Essentials of Data Science with R Software – 1
Professor. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture No. 34
Poisson distribution

Hello friends, welcome to the course of Essentials of Data Science with R Software – 1; in which we are trying to understand the basic concepts of probability theory and statistical inference. So, now you can recall that in the last lecture we had made a discussion on the binomial distribution. So, in this lecture we are going to introduce our new probability mass function, which is Poisson probability mass function or Poisson distribution. Similarly, this Poisson distribution also has the situations, where it can be used. λ

Have you observed a book which is published by a reputed publisher or a local publisher? You will see that if the book is from a reputed company, which has taken at most care in the printing of the book. The number of printing mistakes are very very low; possibly the number of mistakes are going to be large. And the difference you can realize that if you are trying to solve an unsolved exercise or problem and if your answer is not matching with the answer given in the book on the back side, then if you are doing it from a very good book, which is published by a reputed publisher; you will try to check yourself again and again, whether you are trying to make any mistake or not.

But, if if that is from some unprofessional publisher, who has not taken at most care; you will say simply that, the answer maybe wrong. Now, what is this? Can you observe this phenomena? What is really happening? Suppose we have a 500 pages book from both the publishers. Now, you try to count the number of mistakes. It might be possible that in the book from the reputed publisher, that there maybe 1 or 2 mistakes, in thousands of words spread over in the 500 pages. But, in that lower quality book, it is possible that every page may have 1 or 2 mistakes.

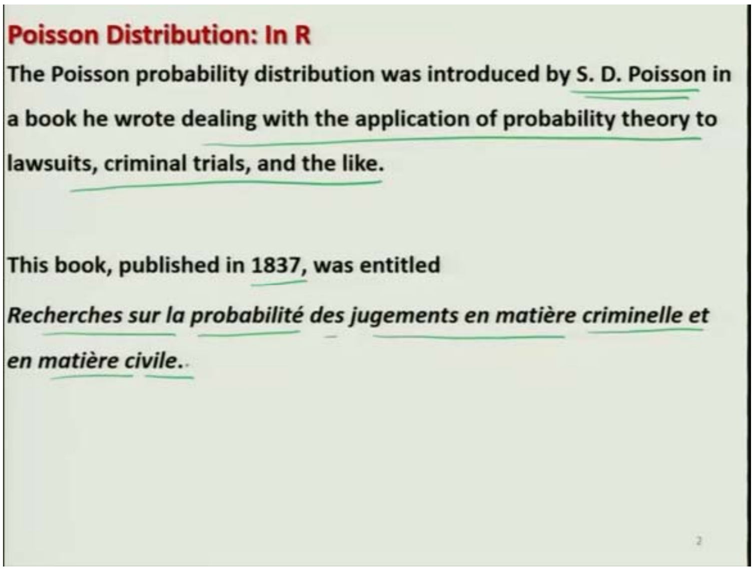
So, now what is happening of you try to compute the probability in both the cases; the probability is going to be simply that total number of mistakes, divided by the total number of words in the book. In the first case what will happen? The number of words are very high; and the number of mistakes are very low. So, the probability will become extremely small; but, in the second case the number of words are high. But, the number of mistakes are also high; so the probability will not be is very small. So, now under this type of situation if you try to employ the binomial distribution; it will not give you the correct results.

Similarly, if I take one more example suppose there is some radioactive experiment. And there is a very small hole in which some radioactive active material is placed and some electron, photon, neutron. Although I am not a conversion with the nuclear physics, but suppose they are emitting. There there will be millions, billions and trillions of such particle; but those particles which are crossing that small hole, they are very very less. So, the probability that a particle is crossing the hole that is going to be extremely small; and the number of particles are going to be very large. Same as the number of words are going to be very high; but the probability of occurrence is very low.

In all such cases, you can use the Poisson distribution, where the number of observations are very high; and the probability of occurrence is very low. Now, this is your job that in data science you have to; means see that where this situation arises. For example, if you are trying to make a shopping from some reputed websites. The mistakes in the transaction, means financial transaction or the refund of the money, they are very, very less. It may be possible that out of 1 million; there maybe 1 or 2 mistakes.

But, if you are trying to do the shopping from a website, which is not well designed; then means possibly the number of mistakes are going to be very large. So, in the first case, you can always use the Poisson distribution to approximate various types of probabilities. So, with this objective now let us try to begin the Poisson distribution.

(Refer Slide Time: 05:19)



Poisson Distribution: In R

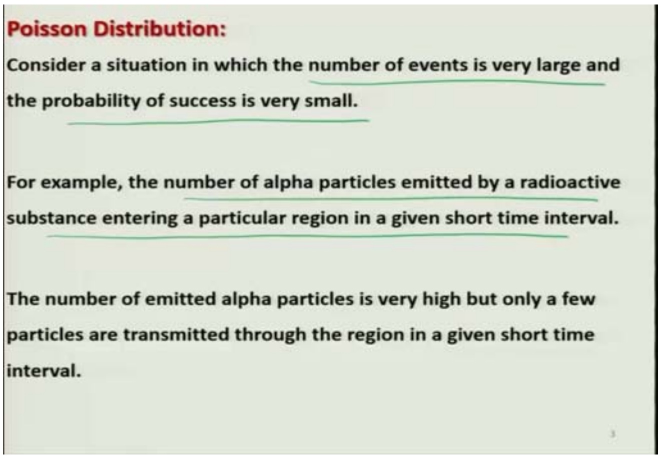
The Poisson probability distribution was introduced by S. D. Poisson in a book he wrote dealing with the application of probability theory to lawsuits, criminal trials, and the like.

This book, published in 1837, was entitled

Recherches sur la probabilité des jugements en matière criminelle et en matière civile.

So, now let us come to our slides, and you can see here that this Poisson probability distribution was introduced by S. D. Poisson. And it was introduced in a book that he wrote dealing with the application of probability theory to lawsuits, criminal trials and the like. The book was published in 1837 and the title of the book was that was in actually, means I believe in French. Some Recherches sur la probabilite des jugements en matiere criminelle et en matiere civile. I am sorry I do not know the French language; I am just trying to read it. So, please excuse me from the correct pronunciation.

(Refer Slide Time: 06:04)



Poisson Distribution:

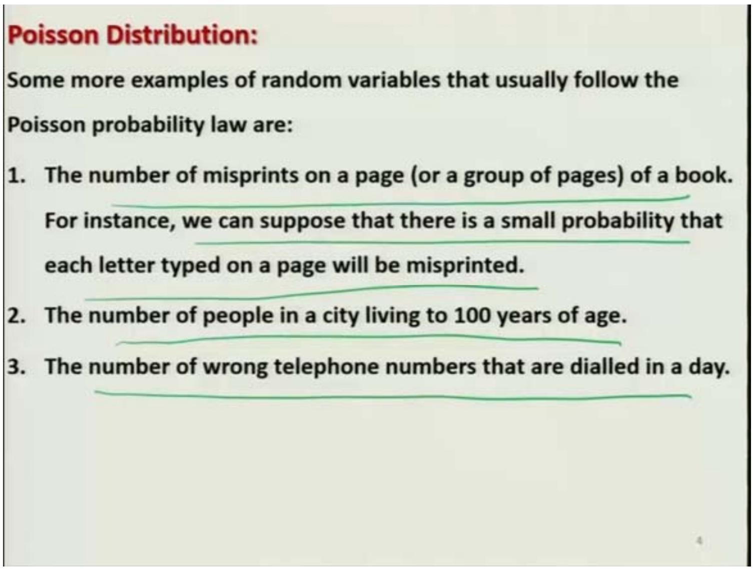
Consider a situation in which the number of events is very large and the probability of success is very small.

For example, the number of alpha particles emitted by a radioactive substance entering a particular region in a given short time interval.

The number of emitted alpha particles is very high but only a few particles are transmitted through the region in a given short time interval.

So, now let us try to consider a situation in which the number of events is very large, and the probability of success is very small. For example, the number of alpha particles emitted by a radioactive substance entering a particular region in a given short time interval. They are going to be very small, but this by the number of alpha particles emitting emitted will be very very large. So, that is what I was trying to explain you in the examples.

(Refer Slide Time: 06:34)



Poisson Distribution:

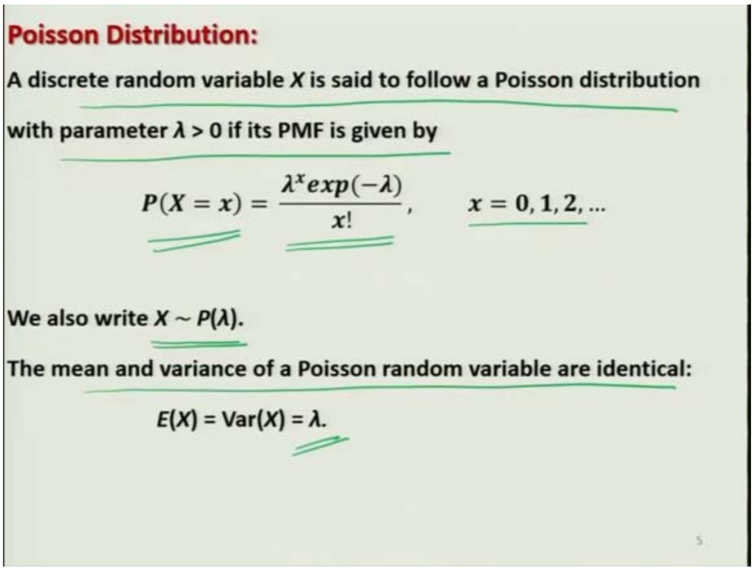
Some more examples of random variables that usually follow the Poisson probability law are:

1. The number of misprints on a page (or a group of pages) of a book. For instance, we can suppose that there is a small probability that each letter typed on a page will be misprinted.
2. The number of people in a city living to 100 years of age.
3. The number of wrong telephone numbers that are dialled in a day.

So, some more examples where this the random variable follow a Poisson distribution will be like this. That the number of misprints on a page or a group of pages of a book. For example, if we can suppose that there is a small probability that each letter typed on a page will be misprinted; that will that is going to be very small, if the book is from some reputed publisher. The number of people in a city living to 100 years of age, you know well; this number is going to be extremely small. And the probability of occurrence of such an event is going to be extremely low.

So, and similarly if you try to see now what is the total number of wrong telephone numbers that are dialed in a day. They are very very less, because you are simply trying to to press or click on the name of the person; and you do not have to remember the the 7 digit, 9 digit, 10 digit phone numbers. You simply have to just press and the call is there, but earlier before the cell phones, one has to dial the numbers each and every number manually. So, there so the chances of dialing wrong number was very very high earlier. But, now with the new this inventions and this developments; this probability has become very low. But, the number of calls in a day, they have become very large.

(Refer Slide Time: 08:03)



Poisson Distribution:

A discrete random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ if its PMF is given by

$$P(X = x) = \frac{\lambda^x exp(-\lambda)}{x!}, \qquad x = 0, 1, 2, \ldots$$

We also write $X \sim P(\lambda)$.

The mean and variance of a Poisson random variable are identical:

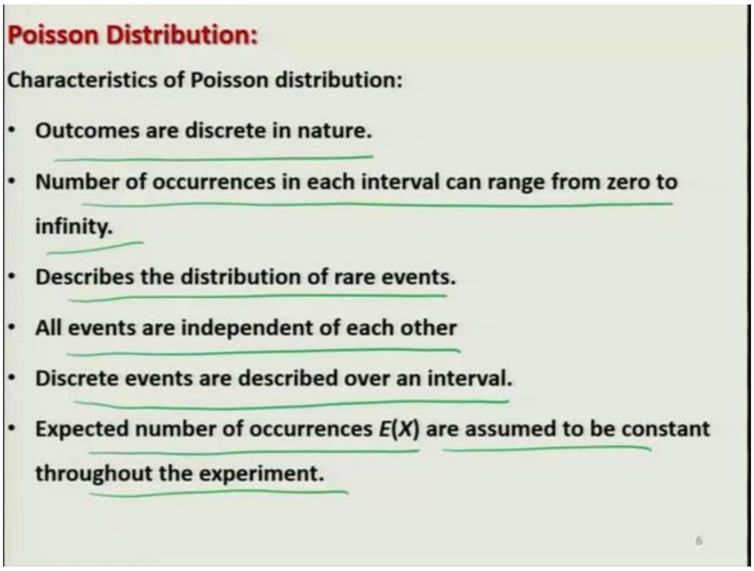
$E(X) = Var(X) = \lambda$.

So, in such cases you can think of using the Poisson distribution to compute different types of probabilities etc. So, a discrete random variable X is said to follow a Poisson distribution with parameter λ is greater than 0. If its probability mass function is given by $P(X = x) = \frac{\lambda^x exp(-\lambda)}{x!}, \ x = 0,1,2, \ldots$ So, in this case the parameter is λ; so we also write that x follows a Poisson distribution with parameter λ like this, X ~ P (λ).

One of the very peculiar characteristic of Poisson distribution is that the mean and variance of Poisson random variable are identical. Same as λ, mean is λ variance is λ. So, this also gives you a sort of some evidence based on which you can decide in data sciences, that which of the distribution is going to be fitted over there. For example, if mean and variance are coming out to be same; you can imagine that this is the case in the case of Poisson distribution, and if the value of mean is coming out to be greater than the value of variance. You can think about the that these things happen in the binomial distribution.

I am not saying that these things are going to happen in each and every one; or a binomial is the only one, which has such property. But, these are only the hints, just like when somebody goes to the doctor; doctor like likes to examine the mouth, eyes, tongue etc. etc. and based on that the doctor takes a call that what type of ailment the person can have.

(Refer Slide Time: 09:55)



Poisson Distribution:

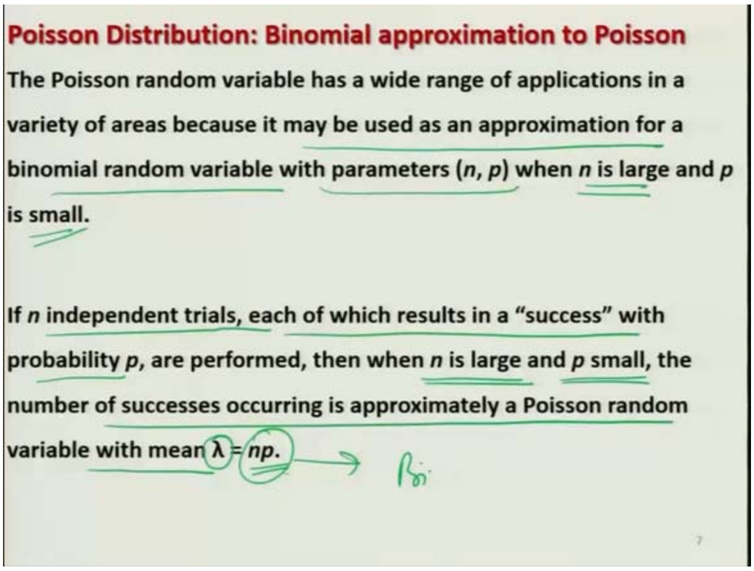
Characteristics of Poisson distribution:

- **Outcomes are discrete in nature.**
- **Number of occurrences in each interval can range from zero to infinity.**
- **Describes the distribution of rare events.**
- **All events are independent of each other**
- **Discrete events are described over an interval.**
- **Expected number of occurrences $E(X)$ are assumed to be constant throughout the experiment.**

So, similarly this Poisson distribution has some characteristics and these characteristics if you try to compare with the real life data process; which you are handling in the data science, possibly that may also help you in giving some sort of hint, that whether Poisson distribution can be used there or not. For example, the outcomes are discrete in nature; the number of occurrence in each interval can range from 0 to infinity. And and it describes the distribution of rare events, which are not so popularly happen; or their frequency is extremely low.

All the events are independent of each other; the discrete events are described over an interval. And expected number of occurrence that is E(X) are assumed to be the constant throughout the experiment. So, these types of says symptoms or properties of the or the characteristic of the Poisson distribution can help you.

(Refer Slide Time: 10:55)



Poisson Distribution: Binomial approximation to Poisson

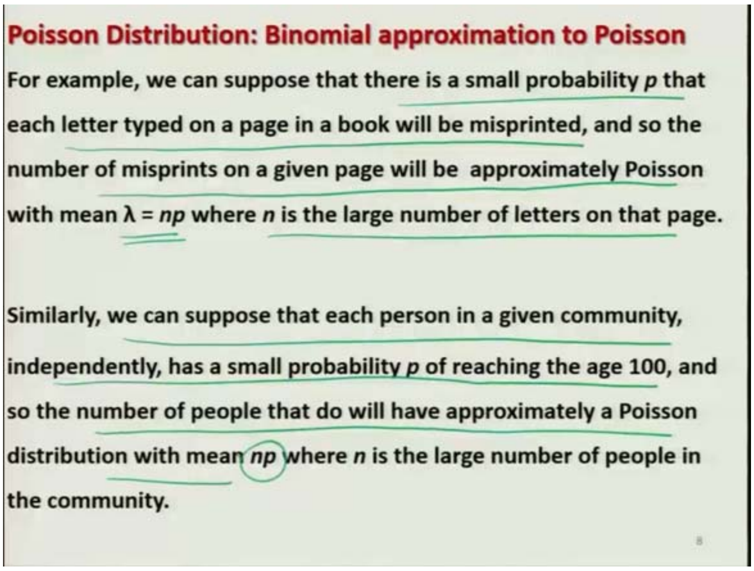
The Poisson random variable has a wide range of applications in a variety of areas because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small.

If n independent trials, each of which results in a "success" with probability p, are performed, then when n is large and p small, the number of successes occurring is approximately a Poisson random variable with mean $\lambda = np$. $\longrightarrow$ Bi

So, this Poisson this random variable has a wide range of applications in a variety of areas. Why? Because it may be used as an approximation for a binomial theorem variable; so binomial random variable has got a parameter n and p. And in the case of binomial distribution, if n is very large and p is very small; then possibly this distribution can be approximated by other Poisson distribution. So, in the case of Poisson distribution, we say that if n independent trials are there, and each of the trials results in a success with some probability say p.

And in case if the number of trial that is n is very large and the probability of occurrence of the event is very small; then the number of successes occurring is approximately a Poisson random variable with mean λ which is equal to n into p; and np they are coming from binomial distribution with parameters n and p.

(Refer Slide Time: 12:20)



Poisson Distribution: Binomial approximation to Poisson

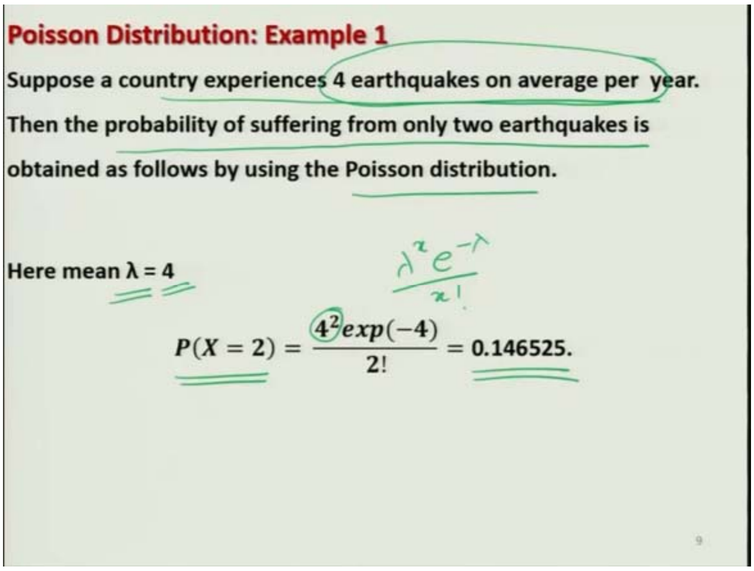
For example, we can suppose that there is a small probability *p* that each letter typed on a page in a book will be misprinted, and so the number of misprints on a given page will be approximately Poisson with mean λ = *np* where *n* is the large number of letters on that page.

Similarly, we can suppose that each person in a given community, independently, has a small probability *p* of reaching the age 100, and so the number of people that do will have approximately a Poisson distribution with mean *np* where *n* is the large number of people in the community.

For example, we can suppose that there is a small probability say p that each letter typed on a page in a book will be misprinted. And so the number of misprints on a given page will be approximately Poisson distribution, with mean λ is equal to np; where n is the large number of letters on that page. And similarly, if you try to take the other example that we considered earlier. It says that we can suppose that each person in a given community, independently, has a small probability p of reaching the age 100.

And so the number of people that will have approximately a Poisson distribution with mean np, where n is the large number of people in the community, which is very large. So, these types of statements you can actually make.

(Refer Slide Time: 13:21)



Poisson Distribution: Example 1

Suppose a country experiences 4 earthquakes on average per year.

Then the probability of suffering from only two earthquakes is obtained as follows by using the Poisson distribution.

Here mean λ = 4

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

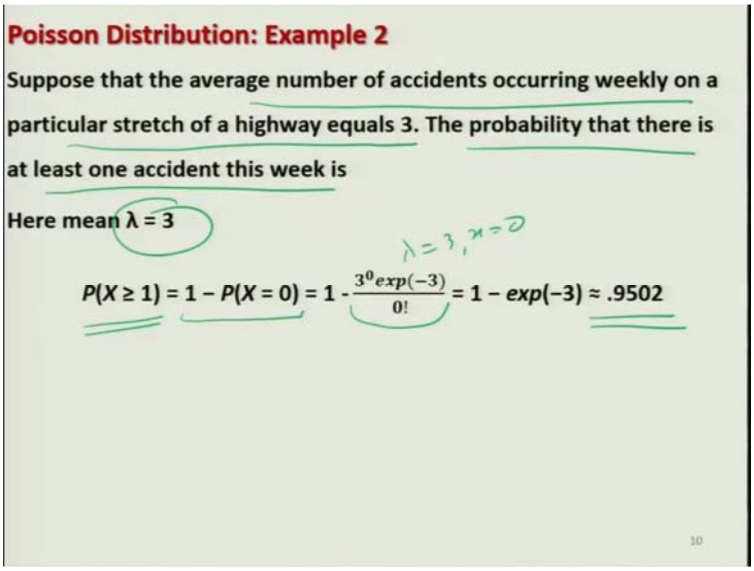
$$P(X = 2) = \frac{4^2 exp(-4)}{2!} = 0.146525.$$

9

Now, I try to take here some example and try to show you that how you can compute the different types of probabilities using the Poisson distribution. And the next lecture I will try to show you the similar things on the R software. So, suppose a country experiences four earthquakes on an average every year. Then the probability of suffering from only 2 earthquakes is to be obtained. So, you know that means the earthquake usually come for a very short time; and the total in a year is very very large. So, the probability of earthquake is very very low; so, we can assume that this is following a Poisson distribution.

And here since we are given the information that that on an average, there are 4 earthquakes per year. So, we know in the Poisson distribution λ is the mean; so λ is considered here as 4. So, now we want to find out the probability of exactly two earthquakes; so probability of X equal to 2. So, that will become here λ^2, λ^x $e^{-\lambda}$/ x!. So, you this will become here 4 square exponential of minus 4, divided by factorial of 2; and if you try to approximate it, this will come out to be 0.14. So, there is a only 14 percent approximate probability that there will be 2 earthquakes in the in that country.

(Refer Slide Time: 14:57)



Poisson Distribution: Example 2

Suppose that the average number of accidents occurring weekly on a particular stretch of a highway equals 3. The probability that there is at least one accident this week is

Here mean $\lambda = 3$

$\lambda = 3, x = 0$

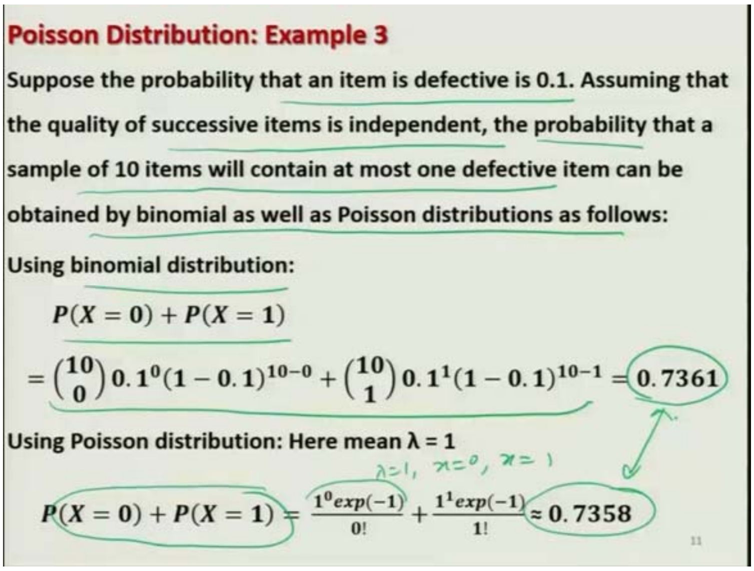
$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{3^0 exp(-3)}{0!} = 1 - exp(-3) \approx .9502$$

10

And now let me take one more example, suppose that the average number of accidents occurring weekly on a particular stretch of a highway is 3. So, that means you are trying to say that on an average there are 3 weekly accidents. And you know that the number of cars going on the highway are very very large. So, we we can assume that this process is possibly following a Poisson distribution; and in this case if we try to find out the probability that there is at least 1 accident this week is to be found. So, obviously in this case λ will become equal to 3, because this is the average value of the number of accidents in a week.

And what you want here at least 1 accident; so that will become here probability that X is greater than or equal to 1. So, this I can computed by 1 minus probability of X equal to 0; and simply I try to use λ equal to 3 and x equal to 0. And the probability mass function of Poisson distribution, I try to put it here; and then I try to simplify it. This will be very close to point 95 or there is a 95 percent probability.

(Refer Slide Time: 16:15)



Poisson Distribution: Example 3

Suppose the probability that an item is defective is 0.1. Assuming that the quality of successive items is independent, the probability that a sample of 10 items will contain at most one defective item can be obtained by binomial as well as Poisson distributions as follows:

Using binomial distribution:

$$P(X=0)+P(X=1)$$

$$=\binom{10}{0}0.1^{0}(1-0.1)^{10-0}+\binom{10}{1}0.1^{1}(1-0.1)^{10-1}=0.7361$$

Using Poisson distribution: Here mean λ = 1

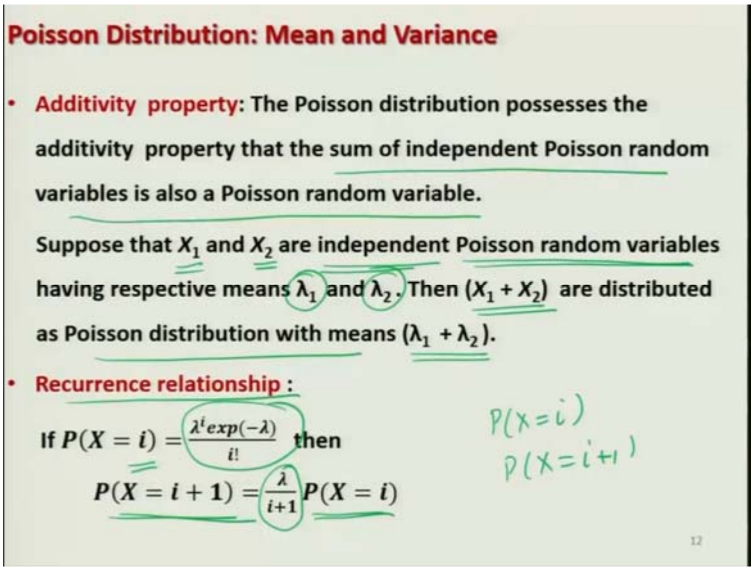
$\lambda=1,\ x=0,\ x=1$

$$P(X=0)+P(X=1)=\frac{1^{0}exp(-1)}{0!}+\frac{1^{1}exp(-1)}{1!}\approx 0.7358$$

Now, we try to consider one more example, where we try to find out the probabilities of the same event using the binomial distribution and the Poisson distribution; and we try to compare their probabilities. So, suppose the probability that an item is defective is 0.1, and assuming that the quality of successive items is independent. The probability that a sample of 10 items will contain at most 1 defective item can be obtained by binomial as well as Poisson distribution. So, now first we try to use the binomial distribution, so the required probability is simply the probability of X equal to 0 plus probability of X equal to 1.

So, if you simply try to put here all the values in the probability mass function of the binomial distribution, this will come out to be 0.7361. And in case if you try to use the Poisson distribution to to find out the same probability that probability X equal to 0, plus probability of X equal to 1. You simply have to substitute here say λ is equal to 1 and x equal to 0, and x equal to 1. And then we try to sum them up there and then we obtain here approximately 0.7358. So, you can see that both these probabilities are going to be nearly the same.

(Refer Slide Time: 17:34)



Poisson Distribution: Mean and Variance

- **Additivity property**: The Poisson distribution possesses the additivity property that the sum of independent Poisson random variables is also a Poisson random variable.

  Suppose that X_1 and X_2 are independent Poisson random variables having respective means λ_1 and λ_2. Then $(X_1 + X_2)$ are distributed as Poisson distribution with means $(\lambda_1 + \lambda_2)$.
- **Recurrence relationship** :

  If $P(X = i) = \frac{\lambda^i exp(-\lambda)}{i!}$ then

  $P(X = i + 1) = \frac{\lambda}{i+1} P(X = i)$

$P(X=i)$

$P(X=i+1)$

12

Now, let me try to give you some important properties of Poisson distribution. There are many many properties, but I just want to illustrate here two properties; which I feel that this will be useful for you, when you are trying to work in the data science. So, one is the additive property of the Poisson distribution that sum of independent Poisson random variable is also a Poisson random variable. That means in case if you try to take two random variables X_1 and X_2 both are independent; both are following the Poisson distribution.

But, they have got different parameter X_1 is following a Poisson distribution with parameter λ; X_2 and X_2 is following the Poisson distribution with parameter λ_2 . Then their sum $X_1 + X_2$ will be following the Poisson distribution; and the parameter is going to be λ_1 plus λ_2. So, this $X_1 + X_2$ will have a probability mass function, whose mean and variance are going to be $\lambda_1 + \lambda_2$. And the next property is the Recurrence relationship and this type of relationship actually helps in computing different types of probabilities.

So, if I say the $P(X = i) = \frac{\lambda^i exp(-\lambda)}{i!}$ then $P(X = i + 1) = \frac{\lambda}{i+1} P(X = i)$. So, if somehow you can find out the value of P(X=i); then after that you can compute the P(X = i + 1); and then the subsequent probabilities without any problem.

So, now we come to an end to this lecture, and I have given you a brief background about the Poisson distribution. And I have tried my best to explain you that under what type of condition this can be used. So, now I will request you once again that you try to look into some problems from your assignment, from your books; and try to solve that probability. The main thing is this you have to understand that under this condition, the distribution can be considered as Poisson; and how to compute different types of probabilities and the properties.

And in the next turn I will try to take the example, where I will try to show you that how these probabilities can be computed directly on the R software. So, you try to practice it and I will see you in the next lecture; till then good bye.