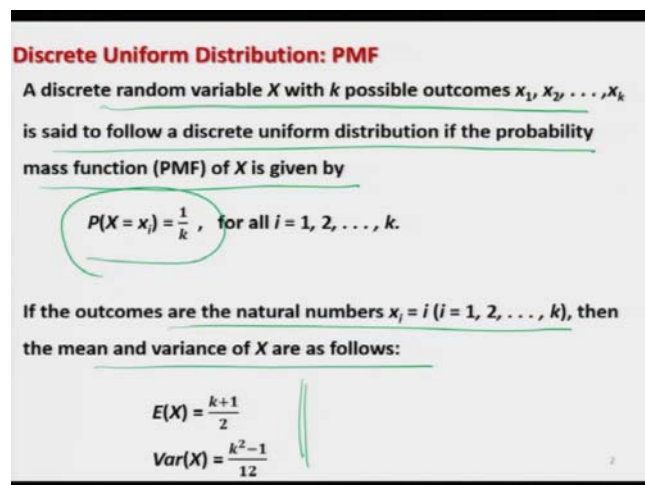


Essential of Data Science with R Software – 1
Probability Theory and Statistical Inference
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture – 31
Discrete Uniform Distribution in R

Hello friends welcome to the course essentials of data science with R software 1, in which we are trying to understand the basic concepts related to the probability theory and statistical inference. So, you can recall that in the last lecture, we had discussed about the discrete uniform distribution. We understood the probability mass function that how it is coming into existence and what is the interpretation and we were convinced that it is trying to reflect the phenomena in which the probabilities of all the outcomes are going to be the same.

And we also had indicated or we have found the mean and variance of those things. But now, the question is when I say that X is a random variable, which is following the following probability mass function, for example, in this case, it is a uniform, discrete uniform distribution, what does this mean, what is this X , how it looks like? And then what is the meaning of this mean and variants and how these things can work. So, to illustrate those things, now, we will try to take the help of R software and we try to understand what is the meaning of these concepts. So, let us try to begin this lecture and try to learn a couple of things that related to the discrete uniform distribution in this lecture.

(Refer Slide Time: 01:39)



Discrete Uniform Distribution: PMF

A discrete random variable X with k possible outcomes x_1, x_2, \dots, x_k is said to follow a discrete uniform distribution if the probability mass function (PMF) of X is given by

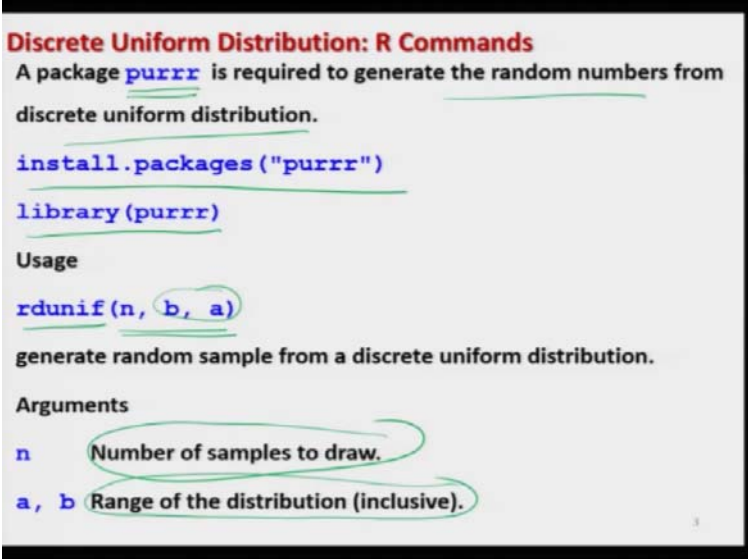
$$P(X = x_i) = \frac{1}{k}, \text{ for all } i = 1, 2, \dots, k.$$

If the outcomes are the natural numbers $x_i = i$ ($i = 1, 2, \dots, k$), then the mean and variance of X are as follows:

$$E(X) = \frac{k+1}{2}$$
$$Var(X) = \frac{k^2-1}{12}$$

So, first let us try to have a quick review what we had given in the last lecture as the probability mass function of X . So, we had discussed that a discrete random variable X with k possible outcomes x_1, x_2, \dots, x_k is set to follow a discrete uniform distribution if the probability mass function of x is given by this function probability of x equal to x_i is $1/k$ for all i goes from 1 to k . And we also had to discuss that if all the outcomes are the natural numbers 1, 2, 3, 4 up to k then the mean and variance of x are obtained like this mean is $(k + 1)/2$ and variance is $(k^2 - 1)/12$.

(Refer Slide Time: 02:19)



Discrete Uniform Distribution: R Commands
A package **purrr** is required to generate the random numbers from discrete uniform distribution.

```
install.packages("purrr")  
library(purrr)
```

Usage

```
rdunif(n, b, a)
```

generate random sample from a discrete uniform distribution.

Arguments

- n** Number of samples to draw.
- a, b** Range of the distribution (inclusive).

Now, let us try to do the same thing in the R software, well in the base software, details of discrete uniform distribution are not available. And in order to understand or if we want to employ the discrete uniform distribution, we need a package whose name is purrr means p u triple r, that is how is the name of the package. So, this package is required to generate the random numbers from discrete uniform distribution and other types of things can be done.

So, first we try to install this package using the command install dot packages p u triple r and then we try to load this package by using the command library. Now, once you do it, then the syntax for generating the random numbers from this discrete uniform distribution is rdunif r d u n i f, inside the parenthesis you have to write three parameters n , b and a where n is the number of samples to be drawn and b and a they are going to be the values, are going to be the values, which are going to provide the range of the distribution and they are, a and b are inclusive.

(Refer Slide Time: 03:45)

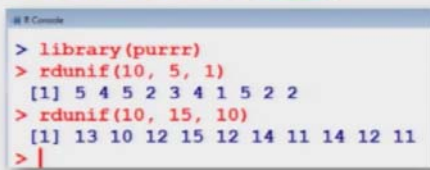
Discrete Uniform Distribution: Example 2 in R
Suppose we want to generate 10 observations out of 1, 2, 3, 4 and 5.

Then

```
rdunif(10, 5, 1)
[1] 5 4 5 2 3 4 1 5 2 2
```

Similarly, suppose we want to generate 10 observations out of 10, 11, 12, 13, 14 and 15. Then

```
rdunif(10, 15, 10)
[1] 13 10 12 15 12 14 11 14 12 11
```



```
> library(purrr)
> rdunif(10, 5, 1)
[1] 5 4 5 2 3 4 1 5 2 2
> rdunif(10, 15, 10)
[1] 13 10 12 15 12 14 11 14 12 11
> |
```

Discrete Uniform Distribution: PMF
A discrete random variable X with k possible outcomes x_1, x_2, \dots, x_k is said to follow a discrete uniform distribution if the probability mass function (PMF) of X is given by

$$P(X = x_i) = \frac{1}{k}, \text{ for all } i = 1, 2, \dots, k. \quad P(X = x_i)$$

If the outcomes are the natural numbers $x_i = i$ ($i = 1, 2, \dots, k$), then the mean and variance of X are as follows:

$$E(X) = \frac{k+1}{2}$$
$$Var(X) = \frac{k^2-1}{12}$$

So, first let us try to understand what does this actually mean? So, suppose we want to generate 10 observations out of 1 2 3 4 and 5. Suppose I have a population of five numbers who can assume you can imagine and we want to generate say 10 observations, which are following the probability mass function of this discrete uniform distribution like this 1. So, in this case, k is equal to here 5.

So, the probability distribution will be probability that x equal to x_i is 1 upon 5. So, now means how to generate the random numbers that mean what are the possible values of x that you can obtain which will look like as if these values are generated from this discrete uniform

distribution 1 upon 5. So, for that doing in R is very simple. Just try to write down here `rdunif` the number of observations 10 and the value of b and a. So, these are here 5 and 1, because 5 is the maximum value and 1 is the minimum value and 1 2 3 4 5 all are integer.

So, now you can see here you are getting here the values like 5 4 5 2 3 4 1 5 2 2, you can see here one, two, three, four, five, six, seven, eight, nine, ten. So, you have got here ten values which are lying between 1 and 5 and you can believe that, you have conducted an experiment and these are that ten outcomes. And similarly, if you want to generate say ten observation from some phenomena in which the values of x are say 10, 11, 12, 13, 14 or 15.

So, in that case I can write my command here as a `rdunif` to 10 and then the value of b will be here 15, a is going to be here 10 and if you try to see here that you can get here in this type of outcome without any problem, and you can see here all the values are lying between 10 and 15. And definitely in case if you try to repeat this experiment, do you think that are you going to get the same cutoff values? Certainly not. And even when I will try to show you it on the R console you will not get the same values what you are getting here.

(Refer Slide Time: 06:25)

Discrete Uniform Distribution: Example 2 in R
Suppose we want to generate observations out of 1, 2, ..., 100.
Then we find its mean and variance and compare with theoretical results.

Theoretical mean = $E(X) = \frac{100+1}{2} = 50.5$

Theoretical variance = $Var(X) = \frac{100^2-1}{12} = 833.25$

So, I will try to first complete this part and then I will try to come to the R console to show you and then I will try to convince you with all these arguments. Now, we have understood the meaning of mean and variance and we try to indicate them as the first raw moment and second central moment and I had told you at that time, these are the values in the population that means,

if you try to draw all possible samples or in case if you try to consider all possible values in the population, then the mean will come out to be 50.5 in this case, and that is possible.

Because if you try to see here that there are 100 values. So, if you try to find out the mean of 1 plus 2 plus up to 100 and then you can find out that this is going to be 50.5 and you can see that this is the middle value which is lying somewhere in the mid. And similarly, if you try to find out the variance, the value will come out to be 833.25. But, my question is this, when you are trying to draw a sample, the sample is not always going to consist of all the 100 values.

Suppose your sample is less sample can be more than 100 also. So, what is happening? How does this make any sense? What is the interpretation of this theoretical mean and theoretical variance when you are trying to use them in a real life, this I would like to show you through some simulations, which I have done here.

(Refer Slide Time: 08:20)

Discrete Uniform Distribution: Example 2 in R
Observe the difference with theoretical $E(X) = 50.5$, $Var(X) = 833.25$

$b=$ $a=1$

```
> x = rdunif(10, 100, 1) # 10 observations
```

<pre>> mean(x) [1] 50.7</pre>	<pre>> var(x) [1] 839.1222</pre>
----------------------------------	-------------------------------------

```
> x = rdunif(10, 100, 1) # 10 observations Repeat
```

<pre>> mean(x) [1] 63.3</pre>	<pre>> var(x) [1] 910.4556</pre>
----------------------------------	-------------------------------------

```
> x = rdunif(10, 100, 1) # 10 observations Repeat
```

<pre>> mean(x) [1] 59.1</pre>	<pre>> var(x) [1] 1136.989</pre>
----------------------------------	-------------------------------------

The screenshot shows three rows of R console output. Each row starts with a command to generate 10 observations from a discrete uniform distribution with parameters (10, 100, 1). The first row shows a mean of 50.7 and a variance of 839.1222. The second row, labeled 'Repeat', shows a mean of 63.3 and a variance of 910.4556. The third row, also labeled 'Repeat', shows a mean of 59.1 and a variance of 1136.989. Handwritten annotations in green and blue highlight the theoretical values 50.5 and 833.25, and checkmarks are placed next to the sample results. The parameters 'b=' and 'a=1' are also annotated in green.

Discrete Uniform Distribution: Example 2 in R
 Suppose we want to generate observations out of 1, 2, ..., 100.
 Then we find its mean and variance and compare with theoretical results.

Theoretical mean = $E(X) = \frac{100+1}{2} = 50.5$

Theoretical variance = $Var(X) = \frac{100^2-1}{12} = 833.25$

So, now, what I try to do here that, in this case, for example, suppose I have found that expected value of X is 50.5 and variance of x is 833.25. So, what I try to do here, I try to simulate the same thing inside the R software you can see here that these are the observation between 1 and 100. So, means I can take the value of a and b to be 1 and 100 and I can see here that a equal to 1 b is equal to 100 and we try to generate the observation from discrete uniform. So, the command here will be `rdunif` inside parenthesis 10, 100, 1, for generating the 10 observations.

So, I try to generate these 10 observations and then I try to find out the mean of those observations. So, what will happen? What are those numbers that I am not considering, I am not bothered about them, but I simply want to generate that sample and I am trying to find out the mean and variance of those values which are generated and stored under this a data vector x.

So, once I try to do it, I get here a value here 50.7 for the mean and 839.1222 for the variance. Now, can you compare here these values this 50.7 and this 50.5. You can see here that the true value here is 50.5 but then you are getting here a value 50.7 So, that is going to be different and the value of variance should be 833.25, but then you are trying to get here 839.1222. So, this is also different, but they are pretty close.

Now, in case if you try to repeat the same experiment again and again and if you try to find out the value of mean and variance, do you think that are you going to get the same thing for example, can we try to show you here I try to repeat the same experiment here I try to generate

once again can 10 observations. I try to repeat it, then you can see here this time I am getting here a value of mean as 63.3 and the variances as 910.4556.

Which is very different from the value which is here 50.5 the true value and the true theoretical variance is 833.25 and similarly, if I try to repeat this experiment once again. So, this time I am getting here the values here of the mean as 59.1 and the variance 1136.989. So, this is changing. So, now, with this example, I want to convey a couple of things, number one, you can see here that this mean or variance, they are depending on the random numbers which are getting generated.

And since this is a random sample, that means, before conducting the experiment, we do not know what is the outcome that is why these observations will be changing in every draw. And then once these observations are changing in every draw, the mean and variance are also going to be changed. So, that means, mean and variance are also random variable. Do you agree with me?

Well, we have done that, if x is a random variable, then any function of the random variable is also a random variable. So, since mean and variance they are the function of the random variable, so, that is why the mean and variance that you are trying to compute on the basis of a random sample they will also become random. So, in fact, the mean and variance they are, they are the functions of random variables, so, they are random.

And whenever you are trying to generate the value of mean and variance from different sets of samples from different sets of observation, they will also be changing, because they are random. Now, question comes here that sometimes you are getting the value of your mean as 50.7, sometimes you are getting 63.3 sometimes you get 59.1 and same is true for the variance also. So, but that true value is 50.5. So, how can you ensure or what should you do so, that this value is coming close to 50.5 and the variances close to 833.25.

(Refer Slide Time: 13:10)

Discrete Uniform Distribution: Example 2 in R

Observe the difference with theoretical $E(X) = 50.5$, $Var(X) = 833.25$

```
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.5494
> var(x)
[1] 838.4448
```

```
> x = rdunif(10000, 100, 1) #10000 observations
> mean(x)
[1] 50.7379
> var(x)
[1] 833.3747
```

```
> x = rdunif(10000, 100, 1) #10000 observations
> mean(x)
[1] 50.5403
> var(x)
[1] 824.5292
```

So, now, what we try to do here, that we try to make an experiment and we try to increase the number of observation from the same distribution. So, here once again a is equal to 1, b is equal to here 100. And with now, we are trying to generate 10,000 observations and we try to find out the mean and variance based on 10,000 observations. So, you can see here, when I tried to generate 10,000 observation, which are stored in the variable x and then I tried to find out their mean it comes out to be 50.5494 and variance comes out to be 838.4448.

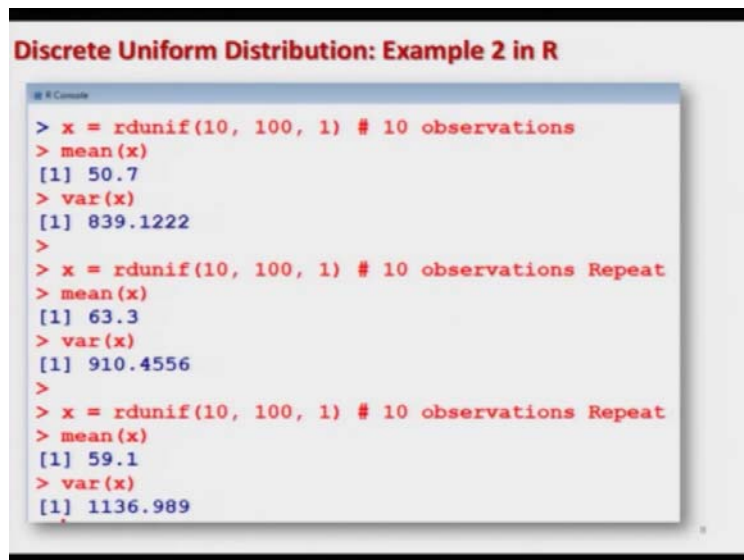
So, that means, this value 50.54 you can see this is very close to 50.5 and variance 838.4448 is very close to 833.25. And if you try to repeat this experiment that you try to repeatedly generate 10,000 observations and again compute the mean and variance these values you can see here they are coming out to be mean as 50.73 and the variance is coming out to be 833.37 and similarly, if you try to repeat it, the mean is coming out to be 50.54 and the variances coming out to be 824.

So, you can see here that when you are trying to generate more number of observations, the computed values are getting closer to the true theoretical values of mean and variance. Why this is happening, what is the reason behind it that we will try to see, but my objective was here to show you that what is the meaning of theoretical mean and theoretical variance. So, in case if you try to take your all possible samples and then try to find out the mean without any variability that will exactly come out to be 50.5.

And the variance will come out to be exactly 833.25 but it will not happen in real life. But, the theoretical value will give you an idea that whenever you are trying to conduct the experiment, the type of statistical inferences, what you are going to draw based on different statistical tool, you can judge about them, for example, here I can judge that if I do not know the mean and variance and if I try to find out the simple arithmetic mean and variance of the sample observation, then are they going to determine the true value of mean and variance or not?

So, this is how we try to proceed in data science that we try to look into the data we try to make different types of simulation and we try to conclude something which is going to help us. So, now, we can just try to do these things on the R console, so, that I can show you.

(Refer Slide Time: 16:04)



```
Discrete Uniform Distribution: Example 2 in R
R Console
> x = rdunif(10, 100, 1) # 10 observations
> mean(x)
[1] 50.7
> var(x)
[1] 839.1222
>
> x = rdunif(10, 100, 1) # 10 observations Repeat
> mean(x)
[1] 63.3
> var(x)
[1] 910.4556
>
> x = rdunif(10, 100, 1) # 10 observations Repeat
> mean(x)
[1] 59.1
> var(x)
[1] 1136.989
```

Discrete Uniform Distribution: Example 2 in R

```
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.5494
> var(x)
[1] 838.4448
>
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.7379
> var(x)
[1] 833.3747
>
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.5403
> var(x)
[1] 824.5292
```

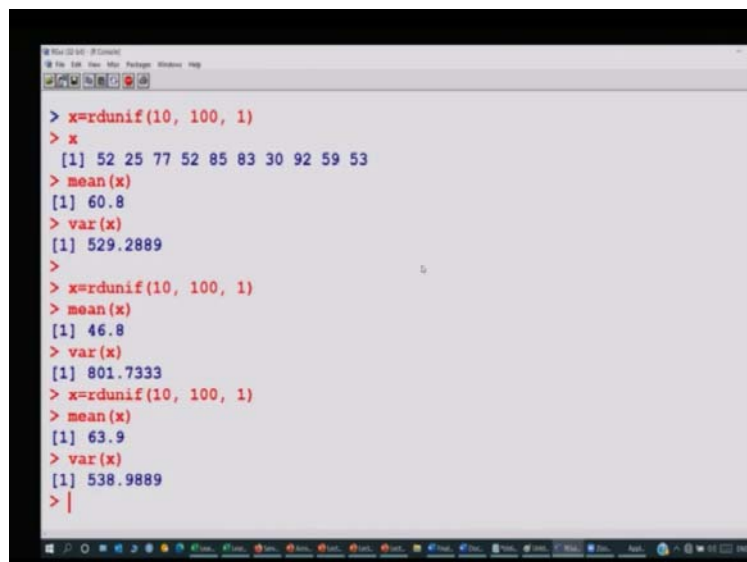
```
> library(purrr)
> rdunif(10, 10, 1)
[1] 5 5 6 2 10 1 6 7 3 5
> rdunif(10, 10, 1)
[1] 3 4 5 7 2 10 3 7 5 10
> rdunif(10, 10, 1)
[1] 2 1 3 6 10 7 1 7 7 4
> rdunif(10, 10, 1)
[1] 7 9 7 6 4 8 3 5 4 4
> |
```

And you can see here, this is the, these are the screenshot of the same observation, which I have just shown you. So, let us now try to come to the R console and try to do these things who are here, first, we need to install the package, although this package is already installed on my computer. So, I do not need to install it, but I need to upload it purrr. You can see here now this thing and now if I tried to generate here say 10 observation from a uniform distribution whose starting value is 1 and last value is 10.

So, for example, you can see here like this, so, you can see here you are getting here one, two, three, four, five, six, seven, eight, nine, ten values from the discrete uniform distribution and if you try to repeat this experiment, you can see here that these values are getting changed and

every time you are trying to repeat this you are getting here a different value. So, now, what I try to do here, I try to save these values and I have never and I try to find out the values of the mean and variance.

(Refer Slide Time: 17:31)



```
> x=rdunif(10, 100, 1)
> x
[1] 52 25 77 52 85 83 30 92 59 53
> mean(x)
[1] 60.8
> var(x)
[1] 529.2889
>
> x=rdunif(10, 100, 1)
> mean(x)
[1] 46.8
> var(x)
[1] 801.7333
> x=rdunif(10, 100, 1)
> mean(x)
[1] 63.9
> var(x)
[1] 538.9889
> |
```

So, what I try to do here that I try to generate the observation and then I try to find out their mean and variance. So, let x be equal to $rdunif$. And there we try to generate 10 observation from discrete uniform distribution whose minimum value is 1 and maximum value is 100. So, you can see here these values will look like this and if I try to find out here the mean of x , this will come out to be 60.8 and variance of x will come out to be here 529.2889 and you can see that these values are very different from the theoretical mean and theoretical variance.

And if you try to repeat it here, for example, I try to generate here one more set of observation and then I try to find out their mean and their variance, you can see here that these values are coming out to be very different, in the two cases, the mean and variance both are very much different.

And if you try to hear repeat it, these values will be changing and now, if I ask you what is the mean and what is the variance before I find that out, you cannot tell me, in the sense, this mean and variance they are the function of random variables. So, they are random and actually they are called as statistic that we will try to discuss later on. But you can see that here, what is the meaning of randomness and what do we really mean by sample dependence?

(Refer Slide Time: 19:11)

```
> x=rdunif(10000, 100, 1)
> mean(x)
[1] 50.1414
> var(x)
[1] 833.2211
>
> x=rdunif(10000, 100, 1)
> mean(x)
[1] 50.6026
> var(x)
[1] 842.836
>
> x=rdunif(10000, 100, 1)
> mean(x)
[1] 50.3227
> var(x)
[1] 837.4721
> |
```

Discrete Uniform Distribution: Example 2 in R

```
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.5494
> var(x)
[1] 838.4448
>
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.7379
> var(x)
[1] 833.3747
>
> x = rdunif(10000, 100, 1) # 10000 observations
> mean(x)
[1] 50.5403
> var(x)
[1] 824.5292
```

Now, what I try to do here I try to increase this number of observations. And I try to get here suppose here 10,000 observations and I try to find out their mean of x and variance of x, you can see here they are coming out to be they are now closer to the theoretical mean and theoretical variance and if I try to repeat and try to get 10,000 more observations, and I try to find out their here mean and variance, this is coming out to be pretty close to say 50 and 842.

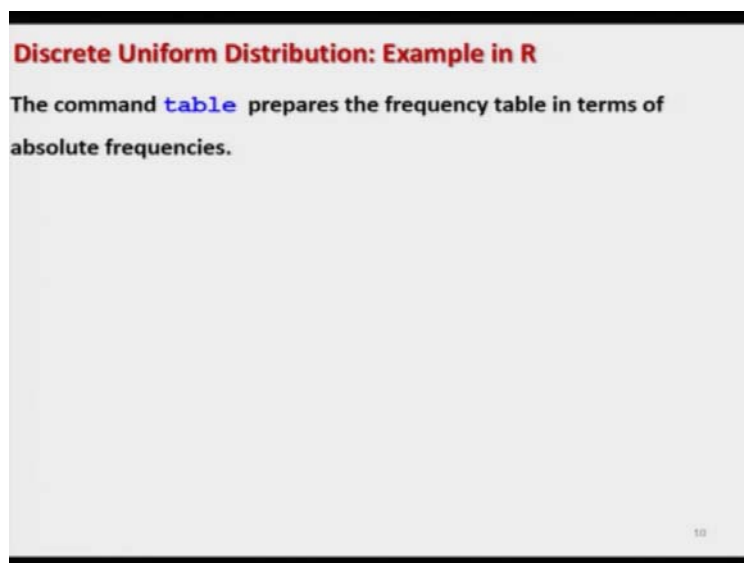
So, now I try to ask you one thing here more that suppose you have got some data and where you do not know the real mean and variance. Do not you think that if you try to use this mean and variance to get remind the theoretical values of the distribution possibly they are going to give

you the good value? Now, you can see in this case I am, when I am trying to do a simulation, you know the true values and now you are trying to find the same values by generating different types of samples.

For example, if you try to generate here one more sample here, you can see here that the values of mean and variance they are not very different. In this case, you can see here the value of mean is just 50.14, 50.60, 50.32 and the variance is 833.22, 842.8 and 837.4. So, means either you are assuming the variance to be 833 or 842 or 837, possibly, it will not make much difference. And definitely if you try to increase the sample size to a more and to a larger value, possibly, they will give you a fair idea about the true value of the theoretical mean and theoretical variance.

So, now, this is the first place where I am trying to give you such an idea that how things work in data science. So, now, let us try to take an example and try to see how the meaning of this probability mass function and its values is coming into picture. So, I will try to take the same example that I consider in the last lecture.

(Refer Slide Time: 21:27)



So, we come back to our slides. And now, actually, I am going to use here a command `table` that you know that the `table` is the command in R software, which prepare the frequency table in terms of absolute frequency and you can also compute the relative frequency and relative frequencies are a good estimate of probabilities that we already have discussed.

(Refer Slide Time: 21:46)

Discrete Uniform Distribution: Example 2 in R
Consider the rolling of a dice.

X : number of dots observed on the upper surface of the die
has a uniform discrete distribution with PMF

$$P(X = i) = \frac{1}{6}, \text{ for all } i = 1, 2, 3, 4, 5, 6.$$

The theoretical mean and variance of X are as follows:

$$E(X) = \frac{6+1}{2} = 3.5$$
$$\text{Var}(X) = \frac{6^2-1}{12} = \frac{35}{12} = 2.92$$

11

Now, you see, I try to take here the same example do you remember, we had considered this example in the last lecture, where we are trying to roll a dice and which has six possible points on the upper surface and the probability of each and every point is the same. The 1 by 6 and its mean and variance they are coming out to be 3.5 and 2.92, which are the theoretical mean and theoretical variance. Now, we try to simulate the same thing in the inside the R software.

Once these values are between 1, 2, 3, 4, 5 and 6, they have got an equal probability 1 by 6, do not you get the first idea from here that possibly in this case, discrete uniform distribution can be used, which is going to give you a good idea about the real phenomena? Yes, this is how we get such an information.

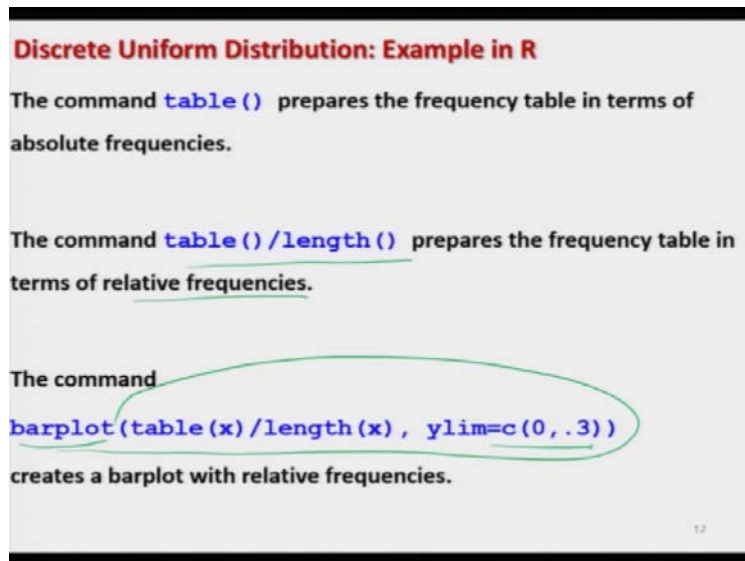
(Refer Slide Time: 22:44)

Discrete Uniform Distribution: Example in R

The command `table()` prepares the frequency table in terms of absolute frequencies.

The command `table()/length()` prepares the frequency table in terms of relative frequencies.

The command `barplot(table(x)/length(x), ylim=c(0, .3))` creates a barplot with relative frequencies.



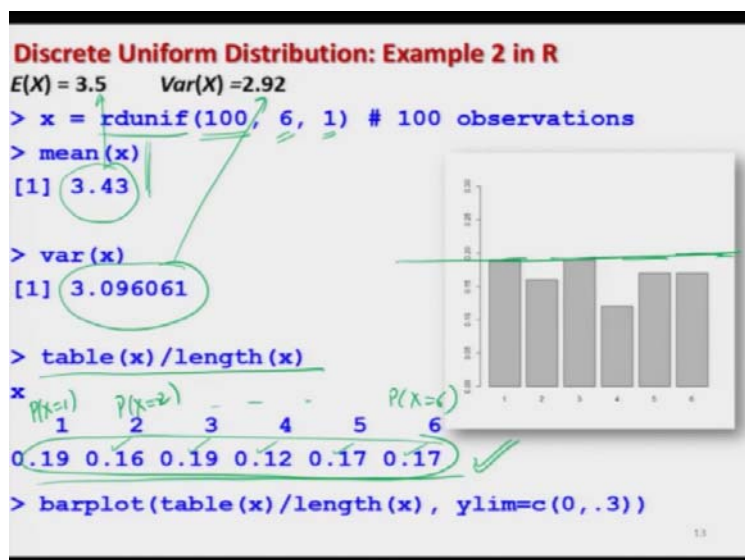
Now, mean say you know that this table command will prepare the frequency table in terms of absolute frequency and table divided by length that will prepare the frequency table in terms of relative frequency. And if you try to create the bar plot here of this tabulated values of the frequency table, then you have to use the command here bar plot. So, I am not going into that detail, because I believe that you know. So, now, I will try to create a here the bar plot where the limits on the Y are between 0 and 0.3.

(Refer Slide Time: 23:20)

Discrete Uniform Distribution: Example 2 in R

$E(X) = 3.5$ $Var(X) = 2.92$

```
> x = rdunif(100, 6, 1) # 100 observations
> mean(x)
[1] 3.43
> var(x)
[1] 3.096061
> table(x)/length(x)
x
 1  2  3  4  5  6
0.19 0.16 0.19 0.12 0.17 0.17
> barplot(table(x)/length(x), ylim=c(0, .3))
```



Outcome (x)	Relative Frequency
1	0.19
2	0.16
3	0.19
4	0.12
5	0.17
6	0.17

So, now, I try to simulate the same thing. So now I have an experiment where the dice is being rolled, and the probability of every point is 1 by 6. So, I can fit here a discrete uniform distribution. So, what I try to do here, I try to conduct the experiments say 100 time. So, and I get here the 100 values, which can be generated using the command `rdunif` with the range 1 to 6. And then I tried to find out here the value of the, mean of the these random observations, and this comes out to be 3.43.

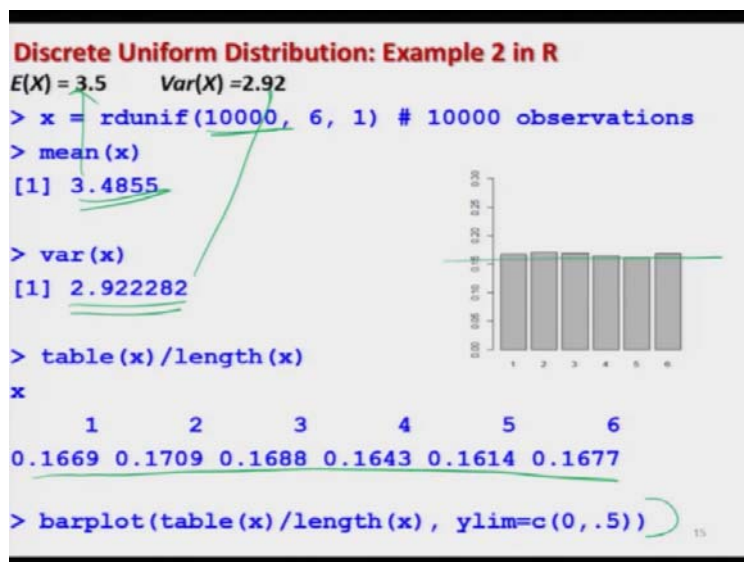
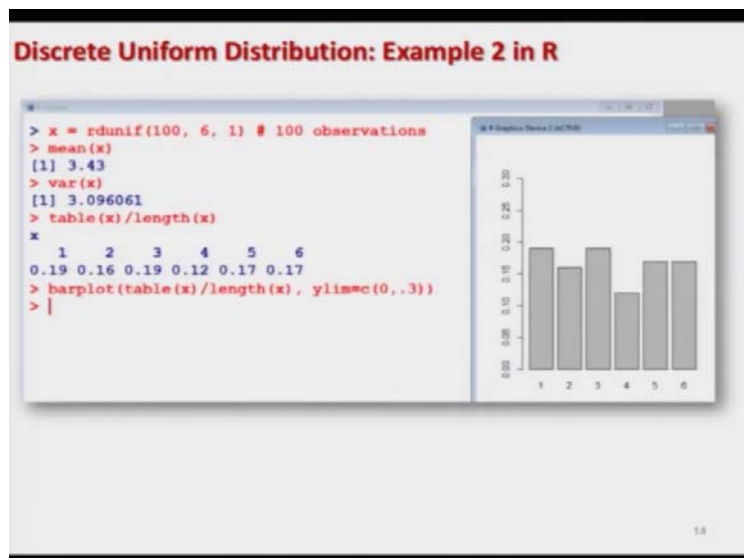
So, you can see here that is close to 3.5. And I tried to find out here the variance of x , that I just shown you this is coming out to be 3.09, which is quite close to 2.92. And then I try to prepare here the relative frequency table. By using this command `table` divided by `length`. Now tell me one thing, what is this giving you here? What are these things? It is 0.19, 0.16, 0.19, 0.12, 0.17, 0.17.

What are these thing? Do not you think that these are the probabilities of observing probability that x equal to 1 and probability of x equal to 2. And similarly here, probability of x equal to 6, but the difference is this, they are based on the sample values. And you can see here that theoretical values of the probabilities are 1 by 6, but here in this case, the values are not the same, but this is trying to imitate the real phenomena.

Now, the question is this that, why this difference is coming and how you can make it as close as possible this difference is coming because it is dependent on the sample size and your sample size here is just 100 and value which you are trying to compute for the mean and variance as theoretical mean and theoretical variance, they are based on the entire population. So, in statistics, our objective is this, that we want to know the value of the theoretical mean, theoretical variance and some other parameters, which are unknown to us based on the sample of data.

But here I want to show you that, what is the meaning of expected value of x and variance of x , and what is the meaning of these probabilities that are going to be determined by the probability mass function. So, you can see here these are the probabilities which are something like estimated probabilities of the probability mass function for different values of x equal to 1, 2, 3, 4, 5 and 6. And if you try to create here a bar plot you can see here that ideally these probabilities are not the same and ideally this uniform distribution should have all the probabilities which are at the same level, this is not happening.

(Refer Slide Time: 26:26)



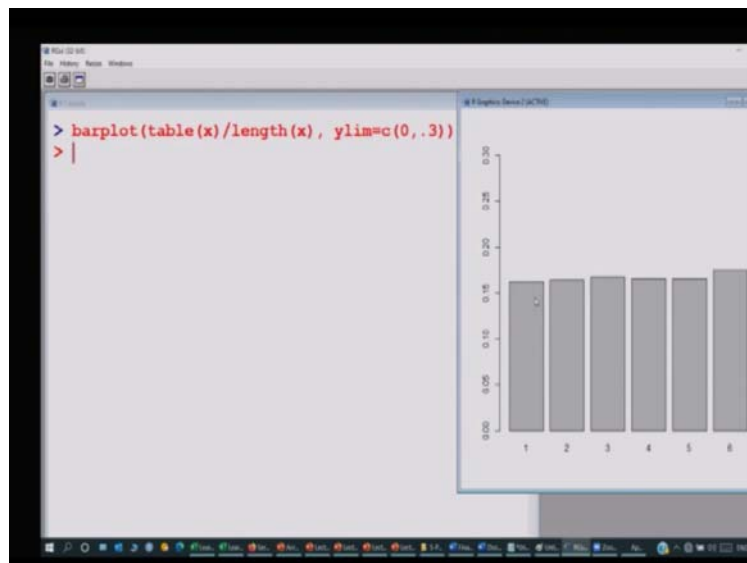
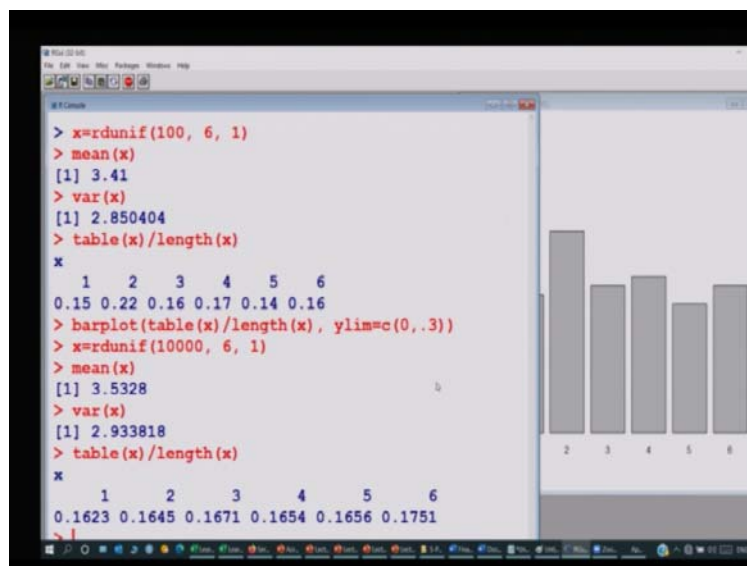
Now, if I try to repeat this experiment well, this is the screenshot and we try to now draw here 10,000 observation from the same example. Now, you can see here that the mean is coming out to be 3.48 and the variance is coming out to be 2.92, very close to the true values. And if you try to compute here the relative frequencies that will give you an idea about the theoretical probabilities you can see here that these values are coming out to be like 0.16, 0.17 and so, on and they are very close to each other.

And do not you think that this is going to indicate that your probability mass function is 1 upon 6 for each value of x and if you try to create here the bar plot using this command here, this will

look like this and you can see here that most of the probabilities are coming out to be the same. So, now, this is what I wanted to show you that how this probability mass functions and theoretical mean and theoretical variance they are going to represent a real life phenomenon.

So, you can see here, by this simulation experiment you can get more convinced that when you are trying to do such experiments in real life, possibly the results what you are going to get they will be close enough and then you can justify them using your theoretical knowledge and this is the here the screenshot.

(Refer Slide Time: 27:59)



Now, I try to show you these things on R console and then I will try to finish this lecture. So, now come we come to our R console. So, let me try to clear the earlier screen and now, I will try to generate here suppose here suppose here hundred observation from a discrete uniform distribution with the lower value as 1 and upper value as 6 and now, I try to find its mean and variance you can see it is here 3.41 and the variance here is 2.85 and if you try to repeat it, these values will be getting repeated.

So, but if you try to create here the relative frequency table, so, this will be here table of x divided by length of x , you can see here that these values are not coming out to be exactly equal to the means, 1 upon 6. And in case if you try to create here the bar plot of the same thing you can see here this is like this. So, there is a variation in the values of the heights of the bar which are indicating the relative frequencies which is indicating the estimated probabilities which are involved in the probability mass function.

And now, let us try to repeat this experiment say for say 10,000 time, you can see here you have got here the values of x you can get here the value of mean as 3.53 and variance of 2.93. So, you can see that here these values are pretty close to the theoretical values and if you try to find out here the relative frequencies, you can see here that you are getting here, the values here as 0.1623 and so on.

You can see that most of the values are very close to 1 by 6. And in case if you try to create here the, this bar plot for this one you can see here, now this bar plot is coming out to here like this where the probabilities are just very close to 1 upon 6. So, you can see here now that with this very small example, I have done two things. The first thing I have shown you that whatever was my theoretical probability function and there are complete theoretical setup, I have imitated that thing in a with the help of a data set.

When I try to write down a probability function, it can be discrete or continuous whatever you want. Now, you can see that they are going to generate value of a probability of an event which can really occur in real life. And if you try to count and try to find out these different values, you will have to work harder, but if you try to use the probability mass function, they know that if your phenomenon is like that, you can compute the probability from this function directly.

For example, if you want to know the mean value, variability, and similarly, if you wish to try to see for this skewness, kurtosis etc., whatever is the phenomenon, the properties of the phenomenon can be very, very well expressed and approximated by the probability mass function or the probability density function. But the condition is that you have to choose the right and correct form of the probability mass or the probability density function, that is the biggest challenge.

And second thing I have shown you that if you got a data, how you can infer different types of information from the data and data will give you an idea that what type of probability function can be used over here, up to now, you have done only two probability mass function, but there is a long list of the probability mass function and every probability mass function and every probability density function is going to indicate the probability for a certain type of or a special type of phenomenon.

So, one of the biggest challenge that comes in the data sciences is that, how to choose the right probability function, which is going to express my real phenomena in the correct way. And that is the challenge what we are going to do. So, in order to help you out I will try to take up some more probability function and similarly, on the same lengths I will try to show you their implications in the R software also, R software can help you in understanding the phenomena and imitating the phenomena inside a computer.

So, whenever you are trying to work with a real data, you can imitate that data inside the computer and can make different types of experiment to convince yourself that whatever you are going to do on the real set of data that is really going to give you the correct results or not. So, you try to practice it, try to understand what I have explained you and I will see you in the next lecture with a new distribution, new probability mass function. Till then, goodbye.