Essential of Data Science with R Software-1
Probability and Statistical Inference

Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture No. 12

Probability and Relative Frequency - An Example

Hello friends, welcome to the course Essentials of data science with R software-1. Where we are trying to understand the concepts of probability theory and statistical inference. And we already have begun learning of the concepts of probability theory. So, you can recall that in the last lecture I had given you or I introduced you with one of the basic fundamental definition of probability theory.

And I have tried my best to explain you what is the interpretation of that thing. You see whenever you are trying to work in data science there is difference when you are trying to work with the statistics. In statistics you can have a sample of size 5, 10, 15, 20 where you can see what is really happening. But in data sciences, the sample size is going to be very large. And you are going to execute different types of complicated operations where it may not be possible for you to observe the intermediate steps.

And that is why this understanding of the statistical concept is very important. Particularly, if what will happen when they are trying to be executed on a dataset on some numerical value. For example, in the last lecture if you remember I took an example very simple example of head and tails which are obtained by the toss of a coin. You have learnt right from the beginning that the probability of getting head or tail is just 1/2. But then, when I conducted a sort of simulation experiment in the R software where I generated say 100 values 1000 values and so on of head and tails which are indicated by 0 and 1, then you saw that when we are trying to compute the relative frequencies of head and tail they have not exactly coming out to be 0.5 always. And you have observed that this phenomenon is going to change when you try to increase the number of repetitions that is small n, the number of times you repeat the experiment. And you had observed that when you are trying to repeat the experiment for larger number of times the value of the relative frequency is converging towards a particular value. And you can see the fluctuations in the value of relative frequency they are very high, when you are trying to take a smaller sample size.

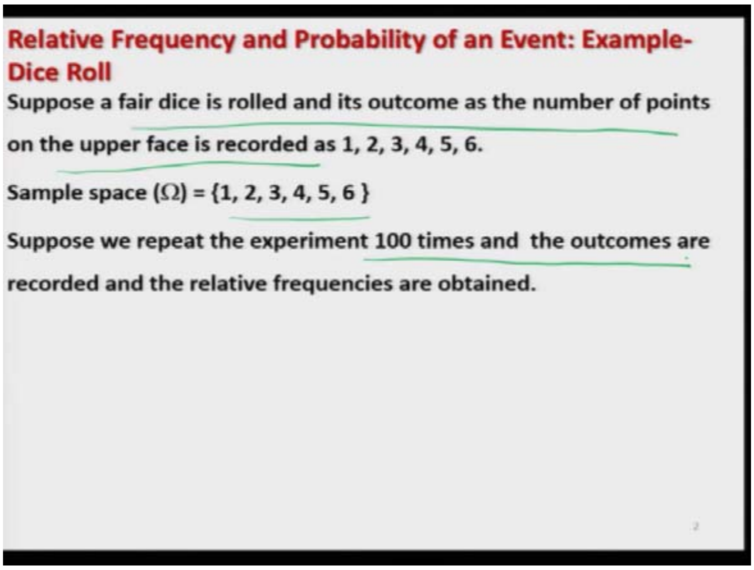
So, now this examples gives you a very good understanding that what it will, what will really happen when you try to compute a certain probability in larger dataset. That may also happen that if you try to increase the sample size also then also probability may converge to a reasonable value. But when you are trying to work with the finite sample very small sample, then this convergence may be little bit tricky.

So, now when you are trying to work in a data science and you really want to understand what will be the probability of certain event and suppose it is not really possible to compute the probability manually using some mathematical formula, exact formula. So, you will try to approximate it. So, this type of understanding will help you in taking a final call whether you are getting the correct value of probability or not.

And definitely, if you practice more you have more experience you will become a better data scientist. So, now in this lecture. I will try to take one more popular example and I will try to explain you that how this definition of relative frequency and probability are interrelated to each other. So, here in this lecture. I am going to take one example of rolling of a die. Where there will be six possible outcomes and I will try to increase the number of repetitions.

We know from the theory that the probability of observing 1, 2, 3, 4, 5, or 6 in a rolling of a die is just 1/6. But I will try to show you that how this 1 by 6 comes into picture. So, let us begin our lecture and try to see how it does and I will try show you in the R console also on the R software also.

(Refer Slide Time: 4:58)



Relative Frequency and Probability of an Event: Example- Dice Roll

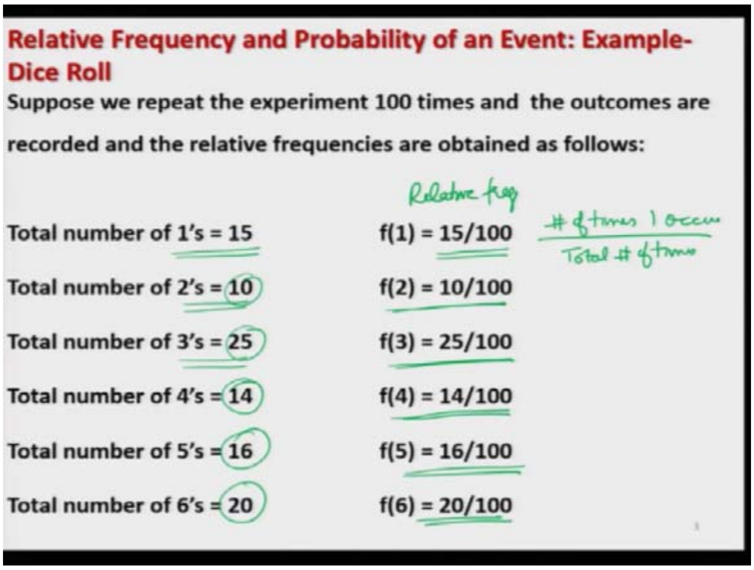
Suppose a fair dice is rolled and its outcome as the number of points on the upper face is recorded as 1, 2, 3, 4, 5, 6.

Sample space (Ω) = {1, 2, 3, 4, 5, 6 }

Suppose we repeat the experiment 100 times and the outcomes are recorded and the relative frequencies are obtained.

So, now let me take here a very simple example and suppose a fair dice is rolled and it outcome as the number of points on the upper face are recorded as 1, 2, 3, 4, 5, or 6. So, the sample space will consist of six possible points 1, 2, 3, 4, 5, or 6. And suppose we try to repeat the experiment say 100 times. And the outcomes are recorded and the relative frequencies are also obtained.

(Refer Slide Time: 5:27)



Relative Frequency and Probability of an Event: Example- Dice Roll

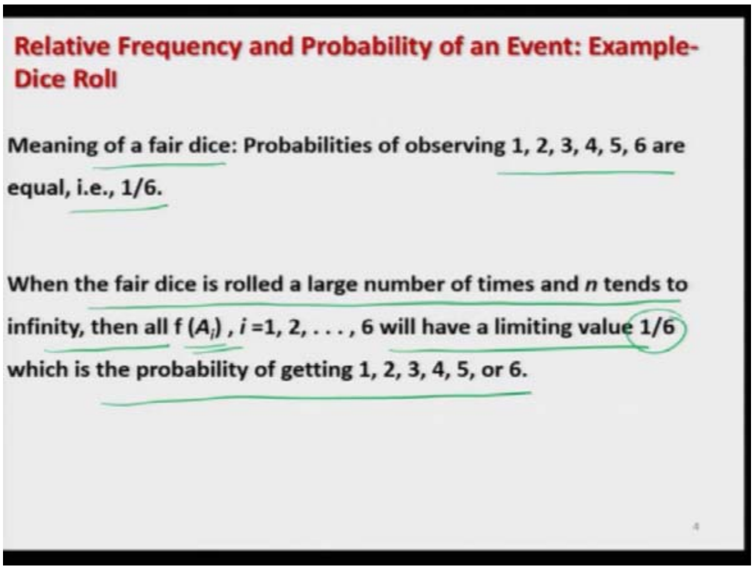
Suppose we repeat the experiment 100 times and the outcomes are recorded and the relative frequencies are obtained as follows:

	Relative freq	
Total number of 1's = 15	f(1) = 15/100	$\frac{\#\text{ of times 1 occur}}{\text{Total }\#\text{ of times}}$
Total number of 2's = 10	f(2) = 10/100	
Total number of 3's = 25	f(3) = 25/100	
Total number of 4's = 14	f(4) = 14/100	
Total number of 5's = 16	f(5) = 16/100	
Total number of 6's = 20	f(6) = 20/100	

And suppose we obtain the following observation. Well these are some hypothetical observations which I have created just to make you understand. Suppose we roll the dice for 100 times and out of 100 times we observe the number of times 1 occur is 15. Number of times 2 occur this is 10, number of time 3 occurs that is 25, number of times for 4 it is 14, and for 5 and 6 this are 16 and 20. Now based on this number of observation. We compute the relative frequency. So, the relative frequency is going to be computed by here like this.

Suppose I try to count the number of times 1 occurs and divided by total number of times. The dice is roll that is 100. So, this will become here 15/100. And similarly, for 2, 3, 4, 5, and 6 the relative frequencies are as follows for 2 the relative frequency is 10/100, for 3 it is 25/100, for 4 it is 14/100, for 5 it is 16/100, and for 6 it is 20/100. Now what do you observed here? When you try to learn the probability of observing 1, 2, 3, 4, 5, or 6. We have learned that this is just 1/6.

(Refer Slide Time: 6:57)



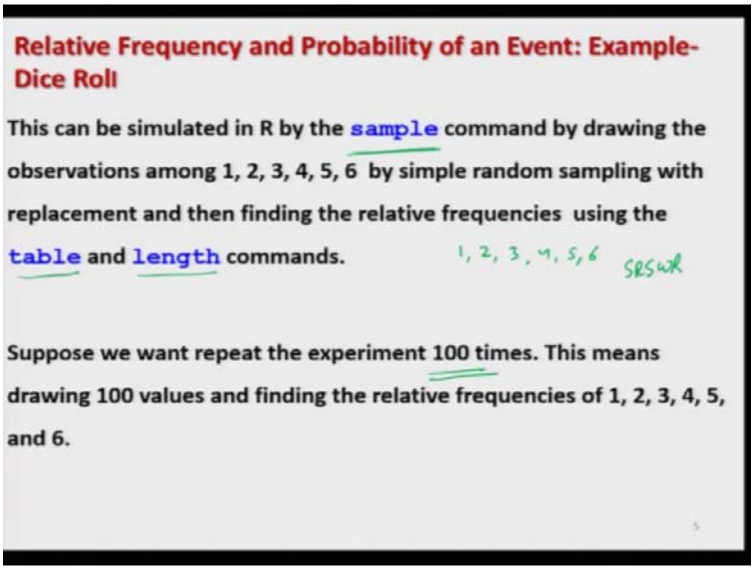
Relative Frequency and Probability of an Event: Example-Dice Roll

Meaning of a fair dice: Probabilities of observing 1, 2, 3, 4, 5, 6 are equal, i.e., 1/6.

When the fair dice is rolled a large number of times and *n* tends to infinity, then all f (A_i) , i =1, 2, . . . , 6 will have a limiting value 1/6 which is the probability of getting 1, 2, 3, 4, 5, or 6.

And this is really the meaning of a fair dice. That is the probabilities of observing 1, 2, 3, 4, 5, 6 are just 1/6. Well, we are assuming here that our dice is fair. So, but now we have understood that what is the meaning of this 1/6. That means, when the fair dice is roll for a large number of times and this number of times is here n. That goes to infinity, then all this relative frequencies f of A_1, A_2, A_3, A_4, A_5 and f of A_6 they will have a limiting value 1/6 and that will indicate the probability of getting 1, 2, 3, 4, 5, or 6. So, now it should be clear in your mind that what is the meaning of observing the probability 1/6 when you try to roll a fair dice.

(Refer Slide Time: 7:49)



Relative Frequency and Probability of an Event: Example-Dice Roll

This can be simulated in R by the `sample` command by drawing the observations among 1, 2, 3, 4, 5, 6 by simple random sampling with replacement and then finding the relative frequencies using the `table` and `length` commands.

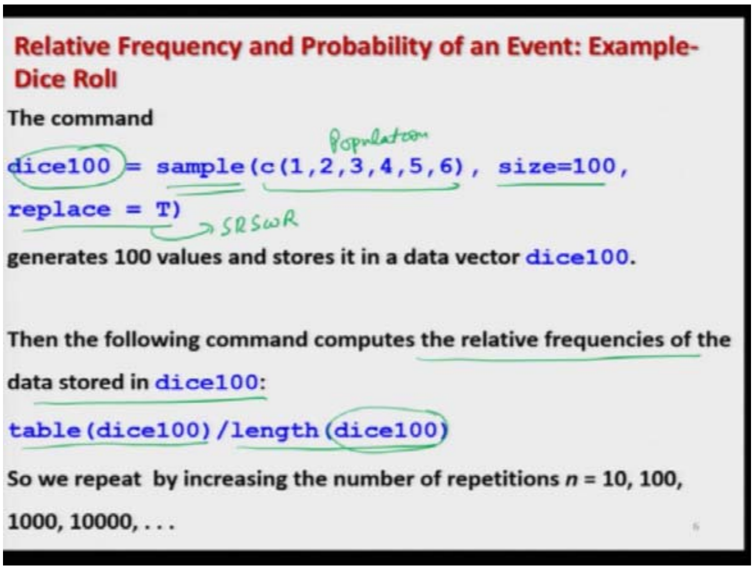
1, 2, 3, 4, 5, 6 SRSWR

Suppose we want repeat the experiment 100 times. This means drawing 100 values and finding the relative frequencies of 1, 2, 3, 4, 5, and 6.

So, now once again I use the same set up that we have use in the case of toss of a coin and we try to compute this probability and we try to observe the behavior that what happens when the number of times of repetitions becomes larger. So, once again we try to use here the sample command from the R software. So, what are we going to do we have here six values 1, 2, 3, 4, 5, and 6 and we will try to draw here a sample by simple random sampling with replacement. Because with replacement has an option that any of the number can occur again and again.

And then we try to count that from that sample how many times 1 occurs. How many times 2, 3, 4, 5, or 6 occur. And based on that we try to find out this relative frequency. In this case the command of mean may not work here because we are going to indicate they said numbers 1 to 6. So in order to find out the relative frequency that we know we have to use the command table and length. So, now I do one thing suppose I repeat the experiment 100 times. This means we have roll the dice 100 times and we have observed the 100 values which are consisting of the values 1, 2, 3, 4, 5, and 6.

(Refer Slide Time: 9:21)



Relative Frequency and Probability of an Event: Example-Dice Roll

The command

```
dice100 = sample(c(1,2,3,4,5,6), size=100,
replace = T)
```

Population
SRSWR

generates 100 values and stores it in a data vector dice100.

Then the following command computes the relative frequencies of the data stored in dice100:

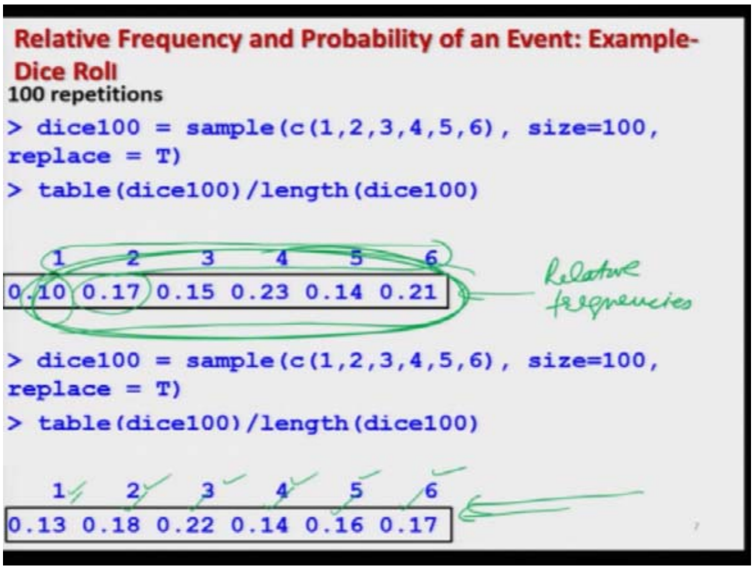
```
table(dice100)/length(dice100)
```

So we repeat by increasing the number of repetitions n = 10, 100, 1000, 10000, . . .

And from there we try to compute that how many times 1, 2, 3, 4, 5, or 6 occurring. So, for that I am using here a command say sample and sample from where this is my here population. Population is consisting of six values 1, 2, 3, 4, 5, and 6. And we want to draw here a sample of size 100. And then we are trying to draw it by simple random sampling with replacement for which I have to give you here replace it is equal to TRUE. So, this will generate 100 values and we are going to store them in the data vector whose name we have given as dice100. So, that will indicate that this is the dice experiment in which 100 values are obtained.

And then from this outcome, we will compute the relative frequencies of this data what we have generated in dice100. So, far that the command will be table dice100 inside the parentheses divided by length of dice 100. We have the dice100 variable is going to be written within the parentheses. So, and then we try to repeat it for say 100 times, 1000 times, 10000 times, and so on. And we try to observe the phenomenon. You see it is very important for you to observe.

(Refer Slide Time: 10:36)



Relative Frequency and Probability of an Event: Example-Dice Roll

100 repetitions

```
> dice100 = sample(c(1,2,3,4,5,6), size=100,
replace = T)
> table(dice100)/length(dice100)

   1    2    3    4    5    6
0.10 0.17 0.15 0.23 0.14 0.21
```

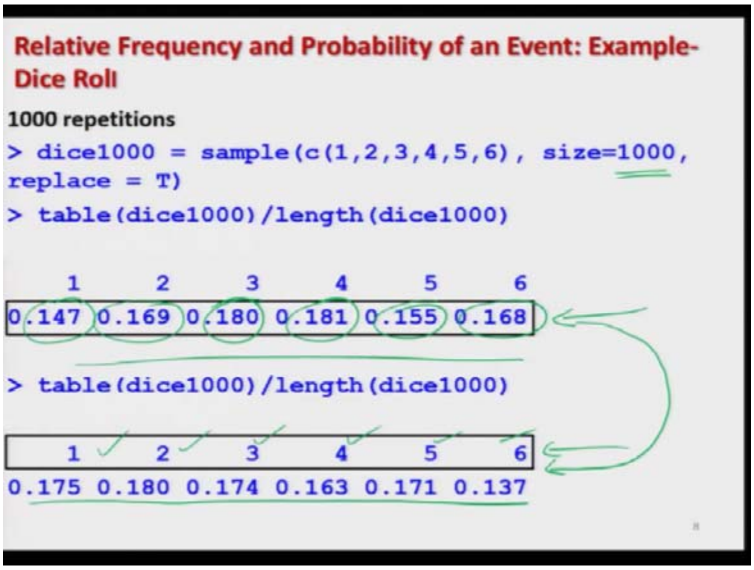
Relative frequencies

```
> dice100 = sample(c(1,2,3,4,5,6), size=100,
replace = T)
> table(dice100)/length(dice100)

   1    2    3    4    5    6
0.13 0.18 0.22 0.14 0.16 0.17
```

And as is said that you have to learn what the data is trying indicate you. So, suppose I execute these things on the R console. And here in this slide, I have given the outcome that I already had the conducted this experiment. But certainly I will try to show this thing on the R console to you also. So first we try to understand the phenomenon. So, I try to obtain here 100 values and you can see here we have the values 1, 2, 3, 4, 5, and 6. And these are relative frequencies. So, you can see here, that these relative frequencies which are the total number of times 1 occur divided by 100 this is here 0.10. That means, out of 100 times the number 1 occurs only 10 times.

And similarly number 2 occurs 17 times. So, that the relative frequencies 0.17 and so on. So, you can see here in these values none of the value is exactly equal to 1/6 and I try to repeat this experiment. The same experiment that I draw 100 values once again. And I try to compute their relative frequency outcome. You can see here, here the relative frequencies of 1, 2, 3, 4, 5, and 6. They are given here like this 0.13, 0.18, 0.22, 0.14, 0.16, and 0.17. Which are very different from these frequencies which we have obtain earlier. But none of the frequencies are indicating to get the probabilities are exactly equal to 1/6.

(Refer Slide Time: 12:14)



Relative Frequency and Probability of an Event: Example-
Dice Roll

1000 repetitions

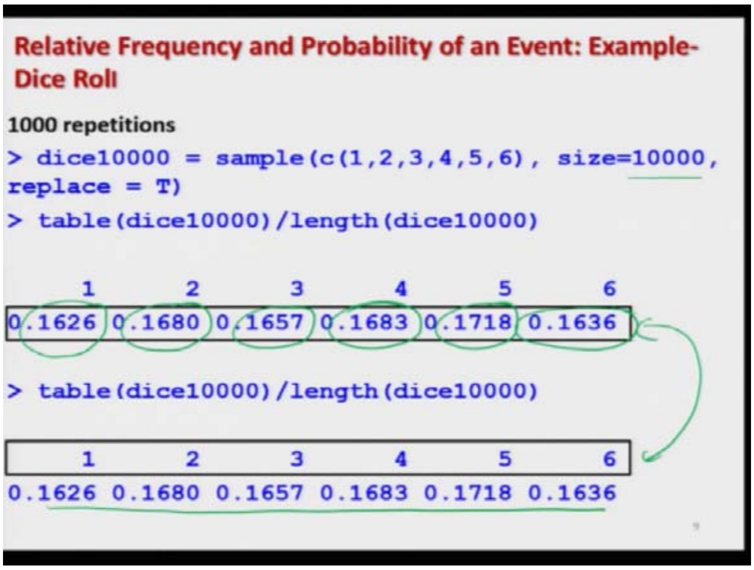
```
> dice1000 = sample(c(1,2,3,4,5,6), size=1000,
replace = T)
> table(dice1000)/length(dice1000)

    1     2     3     4     5     6
0.147 0.169 0.180 0.181 0.155 0.168

> table(dice1000)/length(dice1000)

    1     2     3     4     5     6
0.175 0.180 0.174 0.163 0.171 0.137
```

Now what I do I try to repeat this experiment for say 1000 times. And I try to compute the relative frequency of all the 1000 observations that I have generated from 1, 2, 3, 4, 5, and 6. And this is the relative frequency that we obtain, say out of 1000 observations 1 occurs 147 times, 2 occurs 169 times 169 times, 3 occurs 180 times, 4 occurs 181 times, 5 occurs 155 times, 6 occurs 168 times.

But you can see that the difference among the values of these probabilities is more closer than the values which are here when you try to repeat the experiment. So, do not you think that this is giving you an idea that values are converging towards some unknown value that we do not know. But definitely from our theory, we know that this value is ultimately going to be close to 1/6.

And if I try to repeat the same experiment here once again; then the relative frequency of 1, 2, 3, 4, 5, and 6. This is coming out to be here is 0.175, 0.180, 0.174, 0.163, 0.171, and 0.137. So, you can see here there are still fluctuations among the values. And there is also some difference between the values of the corresponding observations in the two repetitions also. But definitely, the variation is much smaller than when you try to repeat the experiment only 100 times.

(Refer Slide Time: 14:07)



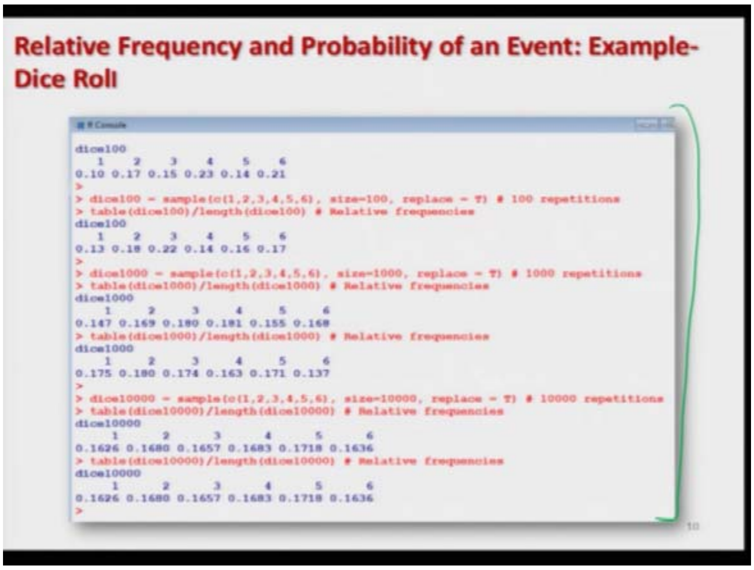
Relative Frequency and Probability of an Event: Example-Dice Roll

1000 repetitions

```
> dice10000 = sample(c(1,2,3,4,5,6), size=10000,
replace = T)
> table(dice10000)/length(dice10000)

     1      2      3      4      5      6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636

> table(dice10000)/length(dice10000)

     1      2      3      4      5      6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636
```

Similarly, if try to see here that in case if I try to repeat the experiment now 10000 times. So when I try to repeat the experiment for such a large number of times 10000. Then the number of 1, 2, 3, 4, 5, and 6 which are observe here they are following. That out of 10000 values we are getting here 1626 number of times 1, 1680 times 2, 1657 times 3, 1683 times 4, 1718 times 5 and 1636 times here 6. And similarly, when I try to repeat this experiment, I try to get here a different number but if you try to see you are getting here the same number here. But that is only by chance that may change.

(Refer Slide Time: 15:05)



Relative Frequency and Probability of an Event: Example-Dice Roll

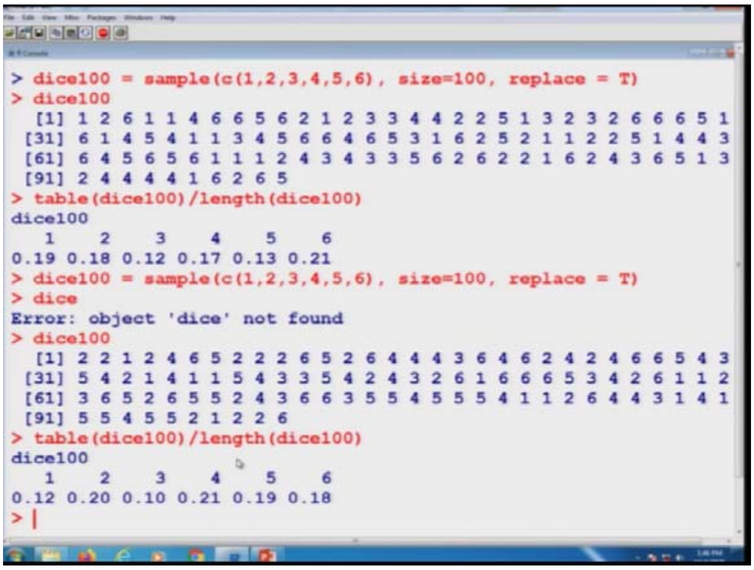
```
dice100
   1    2    3    4    5    6
0.10 0.17 0.15 0.23 0.14 0.21
>
> dice100 = sample(c(1,2,3,4,5,6), size=100, replace = T) # 100 repetitions
> table(dice100)/length(dice100) # Relative frequencies
dice100
   1    2    3    4    5    6
0.13 0.18 0.22 0.14 0.16 0.17
>
> dice1000 = sample(c(1,2,3,4,5,6), size=1000, replace = T) # 1000 repetitions
> table(dice1000)/length(dice1000) # Relative frequencies
dice1000
    1     2     3     4     5     6
0.147 0.169 0.180 0.181 0.155 0.168
> table(dice1000)/length(dice1000) # Relative frequencies
dice1000
    1     2     3     4     5     6
0.175 0.180 0.174 0.163 0.171 0.137
>
> dice10000 = sample(c(1,2,3,4,5,6), size=10000, replace = T) # 10000 repetitions
> table(dice10000)/length(dice10000) # Relative frequencies
dice10000
     1      2      3      4      5      6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636
> table(dice10000)/length(dice10000) # Relative frequencies
dice10000
     1      2      3      4      5      6
0.1626 0.1680 0.1657 0.1683 0.1718 0.1636
>
```

And definitely whatever I am trying to show here on this the screenshot. Which is indicating that I have really conducted this experiment which I am trying to report here that whatever the outcome which you can see on this screen they may not really come once again when you are trying to do here live. And these observation may change when you are trying to conduct yourself but you can see here when these things are getting very close to those values.

So, at 10000 of number of times I can see that these values are very close to 0.16. So, this will indicate that possible the probabilities of observing of 1, 2, 3, 4, 5, or 6 they are very close to 0.16 or 0.17 something like this in some of the things some of the cases. But definitely if you try to increase it beyond 10000 times, then definitely this the difference among the values will become very less. And what we expect that as we try to repeat it infinity number of times all the values of these relative frequencies will become the same.

But in practice what is infinity, this is very difficult to define. So, but if are all practical purposes we assume that the number of repetitions are pretty large. That is what we mean when we try to say limit n tending to infinity. Now just come to the R console and I try to show you these things that how they are going to work here.

(Refer Slide Time: 16:45)



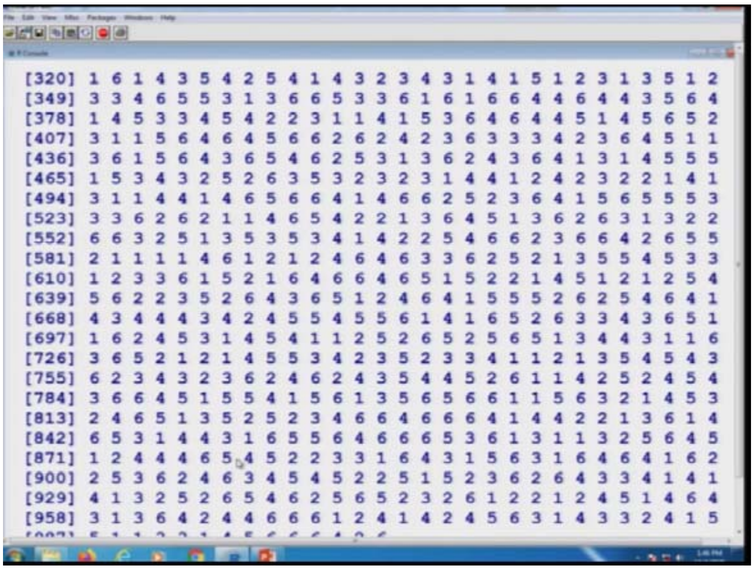
```
> dice100 = sample(c(1,2,3,4,5,6), size=100, replace = T)
> dice100
  [1] 1 2 6 1 1 4 6 6 5 6 2 1 2 3 3 4 4 2 2 5 1 3 2 3 2 6 6 6 5 1
 [31] 6 1 4 5 4 1 1 3 4 5 6 6 4 6 5 3 1 6 2 5 2 1 1 2 2 5 1 4 4 3
 [61] 6 4 5 6 5 6 1 1 1 2 4 3 4 3 3 5 6 2 6 2 2 1 6 2 4 3 6 5 1 3
 [91] 2 4 4 4 4 1 6 2 6 5
> table(dice100)/length(dice100)
dice100
   1    2    3    4    5    6 
0.19 0.18 0.12 0.17 0.13 0.21 
> dice100 = sample(c(1,2,3,4,5,6), size=100, replace = T)
> dice
Error: object 'dice' not found
> dice100
  [1] 2 2 1 2 4 6 5 2 2 2 6 5 2 6 4 4 4 3 6 4 6 2 4 2 4 6 6 5 4 3
 [31] 5 4 2 1 4 1 1 5 4 3 3 5 4 2 4 3 2 6 1 6 6 6 5 3 4 2 6 1 1 2
 [61] 3 6 5 2 6 5 5 2 4 3 6 6 3 5 5 4 5 5 5 4 1 1 2 6 4 4 3 1 4 1
 [91] 5 5 4 5 5 2 1 2 2 6
> table(dice100)/length(dice100)
dice100
   1    2    3    4    5    6 
0.12 0.20 0.10 0.21 0.19 0.18 
> |
```

So, let me try to just copy this command so that I can save some time. So, you can see here that the outcome of this dice 100 is like this. So, what is happening the sample command has sample 100 values from these 1, 2, 3, 4, 5, and 6 by simple random sampling with replacement. But it is giving us an impression as if the experiment has been conducted 100 times. And now from here we simply try to compute the relative frequency of this event.

So, you can see here, the relative frequency of observing 1, 2, 3, 4, 5, 6 is coming out to be like as 0.19, 0.18, 0.12, 0.17, 0.13, 0.21 respectively, but they are not the same and if you try to repeat this experiment by repeating the experiment 100 times once again. So, this data you can see here this is going to be different than the data what we have obtain earlier. You can see here that this data and this data they are different.

And if you try to compute the relative frequency from this data, this is coming out to be like this 0.12, 0.20 and so on. And it is very different from the relative frequency that you have obtained earlier. So you can see here when your sample size is just 100, there is a huge variation in the relative frequencies of observing 1, 2, 3, 4, 5, and 6. Well, I am sure that if you try to repeat this experiment only for size is equal to 10. That means, they are only 10 observation the difference will become actually more. But I have intentionally taken it to be 100 because there are 6 possible values which are occurring.

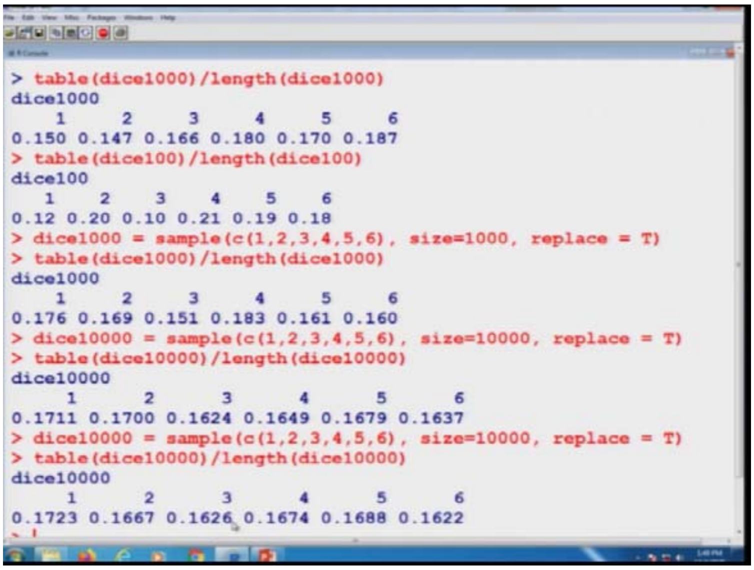
(Refer Slide Time: 18:37)



```
[320] 1 6 1 4 3 5 4 2 5 4 1 4 3 2 3 4 3 1 4 1 5 1 2 3 1 3 5 1 2
[349] 3 3 4 6 5 5 3 1 3 6 6 5 3 3 6 1 6 1 6 6 4 4 6 4 4 3 5 6 4
[378] 1 4 5 3 3 4 5 4 2 2 3 1 1 4 1 5 3 6 4 6 4 4 5 1 4 5 6 5 2
[407] 3 1 1 5 6 4 6 4 5 6 6 2 6 2 4 2 3 6 3 3 3 4 2 3 6 4 5 1 1
[436] 3 6 1 5 6 4 3 6 5 4 6 2 5 3 1 3 6 2 4 3 6 4 1 3 1 4 5 5 5
[465] 1 5 3 4 3 2 5 2 6 3 5 3 2 3 2 3 1 4 4 1 2 4 2 3 2 2 1 4 1
[494] 3 1 1 4 4 1 4 6 5 6 6 4 1 4 6 6 2 5 2 3 6 4 1 5 6 5 5 5 3
[523] 3 3 6 2 6 2 1 1 4 6 5 4 2 2 1 3 6 4 5 1 3 6 2 6 3 1 3 2 2
[552] 6 6 3 2 5 1 3 5 3 5 3 4 1 4 2 2 5 4 6 6 2 3 6 6 4 2 6 5 5
[581] 2 1 1 1 1 4 6 1 2 1 2 4 6 4 6 3 3 6 2 5 2 1 3 5 5 4 5 3 3
[610] 1 2 3 3 6 1 5 2 1 6 4 6 6 4 6 5 1 5 2 2 1 4 5 1 2 1 2 5 4
[639] 5 6 2 2 3 5 2 6 4 3 6 5 1 2 4 6 4 1 5 5 5 2 6 2 5 4 6 4 1
[668] 4 3 4 4 4 3 4 2 4 5 5 4 5 5 6 1 4 1 6 5 2 6 3 3 4 3 6 5 1
[697] 1 6 2 4 5 3 1 4 5 4 1 1 2 5 2 6 5 2 5 6 5 1 3 4 4 3 1 1 6
[726] 3 6 5 2 1 2 1 4 5 5 3 4 2 3 5 2 3 3 4 1 1 2 1 3 5 4 5 4 3
[755] 6 2 3 4 3 2 3 6 2 4 6 2 4 3 5 4 4 5 2 6 1 1 4 2 5 2 4 5 4
[784] 3 6 6 4 5 1 5 5 4 1 5 6 1 3 5 6 5 6 6 1 1 5 6 3 2 1 4 5 3
[813] 2 4 6 5 1 3 5 2 5 2 3 4 6 6 4 6 6 6 4 1 4 4 2 2 1 3 6 1 4
[842] 6 5 3 1 4 4 3 1 6 5 5 6 4 6 6 6 5 3 6 1 3 1 1 3 2 5 6 4 5
[871] 1 2 4 4 4 6 5 4 5 2 2 3 3 1 6 4 3 1 5 6 3 1 6 4 6 4 1 6 2
[900] 2 5 3 6 2 4 6 3 4 5 4 5 2 2 5 1 5 2 3 6 2 6 4 3 3 4 1 4 1
[929] 4 1 3 2 5 2 6 5 4 6 2 5 6 5 2 3 2 6 1 2 2 1 2 4 5 1 4 6 4
[958] 3 1 3 6 4 2 4 4 6 6 6 1 2 4 1 4 2 4 5 6 3 1 4 3 3 2 4 1 5
```

So, now let me try to generate here 1000 observations and I try to compute the relative frequency. So you can see here, these are the 1000 observations which are generated like this.

(Refer Slide Time: 18:53)



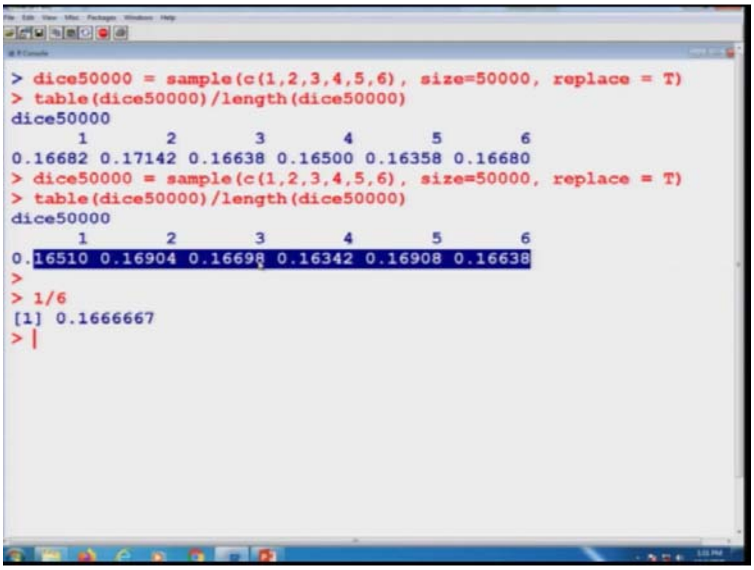
```
> table(dice1000)/length(dice1000)
dice1000
    1     2     3     4     5     6 
0.150 0.147 0.166 0.180 0.170 0.187 
> table(dice100)/length(dice100)
dice100
   1    2    3    4    5    6 
0.12 0.20 0.10 0.21 0.19 0.18 
> dice1000 = sample(c(1,2,3,4,5,6), size=1000, replace = T)
> table(dice1000)/length(dice1000)
dice1000
    1     2     3     4     5     6 
0.176 0.169 0.151 0.183 0.161 0.160 
> dice10000 = sample(c(1,2,3,4,5,6), size=10000, replace = T)
> table(dice10000)/length(dice10000)
dice10000
     1      2      3      4      5      6 
0.1711 0.1700 0.1624 0.1649 0.1679 0.1637 
> dice10000 = sample(c(1,2,3,4,5,6), size=10000, replace = T)
> table(dice10000)/length(dice10000)
dice10000
     1      2      3      4      5      6 
0.1723 0.1667 0.1626 0.1674 0.1688 0.1622 
> 
```

So, anyway I am not bothered about these things. But let me try to compute the relative frequency of this software 1000 observations. So, you can see here now this is like 0.15, 0.147, 0.166 and so on. So, you can see here now the variation between this 1000 observations and this 100 observations, this has become less smaller with 100 observations you have these outcomes and these were the relative frequency. Which you had finally obtain at the end.

You can compare these values and these values they are quite different. And if you try to repeat this experiment with 1000 values once again, then you can compute the relative frequencies once again and you can see here earlier the values were like this 0.15, 0.147 et cetera. But now these values are 0.176, 0.169 et cetera and these two are different. But definitely the difference among the relative frequencies or 1, 2, 3, 4, 5, and 6 is much much smaller than the difference in the relative frequencies of the case when we has only 100 observations.

Now let us try to increase the this number of observations and suppose I make it here 10000. And I try to compute here the relative frequency with 10000 observations. You can see here this values coming out to be very close to 0.17, 0.17, 0.1624 and so on. So, you can see here that this values of relative frequencies are pretty close to 0.16, 0.17 and so on. And if you try to repeat the same experiment once again the relative frequencies are going to be like this 0.17, 0.16, 0.16.

(Refer Slide Time: 20:57)



```
> dice50000 = sample(c(1,2,3,4,5,6), size=50000, replace = T)
> table(dice50000)/length(dice50000)
dice50000
      1       2       3       4       5       6 
0.16682 0.17142 0.16638 0.16500 0.16358 0.16680 
> dice50000 = sample(c(1,2,3,4,5,6), size=50000, replace = T)
> table(dice50000)/length(dice50000)
dice50000
      1       2       3       4       5       6 
0.16510 0.16904 0.16698 0.16342 0.16908 0.16638
> 
> 1/6
[1] 0.1666667
> |
```

And if you try to repeat it here once again for say larger number times suppose I try to make it for 50000 observations. Means computationally you can increase the number of repetitions as much as you want depending on the capability of your computer. So, if you try to see here this is coming out to be like this. Most of the values are very close to 0.1666. And if repeat this experiment once again the relative frequencies are coming out to be very close to 0.16.

And if you try to see the value of 1/6, say actually 0.1666. So, now you can see here when you are trying to repeat the, this experiment for 50000 time most of the values are coming out to be very close to 0.16. And this what we meant when we say that limit n tending to infinity n(A)/n this is called as the probability of the event A. So, for example this value 0.16510 this is the relative frequency of observing 1 and means we have computed it on the basis of 50000 observation which is quite large.

So, you can see here that this value is stabilizing as the number of times the experiment is repeated is increasing. And so one can belief that these values are going to be the real probabilities of observing 1, 2, 3, 4, 5, or 6 when we are trying to roll a dice. Do not you think that this is very convincing? So, now we come to an end to this lecture. And I belief that was a very interesting lecture in the sense that up to now whatever you have studied that probability observing head or tail is 1/2 or the probability of observing say 1, 2, 3, 4, 5, 6 and rolling of a dice is 1/6. What does this mean?

And this really happens in real life that whenever you are trying to find out the probability of any complicated phenomenon. Many times it become difficult to compute the probability mathematically. Just like so easily you compute the, computed the probability of head or tail or probability 1, 2, 3, 4, 5, or 6 it may really be possible. But you can program the phenomenon very easily. That depends on your capability of programming. Means you have to program the phenomenon in such a way such that it resembles with the true phenomenon. And based on that you try to compute this probability.

For example, if there are multiple outcomes means I can use a condition. In case if the outcome is coming out to be like this, then the indicator value or the indicator variable takes the value 1. If that is happening, then the indicator variable takes the value 2. If this is happening, then the indicator value or the indicator variable takes the value 3.

Now I can collect all the values of the indicator variables 1, 2, 3 and I can simply compute the relative frequency of the values of this indicator variable that will give you simple the value of the probability. And for these things, you may not really need the real computation from the Algebra point of view. It may be difficult, it may be complicated. I am not saying impossible because that depends on your capability to compute the probability.

But the question is that suppose I do not know the value of probability of tossing a coin or rolling a die. Suppose I do not know what is the probability of observing 4 or 5. When we are trying to roll a dice. Suppose I just cannot compute it mathematically. But I can program this phenomenon as I have done it. And then I can repeat this experiment for a very large number of time. And suppose if the real value is coming out to, of the probabilities coming out to be 0.1666 and so on.

But if the simulation is giving me the value 0.165 or 160 or say 0.171 etc. is not bad. But definitely it depends on the complicacy of the program the computation how difficult or how it is computationally heavier. The program can be computationally heavy also. That takes a very long time to execute. So, all these things depends on the capability of the computer and how you program it.

But definitely if you are getting an good approximate value of the probability do not you think that we accept it happily. Yes, this is what is happening in data science. So, with this conclusion I will request you to think about this process, think about this phenomenon and one thing you have to understand, can you really depend on the value of 0.1666 as the probability of observing 2 in the rolling of the dice? If you do not know the theory behind it whether you want to compute the relative frequency or the absolute frequency or something else. You will not be confident that whether what you are computing is really the probability or something else.

And now if somebody ask you that I have repeated the experiment only 20 times and you have repeated the experiment 20000 times, whose value is better value or whose value is more dependable value? How will you argue it? So, this statistical theory will give you an idea that how you have to argue. You can say well the classical definition of probability say that n tending to infinity n(A)/n will converge to a value and this value is going to be the probability of that event.

So, in case if my value is based on 20000 observation. I do not know what is the correct value. But definitely this is more dependable than your value which is dependent only on say 20 observation or 20 repetitions. So, this is how you will get confidence. Well, if somebody is making the repetition 20 million times possibly that value will be more dependable. But anyway you have to compromise that how much computation you can effort. And how much time takes for the computation.

So, this is how the theory of probability will help you in finding or the probability through computation in a data science. So, you try to think about it, try to practice try to take very simple example and try experiment them on the software. So, you practice it and I will see you in the next lecture with more topics on probability theory till then goodbye.