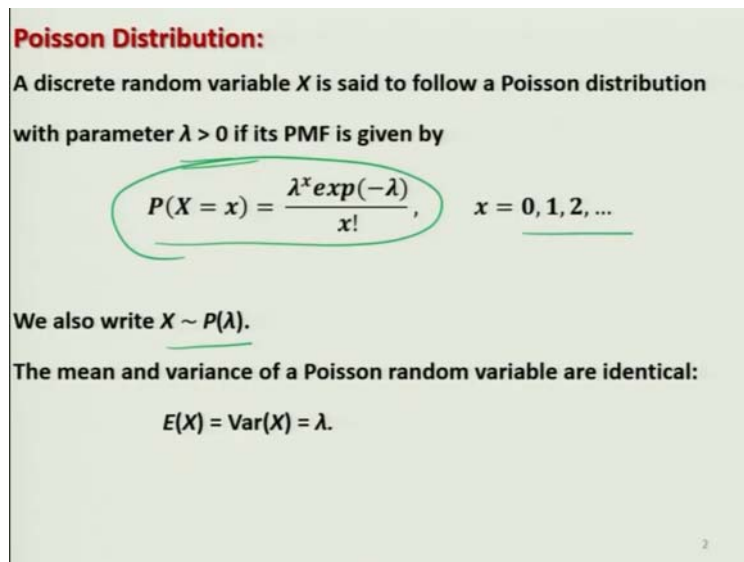


Essentials of Data Science with R Software – 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture No. 35
Poisson Distribution in R

Hello friends, welcome to the course Essentials of Data Science with R Software – 1; in which we are trying to understand the basic concepts with probability theory and statistical inference. So, you can recall that in the last lecture, I have given you the theoretical background on the Poisson probability mass function. Now, the next question is how to implement those concepts in the R software; so that is what we are going to do in this lecture.

And now I will suggest you one thing that if you have not revised the lecture on the binomial distribution; where I implemented that in R, then I would suggest you that if you can have a quick revision, then it will help you and me both in this lecture; because the same thing what I did there. Similar type of things I am going to do here. So, if you are means happy, satisfied and confident with those things; this lecture will be can be understand, it can be understood very easily. So, let us begin our lecture and if you have not revised, please have a quick revision.

(Refer Slide Time: 01:30)



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

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

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

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

$$E(X) = \text{Var}(X) = \lambda.$$

So, now we try to understand how we can implement the Poisson distribution concepts in R software. So, just for a quick revision, you can recall that a discrete random variable X is said to

follow a Poisson distribution, with parameter λ greater than 0. If, its probability mass function is given by this $P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}$, $x = 0, 1, 2, \dots$

And we also write this as x follows a Poisson distribution with parameter λ , that x it tilde Poisson P inside parenthesis λ . The mean and variance of these Poisson random variables are λ and λ ; so they are actually identical.

(Refer Slide Time: 02:15)

Poisson Distribution: Example in R
Parameters `lambda`

Usage

- `dpois(x, lambda)` gives the density,
- `ppois(q, lambda, lower.tail = TRUE)` gives the CDF,
- `qpois(p, lambda, lower.tail = TRUE)` gives the quantile function which is the smallest integer x such that $P(X \leq x) \geq p$ and
- `rpois(n, lambda)` generates the random numbers.

Invalid `lambda` will result in return value `NaN`, with a warning.

So, now in R, we will once again consider the same type of job. What we did in the case of vulnerable distribution that we would like to compute the density CDF, quantiles and random numbers. So, here the parameter λ is now written as say here `lambda`; all in lower case alphabets λ . And there is a command here `dpois`; `pois` means it is coming from P Poisson and `d` means density. So, this command `dpois` gives the density and if you want to write down; this is written as, inside the parenthesis x and then the value of the parameter λ .

And similarly, if you want to find out the CDF, you have to write the command `ppois`. And then you have to write down the data vector q , and then you have to write down the value of the parameter λ . And then you can use lower tail or lower dot tail is equal to `TRUE` or `FALSE`; exactly in the same way as we did in the case of binomial distribution. And then if you want to find out the quantile function, then you have to use the command `qpois` `q p o i s`. And inside the

parenthesis you can write down the value at which you want to find out the percentile; and then value of λ , and then lower dot tail is equal to TRUE or FALSE.

So, this will give you the quantile function, which is the smallest integer x , such that probability that capital X less than equal to small x is greater than or equal to p . And this command `rpois` inside parenthesis and λ will generate small number of random observation, from this Poisson distribution with parameter λ . And if you try to use some invalid value of λ , that will result in NaN, with a warning.

(Refer Slide Time: 04:21)

Poisson Distribution: In R

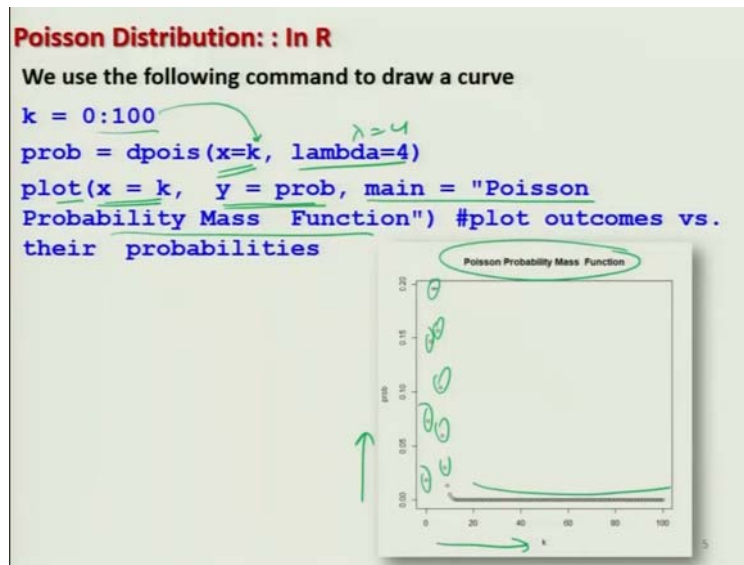
Arguments

- x** vector of (non-negative integer) quantiles.
- q** vector of quantiles.
- p** vector of probabilities.
- n** number of random values to return.
- lambda** vector of (non-negative) means.
- lower.tail** logical; if **TRUE** (default), probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

Handwritten notes: $1 - F(x)$ under $P[X > x]$ and $F(x)$ under $P[X \leq x]$.

So, these are the thing which I just explained you here that x is the vector of non-negative integer. And then q is the vector of quantile, p is the vector of probabilities, n the number of random values to be return, λ is the vector of non-negative means and lower tail this is the logical operator, which takes value TRUE. That will compute the value of $F(x)$, and if it FALSE quick and it will try to compute the value of probability X greater than x , which is 1 minus $F(x)$.

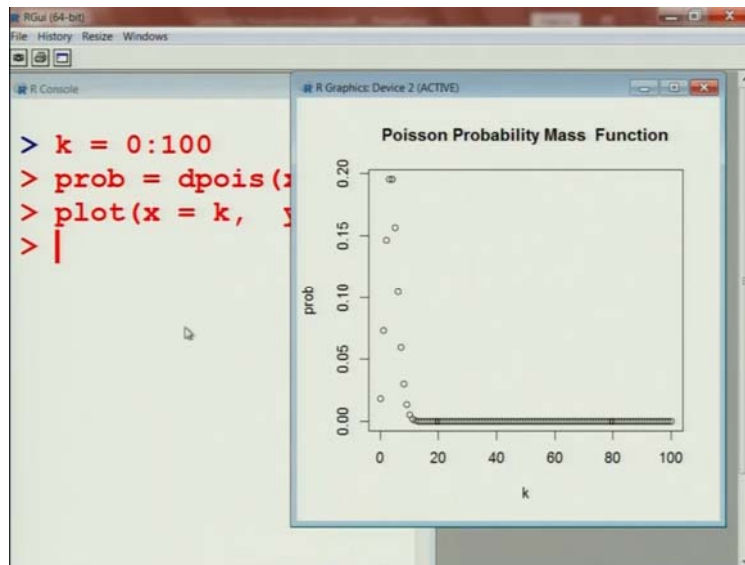
(Refer Slide Time: 04:54)



So, now let us try to first understand how these probabilities will look like. So, what I try to do here that I try to choose here a particular value of λ , say λ is equal to here 4; and then I try to choose the value of x say from 0 to 100, and which is controlled by the variable k . And then I try to plot this function here as plot x equal to k and y equal to $prob$. So, on the x -axis we will have the value of k equal to 0 to 100. And on the y -axis, we will have the value of the probabilities, which are the density values; which are generated from the Poisson distribution for different values of k .

And you can see here that this curve look like, this is the point, this is the point, point; and then it is trying to decrease and after that it becomes like this. So, this will give you an idea that how this curve will look like or how these points will look like, when they are plotted on the x/y axis. And after this, this mean this is going to give you the title of the curve and so on.

(Refer Slide Time: 06:11)



So, now in case if you try to plot this curve on the R console; so you can see yourself. What is really happening? So I simply try to copy and paste these commands on the R console; you can see here this will come out to be. I simply paste and you can see here this type of graphic will be obtained. And if you want to change the value of λ , you can change the value of λ ; and you can yourself observe that what really happens.

(Refer Slide Time: 06:36)

Poisson Distribution: Example in R

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.

```
> dpois(x=2, lambda=4)
[1] 0.1465251
```

computes the density of Poisson distribution with $\lambda = \text{lambda} = 4$ as

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

Alternatively

```
> dpois(2, 4) # default positions are x and lambda
[1] 0.1465251
```

So, this will give you an idea that how the this graphic look like for different values of λ , and for different values of x . So, that is my idea what I wanted to inform you. Now, I try to do the same example which I did in the last lecture; but I had obtained those probabilities using the theoretical concept.

So, the first example that we did was that suppose a country experiences 4 earthquakes on an average per year. Then the probability of suffering from only two earthquakes is obtained by the Poisson distribution; and we had obtained by like the probability that X equal to 2, that was obtained by this function. Now, I try to obtain the same value using the R command; so now we want to compute the density. So, I have to use the command `dpois` and I want to compute the probability at X equal to 2. So, this X equal to 2 will come here; X equal to 2 and then whatever is the value of λ ; this I have to write down λ is equal to 4.

And then you will get this value and you can compare; you are getting the same value that you got earlier. And alternatively, if you do not want to write x or λ , then you can write down `dpois 2` and 4. But, the but you have to be careful that the first position is occupied by x , and the second position is occupied by λ ; these are the default position. If you interchange it, you will get a wrong answer.

(Refer Slide Time: 08:12)

Poisson Distribution: Example in R

`qpois(q, lambda, lower.tail = TRUE)` calculates the quantile.

For example, suppose we want to determine the 60% quantile q which describes that $P(X \leq q) \geq 0.6$ can be obtained by the command

```
> qpois(0.6, lambda=4)
```

[1] 4

or equivalently

```
> qpois(0.6, lambda=4, lower.tail= TRUE)
```

[1] 4

Handwritten notes: 0.6 is circled in the first command. An arrow points from 0.6 to the output 4. A note says 'q=0.6 -> R gives error'.

So, now we try to understand how one can find the quantile of this Poisson distribution from R. So, for that we have the command `qpois`; this calculates the quantile. For example, in the same example where we have taken λ equal to 4; suppose we want to compute the 60 percent quantile. That is we want to compute the value of k of q , such that probability that x less than equal to q is greater than or equal to 0.6. So, this can be obtained by the command `qpois 0.6` and λ equal to 4; and this will give the value 4.

And similarly, if you want to use the option `lower.tail = TRUE`; that will also give you the same value 4. One thing what you have to be little bit cautious that when you are writing q , then you have to write here 06; if you try to write down q is equal to 0.6, then R will give you an error. Why this happened? I do not know, but that is the way it works. So, anyway that is not big deal.

(Refer Slide Time: 09:20)

Poisson Distribution: Example in R

For example, the probability of suffering from two to four earthquakes is obtained as follows:

Let X : Number of earthquakes

To find

$$P(2 \leq X \leq 4) = P(X \leq 4) - P(X \leq 2)$$
$$= F(4) - F(2)$$
$$= \text{ppois}(q=4, \text{lambda}=4) - \text{ppois}(q=2, \text{lambda}=4)$$

[1] 0.3907336

10

So, similarly if you want to compute the probability of suffering 2 to 4 earthquakes in the same example that we just consider. That means we are interested in finding or the probability that X lying between 2 and 4. So, you know from the rules of your cumulative distribution function that this can be written as probability that X less than equal to 4, minus probability X less than equal to 2; which is equal to $F(4)$ minus $F(2)$. And, now how to compute this $F(4)$ and $F(2)$? You

simply have to write down ppois q equal to 4 for here and ppois is equal to 2. And λ will remain the 4 and then you can obtain this value.

So, you can see computing different types of probabilities is very simple, when you are trying to do it in the R software. But, now let me try to first show you that whether you can obtain these values in the R software or not.

(Refer Slide Time: 10:19)

Poisson Distribution: Example in R

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.

```
> dpois(x=2, lambda=4)
```

```
[1] 0.1465251
```

computes the density of Poisson distribution with $\lambda = \lambda = 4$ as

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

Alternatively

```
> dpois(2, 4) # default positions are x and lambda
```

```
[1] 0.1465251
```

Poisson Distribution: Example in R

```
ppois(q, lambda, lower.tail = TRUE)
```

calculate the CDF $F(q) = P(X \leq q)$ at any point q .

For example, the probability of suffering from two or more earthquakes is obtained as follows from $P(X \geq 2) = 1 - F(1)$, then we write

$1 - P(X \leq 1)$

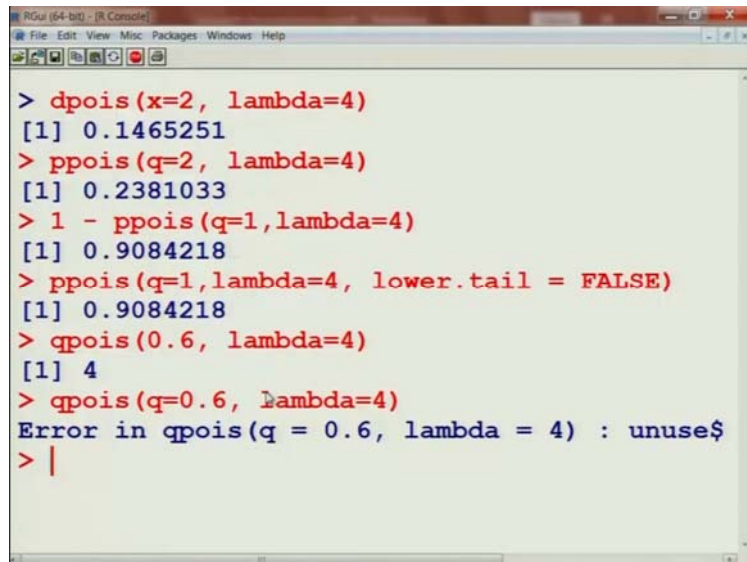
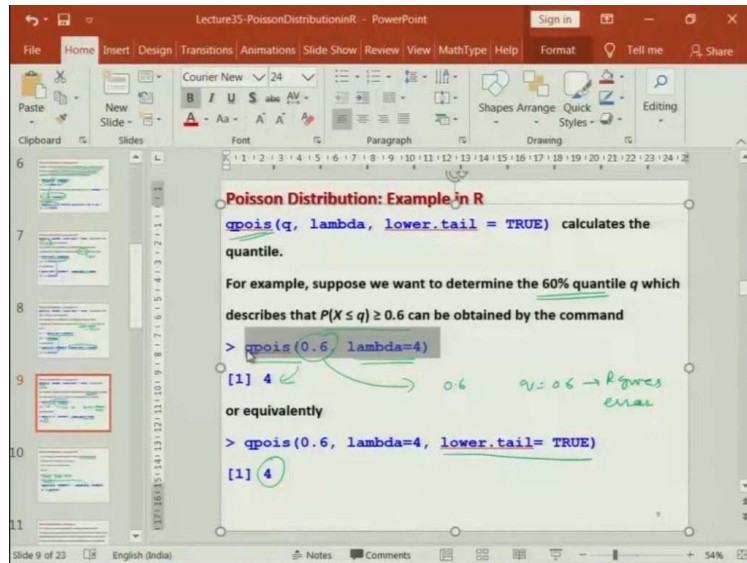
```
> 1 - ppois(q=1, lambda=4)
```

```
[1] 0.9084218
```

or equivalently

```
> ppois(q=1, lambda=4, lower.tail = FALSE)
```

```
[1] 0.9084218
```

So, if I try to take this density at x equal to 2; I try to compute it in the R software. You can see here that density at x equal to 2 with λ equal to 4, is coming out to be without any problem. And similarly if you write try to write down this CDF at x equal to 2; you simply have to use the command `ppois`; that will give you the value of the cumulative probability without any problem. And in case if you want to use the probability that x is greater than or equal to 2; that is also can be obtained without any problem, you can see here.

And if you want to use here this option that you want to use the command `lower.tail` is equal to `FALSE`. You can see that both these values are coming out to be the same; so there is no issue,

there should not be any confusion. And similarly, if you want to find out the the 60 percent quantile, this will come out over 4. And now I would like to show you that in the same command, if you try to write down q is equal to 0.6; this gives a sort of error, so do not worry. So, this is working, now we come back to our slides and try to take one more example.

(Refer Slide Time: 11:47)

Poisson Distribution: Example 4

If an insurance company handles 6 average number of claims everyday, what proportion of days have less than 3 claims? What is the probability that there will be 4 claims in exactly 3 of the next 5 days? Assume that the number of claims on different days is independent.

The company insures a large number of clients, each having a small probability of making a claim on any given day. So it is reasonable to suppose that the number of claims handled daily, call it X , is a Poisson random variable with $E(X) = 6$.

11

So, suppose there is an insurance company that handles 6 on an average number of claims, means every day. Now, the question is what proportion of days have less than 3 claims? And what is the probability that there will be 4 claims in exactly 3 of the next 5 days? So, these types of questions are very popular in data sciences in real life. So, now we can assume that the number of claims on different days, they are independent of each other.

So, the company ensures a large number of clients, and each having a probability of making a claim on any given day; so, that is pretty small actually. So, it is reasonable to suppose that the number of claims handled daily, will be a Poisson random variable. And if this random variable is indicated by X , then the expected value of X is going to be 6; which is the value of λ .

(Refer Slide Time: 12:49)

Poisson Distribution: Example 4

Since $E(X) = \lambda = 6$, the probability that there will be fewer than 3 claims on any given day is

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) = F(3)$$
$$= \text{ppois}(q=3, \text{lambda}=6) = 0.1512039$$

Since any given day will have fewer than 3 claims with probability 0.151, it follows that over the long run 15.1% of days will have fewer than 3 claims.

12

Now, suppose if I want to find out that there will be fewer than 3 claims on any given day; that means I want to find out the probability that X is less than equal to 3. That is probability of X equal to 0, plus probability of X equal to 1, plus probability of X equal to 2; or this is your F_3 . So, I can compute this F_3 using the command `ppois`, for q equal to 3 and λ is equal to 6; which will come out to be 0.15 approximately.

So, you can see that computing the CDF and computing different types of probability is very easy. And since on any given day, we will have fewer than 3 claims with probability only 0.151; very close to 15 percent. So, it follows that over the long run and 15.1 percent of days will have fewer than 3 claims. So, you can see that computing these types of probabilities CDF etc. is not difficult, when you are trying to do it in the R software.

(Refer Slide Time: 13:47)

Poisson Distribution: Example in R

`rpois(n, lambda)` generates n random numbers from $P(\lambda)$.

For example, suppose we want to generate 5 random numbers from a Poisson distribution $P(4)$ which can be obtained by the command

```
rpois(n=5, lambda=4)
[1] 5 3 1 1 4
```

13

Now, we try to see how the random numbers can be generated from this Poisson distribution. So, you know that the command for generating the random number is `rpois`; and you have to give n and the value of λ . And this will generate the small n number of random numbers, from the Poisson with parameter λ . Like for example, if you want to generate 5 random numbers from Poisson distribution with λ equal to 4; then we have to write down `rpois`, n equal to 5; and λ is equal to 4 definitely. When I am trying to show it on the R console, these numbers will not come; but something else will come.

(Refer Slide Time: 14:31)

Poisson Distribution : Example in R

Suppose an experiment is conducted 100 times under Poisson distribution with $\lambda = 4$. We want to plot the density of these observations.

```
n = 100 /
lambda = 4 /
k = 0:n
pmfpoi = dpois(k, lambda)
plot(k, pmfpoi)
```

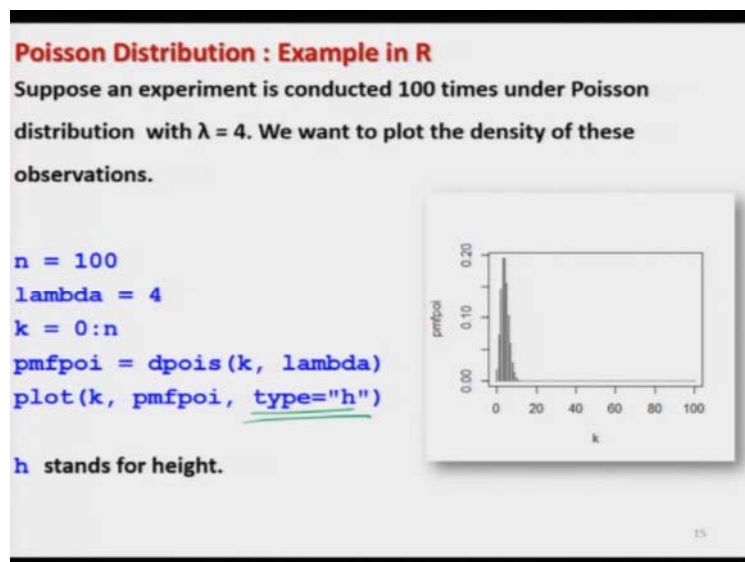
14

So, now let me try to give you some idea that you are considering this type of event, and suppose you want to know that what will happen to the densities as this x is changing. So, it means how the probabilities are going to change. Because once you are trying to work in data sciences, you try to explore various types of aspects of the given experiment. So, now I am assuming that our experiment is now given to us and that means the experiment is suppose conducted 100 times and here the λ is equal to 4 that we know and we want to see how the densities change.

So, if we try to make a plot, so we try to take n equal to 100 λ equal to 4, and we try to change the value of this x by controlled by this k from 0 to n . And so that mean k takes the value 0 to 100; and then we try to obtain the densities; and then we try to plot these densities with respect to k . So, in this curve, you can see we have k and we have the densities.

You can see that at 0, the probability is nearly 0; and then it is increasing and after sometime this starts decreasing. And after that it becomes nearly the constant, so that mean if you try to beyond k equal to 13, 14, 15 like this number, the probabilities are nearly the same. So, this is how you can extract some information about the process using this simple command from the R.

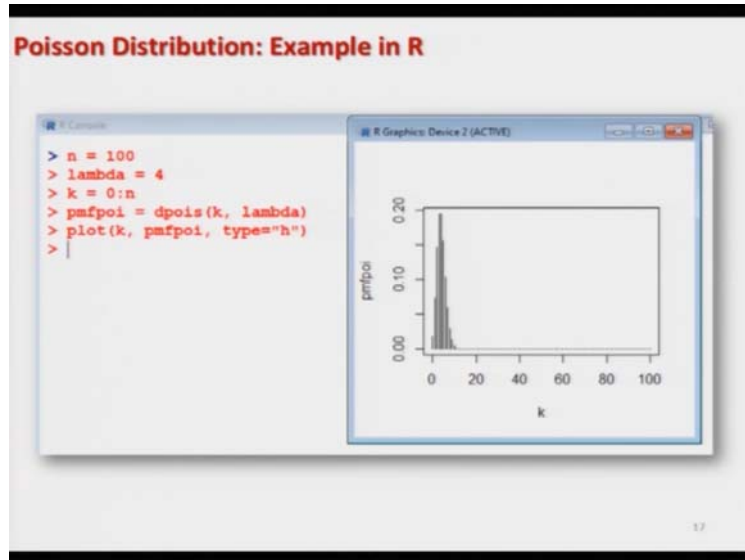
(Refer Slide Time: 16:05)



And if you want to create the say this sketch up plot using the height density line; so you have to change the command and you have to add here type is equal to h; and you will get this type of command. So, if you try to look different commands or different types of graphics; then it will

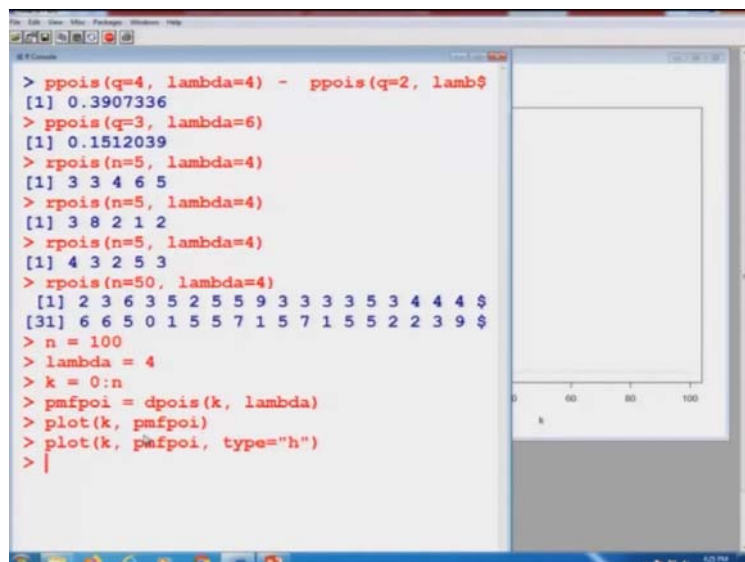
give you different types of information. And based on which you are trying to will get different types of conclusion.

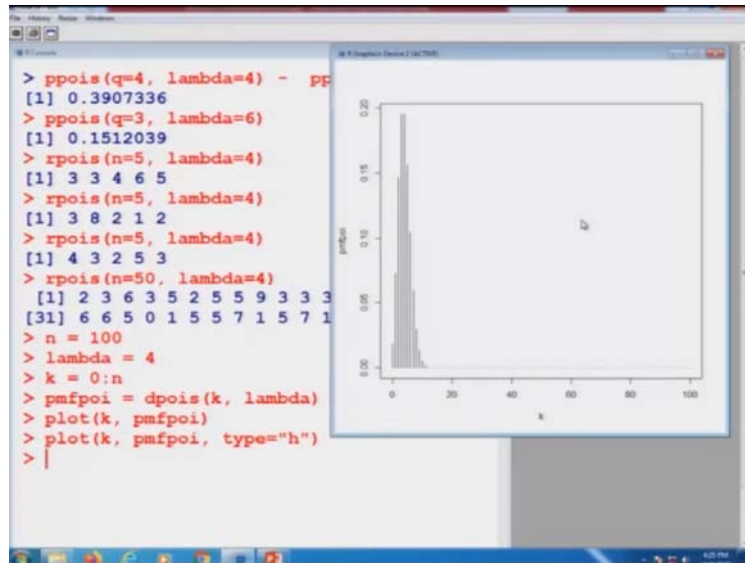
(Refer Slide Time: 16:34)



So, these are the graphics that you have seen and I will try to show you these things on the R consoles also; so that you get more convinced.

(Refer Slide Time: 16:45)





So, you can see that this probability is computed like this, which is the same thing that you have just obtained. And now after this, if you try to compute this probability, here ppois is that you can do very easily I just want to show you here, so that you are confident that whatever is written on the slide is also correct.

Now, in case if you try to generate 5 random observations from the distribution λ equal to 4. You can see that this 5 numbers are 3, 3, 4, 6, 5 and if you try to repeat it; you will get here a different set of number. And if you again repeat it, you will get a different set of number. And if you try to obtain, suppose 50 observations, you can simply say n equal to 50; and you will get 50 this value. So, do not you think that this looks like, like as as if you have repeated the experiment 50 times. So, this is the advantage of the working on the software.

And similarly, if you try to plot the curve or graphic, which I have shown you; you can see this will come out to be like this. Same graphic which I have shown you and if you try to make the type is equal to h; you can see here. Now, this graphic is changed to be like this, they are high density lines. So, now let me comeback to our slides and try to understand the next topic.

(Refer Slide Time: 18:16)

Poisson Distribution: Mean and Variance
`rpois(n, lambda=4)` generates `n` random numbers from $P(\lambda=4)$.
Mean $E(X) = 4$
 $Var(X) = 4$
Now we generate the random numbers and calculate their mean and variance as follows:
`x = rpois(n, lambda=4)`
`mean(x)`
`var(x)`
Compare the simulated mean and variance with the theoretical mean and variance.

So, for example, under the same example if you try to find out the mean and variance; the expected value of X will come out to be same, as λ and variance will come out to be same as the λ . So, now just to have an idea about this process that how this theoretical mean and theoretical variance are close to the values that are obtained in different experiments. So, we try to generate the n number of observation for different values of n , from the same distribution with λ equal to 4 and we try to compute their mean and variance; and we try to essentially simulate this probability model.

(Refer Slide Time: 18:55)

Poisson Distribution: Mean and Variance
Observe the difference with theoretical $E(X) = 4$, $Var(X) = 4$

```
> x=rpois(10, lambda=4) # 10 observations
```

<pre>> mean(x) [1] 3.2</pre>	<pre>> var(x) [1] 5.066667</pre>
---------------------------------	-------------------------------------

```
> x=rpois(10, lambda=4) # 10 observations
```

<pre>> mean(x) [1] 4.4</pre>	<pre>> var(x) [1] 2.044444</pre>
---------------------------------	-------------------------------------

```
> x=rpois(10, lambda=4) # 10 observations
```

<pre>> mean(x) [1] 3.1</pre>	<pre>> var(x) [1] 5.877778</pre>
---------------------------------	-------------------------------------

And if try to see that in this case the expected value of X is 4; that is the mean is population mean is 4, and population variance is 4. Now, if you try to get here 10 values and if you try to find out their mean; this comes out to be 3.2 and variance comes out to be 5.06. So, you can see these values are quite away from the true mean and true variance. And if you try to repeat it, means every time you will get a different value of mean, and different value of your variance. That is obvious because there is a difference in different samples.

(Refer Slide Time: 19:29)

Poisson Distribution: Mean and Variance

Observe the difference with theoretical $E(X) = 4$, $Var(X) = 4$

```
> x=rpois(1000, lambda=4) # 1000 observations
```

<pre>> mean(x) [1] 4.015</pre>	<pre>> var(x) [1] 4.152928</pre>
-----------------------------------	-------------------------------------

```
> x=rpois(1000, lambda=4) # 1000 observations
```

<pre>> mean(x) [1] 3.882</pre>	<pre>> var(x) [1] 3.607684</pre>
-----------------------------------	-------------------------------------

```
> x=rpois(1000, lambda=4) # 1000 observations
```

<pre>> mean(x) [1] 4.003</pre>	<pre>> var(x) [1] 3.814806</pre>
-----------------------------------	-------------------------------------

20

Now, in case if you try to increase the number of observation from 10 to 1000; then you can see this mean values are coming to coming out to be 4.01, 3.8 to 4.0. And they are more close to that two theoretical mean 4. And similarly, the variance comes out to be 4.15, 3.60, 3.81 which is now more closer to 4. But, if you try to increase it more that will become more closer to the true value. So, these types of observations you will be getting from such simulation experiments that you are trying to do.

(Refer Slide Time: 20:03)

Poisson Distribution: Mean and Variance

```
R Console
> x=rpois(10, lambda=4)
> mean(x)
[1] 3.2
> var(x)
[1] 5.066667
>
> x=rpois(10, lambda=4)
> mean(x)
[1] 4.4
> var(x)
[1] 2.044444
>
> x=rpois(10, lambda=4)
> mean(x)
[1] 3.1
> var(x)
[1] 5.877778
```

21

Poisson Distribution: Mean and Variance

```
R Console
> x=rpois(1000, lambda=4)
> mean(x)
[1] 4.015
> var(x)
[1] 4.152928
>
> x=rpois(1000, lambda=4)
> mean(x)
[1] 3.882
> var(x)
[1] 3.607684
>
> x=rpois(1000, lambda=4)
> mean(x)
[1] 4.003
> var(x)
[1] 3.814806
```

22

So, these are the screenshots of the same outcome which I have shown you. So, now we come to an end to this lecture, and you can see that in this lecture that was pretty simple. Exactly on the same line what we did in the case of binomial distribution; we have used the R software to compute different types of probabilities from the Poisson distribution.

But, definitely it will be more helpful for you, if you try to take some example; and try to solve them yourself. Well, I will say once again they try to take very simple values, try to compute the probability manually; so that you understand the basic concepts, how the things are being done

and then you try to use them in the R software. That will give you more confidence that whatever software is doing, that is correct and that and that is according to the rule that you know.

And this will help you when you are trying to deal with much bigger data sizes; particularly when you are trying to work in the data sciences, you have to make automated program, possibly this concepts and this confidence that you have gained with this smaller example they will help you in becoming more assured that whatever you are thinking, that is being done; and whatever the outcome you are getting, you understand it. So, you try to practice it and I will see you in the next lecture; till then good bye.