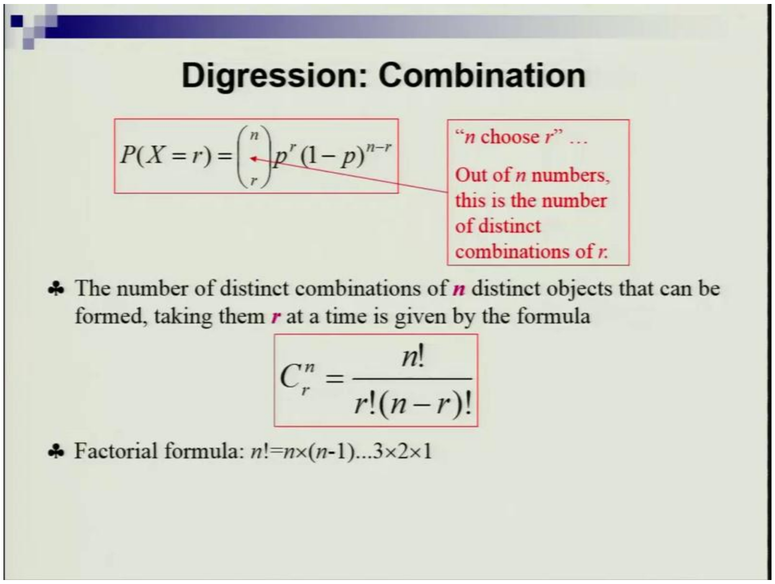
So when you have n number of independent Bernoulli experiments or trials then actually you get the case of a Binomial distribution. And now we are going to have a look the PMF of the Binomial distribution. And from there we will see how one can compute the mean and the variance. So n is basically the fixed number of independent Bernoulli trials. And as I said earlier each experiment results in two outcomes, two possible outcomes, either success with probability p or a failure with probability 1 minus p.

So here we define a discrete random variable X as the total number of successes in n trials. And that is called a binomial random variable with parameter small n and small p. Mathematically it is expressed shortly as capital X follows Bin n comma p. And in the box here we are showing you the PMF, the probability mass function for the Binomial distribution. And here I am showing you how to calculate the probability when you are interested to measure the discrete random variable X taking a particular value small r.

So that is basically n combination r times probability to the power r times 1 minus p to the power n minus r. So that is basically the probability mass function expression. So from this expression one can derive the mean and variance of Binomial distribution. So mean is the expected value which is E X and that is equal to small n times p. The sigma square is variance of X that is equal to small n times p times 1 minus p. And of course, the standard deviation is square root of variance.

So here I am going to skip the derivation of the mean and variance computations for the Binomial distribution with the hope that you can do it yourself because I have shown your earlier for that rolling a die example how to go about it. So I leave it to you. You should be able to find out the steps that will lead you to the expressions that I have shown here for mean and variance.

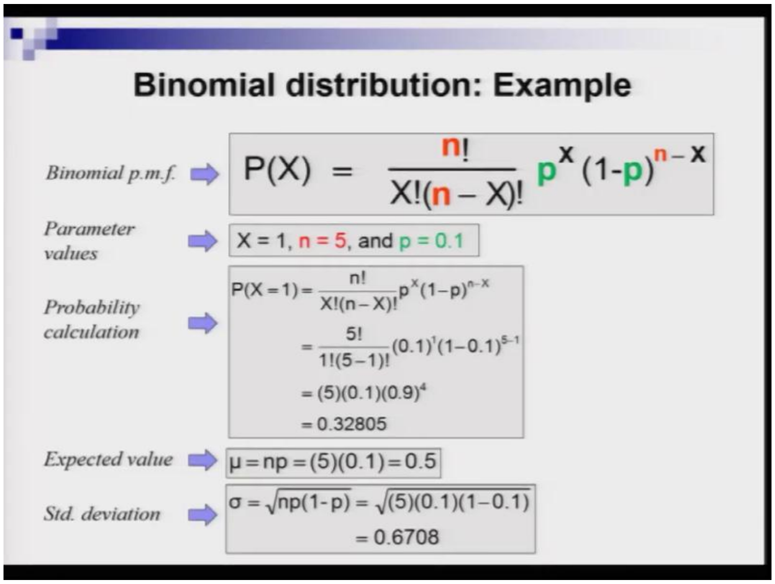
(Refer Slide Time: 25:51)



So before we look at an illustration of the binomial probability calculations let us take a digression for the moment. And let us look at the formula of the PDF in great detail because it involves some mathematical concepts known as combination and factorial. Those of you who do not know, this digression will help them to understand this PDF better. So note that here I have highlighted this n, r within the parentheses or braces. And that is to be read as n choose r. That is the mathematical way to read it. And what does it mean? It means that if I am given n numbers then out of these n numbers or elements this is the number of distinct combinations or r.

So the number of distinct combinations of n distinct objects that can be formed taking them r at a time is given by the formula n combination r which is given by n factorial divided by r factorial times n minus r factorial. Now what is a factorial? So the last formula in this slide is showing you the calculation formula for factorial. So n factorial can be computed as in n multiplied by n minus 1 and so on so forth till we hit number 1. So if we are talking about say, 3 factorial then basically we are talking about a value of 3 times 2 times 1. Next we will look into the binomial probability calculations through an example.

(Refer Slide Time: 27:43)

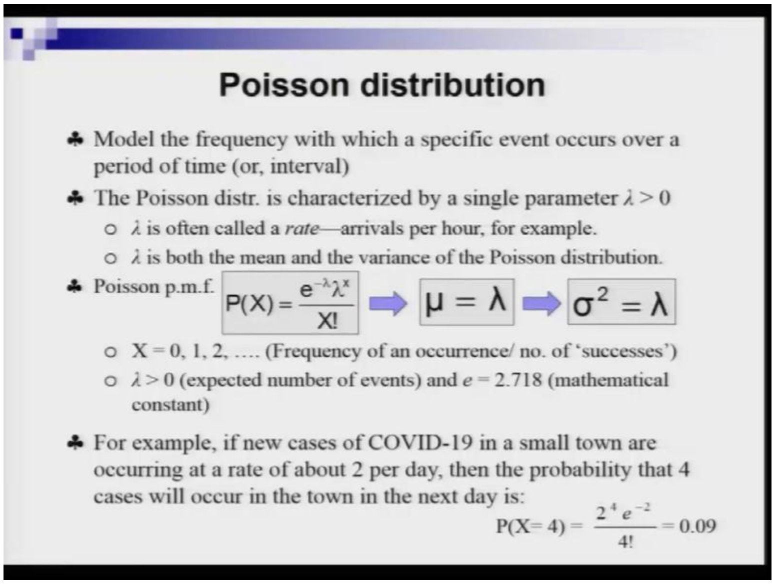


So now we are looking to look at the Binomial distribution again but this time through an example because the calculation of Binomial distribution may not look very easy compared to the computations that we have done earlier in the course. So an illustration will help. So let me show you the binomial PMF again in that first box. Now here you note that the parameter values, I have colored so that you can identify them easily and remember easily.

So n, the number of trials is colored red. And p, the probability of success is colored green. And let us start this illustration or example for some arbitrary values. So let us assume that capital X is equal to 1. So if you, link this to the previous slide then, the r value that I have shown here there, that r is basically 1. But note that, that is not a parameter. So we are interested to know what is the probability of only one success if we conduct n equal to 5 trials or experiments with the probability of success being 0.1.

So n and p parameter values are given. So with these parameters values now one can approach to the probability calculation. And that is what; I am showing you in the third box here. So probability X equal to 1 is that formula and you plug that parameter values in the formula and you get the probability score. Now we will focus on the expected value. And that is given by mu. And, if you remember the formula that was n times p. So again if we plug the values of these two parameters we get the expected value or mean of this discrete random variable X here. And similarly you plug back the values of n and p in that standard deviation formula to get the standard deviation of X. So these are simple calculations.

(Refer Slide Time: 30:24)



Now we will go on to the another important discrete random variable and distribution is called Poisson distribution. So Binomial and Poisson are somewhat related but we are not going to discuss the theoretical linkage between Binomial and Poisson. I am going to end today's lecture by giving a very simple introduction to Poisson distribution.

So where is Poisson distribution is applied? So we will start with a small discussion where Poisson distribution finds use. So it is used to model the frequency with which a specific event occurs over a period of time or interval. A Poisson distribution is characterized by a single parameter lambda. And note that here I have said that lambda is positive. That it is very important because Poisson distribution is undefined for negative or 0 value of lambda.

Sometimes this lambda is called a rate. And the examples could be arrivals per hour. And lambda is both mean and variance of the Poisson distribution. So that is a very, very special feature of Poisson distribution. So if you know the value of lambda, you know the mean and the variance. So with this let us look at the Poisson PMF or probability mass function so that I am showing here in that box. Probability of X equal to e to the power minus lambda times lambda to the power x divided by x factorial.

And from this PMF one can calculate the probabilities for different values of X. Now what could be the values of X? Here the X, discrete random variable will take, frequency or count type data. So one can say that X actually is the frequency of an occurrence or number of successes in some number of trials and all. And X can take values like 0, 1, 2 and it can go up to any integer. And note that positive lambda is expected number of events. And e is basically the mathematical constant. We all know about it and it takes value 2.718 roughly.

So from this Poisson PMF one can derive the expressions for mean and the variance. But note that they will be the same that is lambda. So that is theoretical result. I leave it to you. You can try yourself to see that whether you are able to reach these mathematical expressions by yourself for mu and sigma square. Now I will end discussion on Poisson distribution with a small example. And, I will show you how to compute the Poisson probability.

So let us take an example. So if new cases of this COVID-19 in a small town are occurring at a rate of about 2 per day then what is the probability that there will be 4 cases that will occur in the town in the very next day? And here you can see that lambda actually is taking value 2 and X actually has the value of 4. So if you plug these two values in the Poisson PMF then you get a value of 0.09. So I will end today's lecture here. In the next lecture we are going to start discussion on continuous random variable. Thank you.