

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. 11
Relative Frequency and Probability

Hello friends, welcome to the course Essentials of data science with R software-1 in which we are going to handle the topics of probability theory and statistical inference. So, you can recall that in the earlier lecture I have given you the basic elements which are required to define the probability of an event. Now, in this lecture I will take up the issue and based on those concepts I will try to define the probability.

Probability is such a things which you are learning right from an elementary classes. Means elementary class means like 10 or class 11, 12 etc. One of the basic definition which you have learnt in probability theory is something like m upon n type of definition. That you try to count the total number of favorable cases and you try to divide it by the total number of possible cases.

But now in this definition I personally belief that there is a problem in understanding it. Sometime people ask us that the probability of getting a head or a tail in the toss of coin is $1/2$. What does this mean? In case if I try to toss the coin two times, do you expect that one time I will get head and one time I will get tail? Or in case if I roll a dice where the point 1, 2, 3, 4, 5, and 6 may come, they ask us what do you think if I try to roll a die for the 6 times then you believe that one time it will be 1 then 2, 3, 4, 5, and 6 and there will occur only once. And if you try to say it does not really happen in practice. Even if you try to toss a coin say four times, there is a possibility that there may be 3 heads or 1 tail, there may be 1 head and 3 tails, there may be 2 heads and 2 tails also. So, what does this mean? You are saying that the probability is $1/2$, tut when you are trying to conduct this experiment the total number of favorable cases divided by the total number of cases that is not exactly coming out to be always $1/2$.

So, now this the question which I am trying explain you here. Well, I am not that this definition is wrong. This definition is correct but possibly it requires a better understanding. So, what I will try to do in this lecture, that I will try to introduce these concepts and I will try to explain you that how you have to interpreted. And well it is not possible here to toss a coin say 500 times or

1000 times. But definitely I will try to take the help of R software to show you that when you are trying to compute such probabilities or when you try to say that the probability of occurring of a head or tail is $1/2$. What does this really actually mean?

So, let us begin our lecture and try to understand the probability theory from the relative frequency point of view. Now the question is what is relative frequency that you know. Relative frequency is simply if you try to repeat an experiment for some number of times, you try to see that how many times an event is occurring. For example, if you try roll a dice say this 100 times and you try to count the number of times the number 1 occurring, the number 2 is occurring, 3, 4, 5, and 6 are occurring and you try to divide it by 100. So, that is the your actually a relative frequency.

So, now I will try to connect this relative frequency with the probability theory and I will try to give you a better interpretation that how you should interpret the probability. So, let us begin our lecture.

(Refer Slide Time: 4:30)

Relative Frequency and Probability of an Event:

There is a close connection between the relative frequency and the probability of an event.

A random experiment is described by its possible outcomes.

For example, getting a number between 1 and 6 when rolling a die.

Suppose an experiment has m possible outcomes (events)

A_1, A_2, \dots, A_m and the experiment is repeated n times.

Handwritten notes: $1, 2, 3, 4, 5, 6$ above $A_1, A_2, A_3, A_4, A_5, A_6$; arrow pointing to 100 times, $m=6$; $n=100$.

So, now one thing I can tell you here that there is a close connection between the relative frequency and the probability of an event. We know that whenever we are conducting an experiment, we really mean that our experiment is actually random. Random means because for example if you are tossing a coin then yes you know the possible outcomes that either there will be head or tail. But unless and until you have toss the coin, you do not know whether you are

really go to head or a tail. From that point of view we call this experiment as a random experiment.

So, this random experiment is described by its possible outcomes. For example, when you are trying to roll a die. That means say you will getting a number between 1 to 6. It means 1, 2, 3, 4, 5, or 6. And similarly, suppose there is an experiment which has not only two outcome m possible outcomes. So, obviously these outcomes are going to be call as events. So, let this A_1, A_2, \dots, A_m they are our events and suppose the experiment is repeated.

For example, we assume that suppose the experiment is repeated small n number of times. What does this mean? For example, when you are trying to roll a die then there are possibilities that you may get 1, 2, 3, 4, 5, or 6. So these are your here 6 events $A_1, A_2, A_3, A_4, A_5,$ and A_6 . Now you try to repeat this experiment say suppose I roll a die say 100 times. So in this case you m is going to be 6 and n is going to be 100. So, this is what I mean when I say that the experiment is repeated small n number of times.

(Refer Slide Time: 6:15)

Relative Frequency and Probability of an Event:

Now we can count how many times each of the possible outcome has occurred.

In other words, we can calculate the absolute frequency $n_i = n(A_i)$ which is equal to the number of times an event $A_i, i = 1, 2, \dots, m,$ occurs.

The relative frequency $f_i = f(A_i)$ of a random event $A_i,$ with n repetitions of the experiment, is calculated as

$$f_i = f(A_i) = \frac{n_i}{n}$$

Handwritten notes on the slide include: "H H T H T # of times $n=5$ ", " $n_1=3=n(H)$ # of times $T=2$ ", " $n_2=2=n(T)$ ", "rel. freq of H = $\frac{3}{5}$ ", "rel. freq of T = $\frac{2}{5}$ ", " f_i → function", " $n_i =$ Absolute freq.", and " $n =$ Total # of repetitions".

Now once you have repeated the experiment you can count that how many times each of the possible outcome has occurred. And in case if you try to think in terms of the tools of descriptive statistics, then we are essentially trying to find out the absolute and relative frequency of this experiment or the outcome of this events. So, we simply try to calculate the absolute frequencies. Means absolute frequencies means suppose if you try to toss a coin. Say 5 times you get here

head, head, tail, head and here tail. So, now the number of times head is occurring this is here 3. So, this is the absolute frequency of the occurrence of head.

And the number of times the tail is occurring, this is here 2. So, the absolute frequency of occurrence of tail here is 2. So this is what we mean by absolute frequency. So, this is we I am trying to denote here is n_i which is a function of here A_i . So, for example in this case I can write down here n_1 is equal to 3 and it is function of here H and similarly in this case I write down here, n_2 is equal to 2 which is a function of here tail. And so this n_{A_i} is going to indicate the number of times an event A_i occurs out of this event A_2, \dots, A_m .

Now from this definition I can also find the relative frequency. For example, in this example if I try to find out the relative frequency of occurring of head. So, this is going to be number of time the head is occurring divided by total number of times the event has been repeated 5 times. And similarly the relative frequency of tail is going to be 2 upon 5. That is the number of times the event happens that is the 2 number of times we are getting a tail. And divided by the total number of times the experiment is repeated which is here 5.

So, similarly, I can make it more general and I can define the relative frequency which is denoted here as a f_i . You can see here I am using here two here f. One is here written italics f_i and one here is here f which means here function. So, this f_i is a function of A_i and A_i is random event and we have repeated the whole experiment small n number of times. So, in this case this relative frequency of A_i is computed as say n_i upon small n. So, n_i is the number of times the event occurs. So, n_i is the or I can say here this is simply your here absolute frequency.

And n is the total number of repetitions. So, this is going to be denoted by say f_i . So, now in case if an experiment is consisting of the events A_1, A_2, \dots, A_m so, we can compute here something like here f_1, f_2, \dots, f_m . So, this is how we try to compute the absolute and relative frequencies.

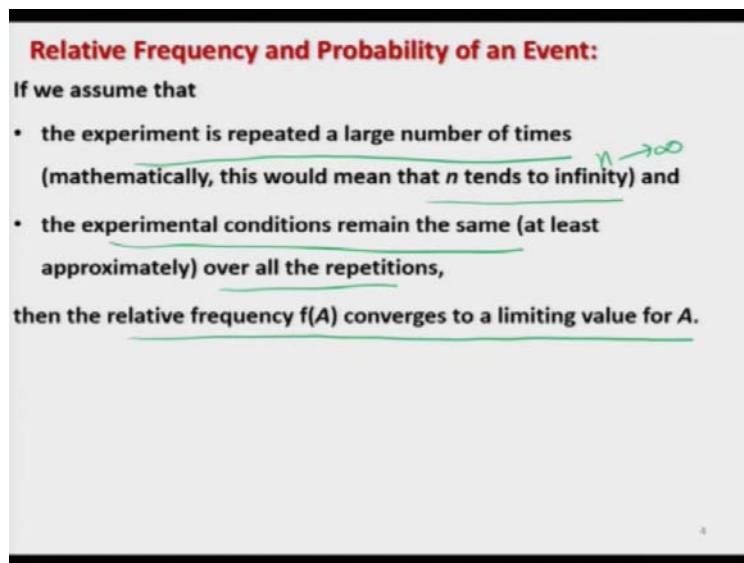
(Refer Slide Time: 9:52)

Relative Frequency and Probability of an Event:

If we assume that

- the experiment is repeated a large number of times
(mathematically, this would mean that n tends to infinity) and
- the experimental conditions remain the same (at least approximately) over all the repetitions,

then the relative frequency $f(A)$ converges to a limiting value for A .



Now in case if you assume that the experiment is repeated for a large number of times. And if you say the same thing in terms of mathematical language, we say that the experiment is repeated infinite number of times or we say that as n goes to infinity. And the second condition is this, this experiment is being repeated under the same conditions. That mean the experimental conditions remains the same. At least approximately I would say, whenever we trying to repeat the experiment. So. this experimental condition remains the same over all the repetitions. In that case we say that the relative frequency converges to a limiting value for A . So, if I try to take an event here A then the relative frequency of this event will be converging to a value A .

(Refer Slide Time: 10:51)

Relative Frequency and Probability of an Event:

This limiting value is interpreted as the probability of A and denoted by

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}$$

where $n(A)$ denotes the number of times an event A occurs out of n times.

$\frac{n(A)}{n} = \frac{m}{n} \rightarrow P(H) = \frac{1}{2} \quad P(T) = \frac{1}{2}$

$P(H) = \frac{\# \text{ of } H\text{'s}}{\# \text{ of Times the expt is repeated as } n \rightarrow \infty} \rightarrow \frac{1}{2}$

What is this actually mean that in case if you are trying to repeat the experiment for a large number of times, you will be computing the relative frequency for each time and as you are trying to increase the number of times of repetition, the relative frequency will stabilize at some point. So, this limiting value is interpreted as the probability of A and it is denoted by P(A). P(A) means P means probability and A is an event. So, which is the probability of an event A. So, this will be something like here limit n tending to infinity and A upon n. We had n A is indicating the number of times an event A occurs out of small n number of times.

So, now you can see this is the probability what we always say, whenever we say that if some event has a probability this. That means if you try to repeat the experiment for a large number of times. And if you try to compute this relative frequency, then incase if you increase the number of repetitions, then this relative frequency will be converging to this particular value. And if you try to see this part which I am writing here say n A upon n this is the same thing where you have studied that the probability is indicated by m upon n that if then experiment is repeated.

Small n number times than small m is the number of favorable events that is the same definition but now you can see that what does this mean, when you are trying to say using the definition of say m upon n and if you try to say that the probability of head is equal to 1/2 or the probability of tail is equal to 1/2. What is this mean? That in case if you keep on continuing the tossing of the experiment for a large number of times, then the value of the probability which is computed as

the total number of heads divided by total number of times the experiment is repeated. This value will be converging to $1/2$. As n goes to infinity. This is what you mean.

(Refer Slide Time: 13:15)

Relative Frequency and Probability of an Event: Example - Coin Toss

Suppose a fair coin is tossed $n = 10$ times. ✓
Number of observed heads $n(A_1) = 3$ times ✓
Number of observed tails $n(A_2) = 7$ times ✓

Meaning of a fair coin: Probabilities of head and tail are equal (i.e., 0.5).

Then, the relative frequencies in the experiment are

$$f(A_1) = \frac{3}{10} = 0.3 \text{ and } f(A_2) = \frac{7}{10} = 0.7.$$

When the coin is tossed a large number of times and n tends to infinity, then both $f(A_1)$ and $f(A_2)$ will have a limiting value 0.5 which is the probability of getting a head or tail in tossing a fair coin.

So, now means I try to first explain you the same example by considering some numerical values and then I will try to show you using the R software that if you try to conduct such experiment or if you try to actually simulate such an experiment. Then you can see how the probability is converging to value $1/2$. So, suppose if I say, suppose a fair coin is toss and equal to 10 times when I am trying to say here fair coin means that the coin is really fair in the sense that there is no impurity in the structure of the coin.

So, that a particular side head or tail comes more frequently. So, now in case if you try to repeat this experiment 10 times then suppose. We observed the number of heads which is here event A_1 . So, the number of times the event A_1 occurs is, suppose 3 times. And then obviously the number of time the tail will occur that is n of A_2 which is the absolute frequency that will become here 7 that means if you try to toss a coin 10 times then out of this, 3 times you get head and 7 times you get tail. And as I said the meaning of the fair coin is that the probabilities of head and tail are the same that is 0.5 or $1/2$.

Now in this case if you try to compute the relative frequencies of the events A_1 and A_2 . They will come out to be like this f of A_1 will become here. The number of times you are getting head divided by the total number of times the experiment is repeated that is 10. So, this will become

here 3 upon 10 which is equal to 0.3. And similarly, the relative frequency of A_2 that is observing the tail will become here 7 upon 10 which is here is 0.7. So, you can see here this is 0.3 and 0.7 but you always assume or you have been told that the probability of getting your head or a tail is simply 0.5.

But now we have understood what is the meaning of this. The meaning of this is this that when the coins is toss for a large number of time and goes to infinity then both f of A_1 and f of A_2 will converse to a value and this value or the limiting value is going to be 0.5. And this is the probability of getting a head or a tail when we are trying to toss a fair coin. This is what we mean.

(Refer Slide Time: 15:49)

Relative Frequency and Probability of an Event: Example - Coin Toss

Suppose the head is denoted by 0 and tail is denoted by 1.

Sample space $(\Omega) = \{0, 1\}$

Suppose we repeat the experiment 5 times and following outcome is observed:

Head, Head, Tail, Head, Tail

which is expressible as 0, 0, 1, 0, 1.

Relative frequencies of Tail = Probability of Tail

Probability of Tail = $\frac{\text{number of 1's (1+1)}}{\text{Total number of repetitions}} = \frac{2}{5} \parallel \rightarrow \frac{1}{2}$ as $n \rightarrow \infty$

Handwritten notes: H T, HHTHT, Sample mean of 00101 = $\frac{0+0+1+0+1}{5}$

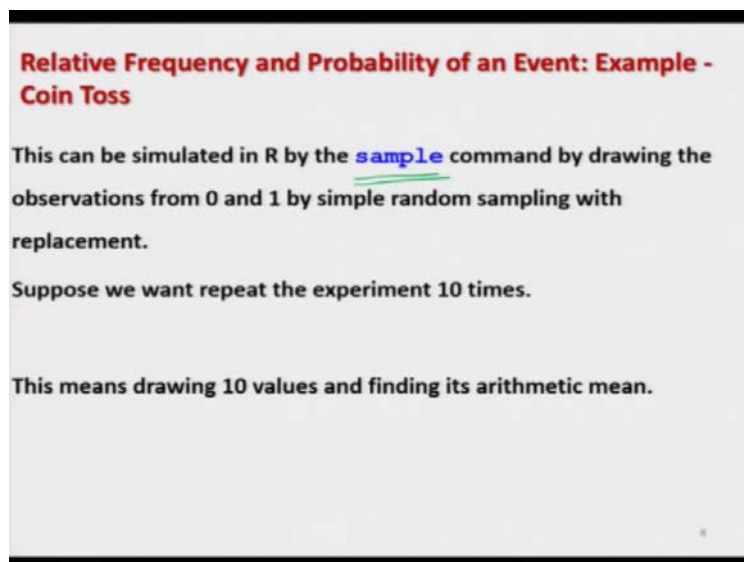
Now suppose we want to compute this probability so definitely we working here only with here two symbols here head and tail. So, with head and tail we cannot make any computations, so let us try to give them a mathematical representation. Suppose whenever the head occurs I can indicated by 0 and suppose if the tail is occurring I denote it by here 1. Suppose if I try to repeat the event say 5 times and suppose I get here H, H, T, H, T.

So, this thing can be represented by head is occurring that means 0, 0, tail is occurring that means 1 then head is occurring 0 and tail is occurring 1. So, sequence like 0, 0, 1, 0, 1 will indicate that in the first toss you get head. Second toss you get head. Third toss you get tails. Fourth toss you get 0 that is head and fifth toss you get 1 that means tail. So, this is what I have

written here exactly. That if you get here head, head, tail, head, tail this is expressible as 0, 0, 1, 0, 1.

Now in case if you try to count the relative frequency of your head or tail. For example, if you want to compute the relative frequency of tail which is actually the probability of tail that counted as number of 1's which is here 1 and 1 two times which is here 2 divided by the total number of repetitions which are here 5. Now in case if you try to repeat this experiments for a large number of times, you will see that this value 2 upon 5 will be practically converging to $1/2$ as n goes to infinity.

(Refer Slide Time: 17:38)



Now I would like to show you here. And in order to show it I will try to use the R software and in the beginning when I introduced you the basic concepts of R software that we are going to use in the course I had explain you the command here sample. I had introduce you the topic of simple random sampling in which I explain you that if you want to draw a sample by simple random sampling then I can use the command here a sample. And you see when you are trying to repeat an experiment you always say this is a random experiment. So, you know whatever be the outcome that is not going to be known unless and until the experiment is conducted.

So, when we are trying to conduct this experiment by indicating the head and tail by 0 and 1, then we can draw different number of zeros and ones from the R software using the sample command. And if you try to see here in this a case itself if you try to see here what are we trying

to do, I am simply trying to find out the sample mean of this 0, 0, 1, 0, 1. If you try to see the mean is going to be simply here 0 plus 0 plus 1 plus 0 plus 1 upon here 5. So, this is 2 upon 5 so that is another way by which you can think that if you simply trying to compute the arithmetic mean of the random number drawn, then possibly this will give you the relative frequency.

(Refer Slide Time: 19:08)

Relative Frequency and Probability of an Event: Example - Coin Toss

For example, the command

```
coin10 = sample(c(0,1), size=10, replace=T)
```

draws a sample of size 10 and stores the values in the data vector coin10. The command

```
table(coin10)/length(coin10)
```

find the relative frequencies of these 10 values.

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

But any way, let us try to use it. Suppose I try to first show you the outcome that experiment I already had conducted and I have prepared the slides and after that I will show you the same thing in the R console also. So, in case if I try to take two numbers here 0 and 1. And suppose I want to draw here or I want to repeat this experiment of drawing head and tail or even till the 0 and 1 say it 10 times.

And for that I here simple random sample with replacement that is replace equal to TRUE. So, this command that sample c 0, 1 size equal to 10 replace equal to T will draw a sample of size 10 and stores the value in the data vector say had coin 10. And in case if you another approach that you can always find out the relative frequency of this outcome. So, that can be obtain by the command table. The name of the variable divided by the length of the variable.

So, my variable here is coin 10. So, this will give us the relative frequencies of this 10 values which we are trying to observe. So, this way you can opt any of the approach to find out the arithmetic mean or the relative frequency. So, we try to repeat this experiment say 10 time then 100 times, 1000 times, 10000 times and so on. And we try to see the outcome.

(Refer Slide Time: 20:27)

Relative Frequency and Probability of an Event: Example - Coin Toss

10 repetitions

```
> coin10 = sample(c(0,1), size=10, replace = T)
> table(coin10)/length(coin10)
```

0	1
0.7	0.3

0,1, ... 10 times

```
> coin10 = sample(c(0,1), size=10, replace = T)
> table(coin10)/length(coin10)
```

0	1
0.3	0.7

10

You see when I am trying to repeat this experiment 10 times, then there are going to be 0, 1, 0, 1 say 10 times. And the relative frequency of those zeros and ones comes out to be here 0.7 and 0.3. And if I try to repeat this experiment once again. Then in the second repetition I am getting here the relative frequency is 0.3 and 0.7. So, what is happening that in the first draw the head are coming out to be 7 and the number of tails that they are coming out to be 3 but in the, but when I try to repeat it this is just getting interchange that we get here 3 heads and 7 tails but here if you try to see we have repeated the experiment only 10 number of times.

(Refer Slide Time: 21:13)

Relative Frequency and Probability of an Event: Example - Coin Toss

100 repetitions

```
> coin100 = sample(c(0,1), size=100, replace = T)
> table(coin100)/length(coin100)
```

0	1
0.55	0.45

```
> coin100 = sample(c(0,1), size=100, replace = T)
> table(coin100)/length(coin100)
```

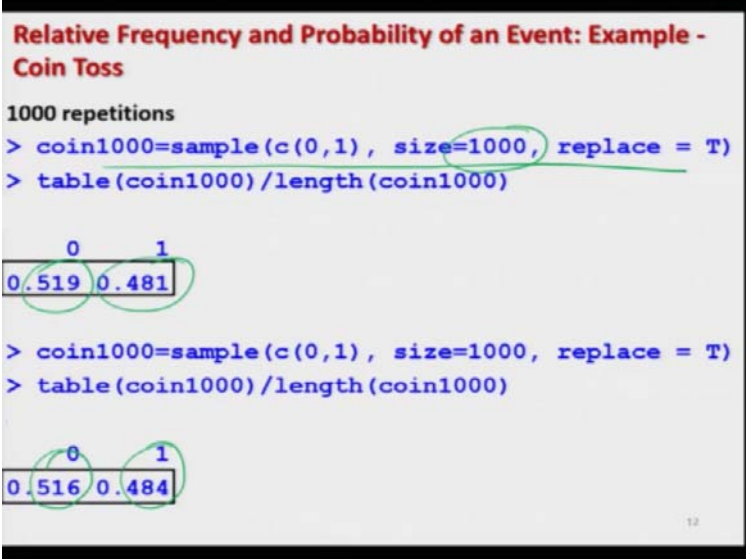
0	1
0.45	0.55

11

Now I try to repeat it. And I try to conduct the experiment 100 times that we get a 100 values of heads and tails. So, in this case you can see here that we use the same command here that table coin 100 divided by length coin 100. But in this case the numbers of head are coming out to be 55. And number of tails are coming out to be here 45. So, you can see here this probability which was earlier something like point 7 point 3 that is now converging to 0.55 and 0.45.

And if I try to repeat the same experiment here once again I am getting here the value 0.45 and 0.55. Well, means I can show you that here it is only by chance that these two outcomes and these outcomes are getting interchanged. But that is just by chance actually I have not done it. And I will show you the screenshot also. But you can see here as we are trying to increase the number of times the experiment is being conducted, the value of relative frequency is converging towards 0.5.

(Refer Slide Time: 22:15)

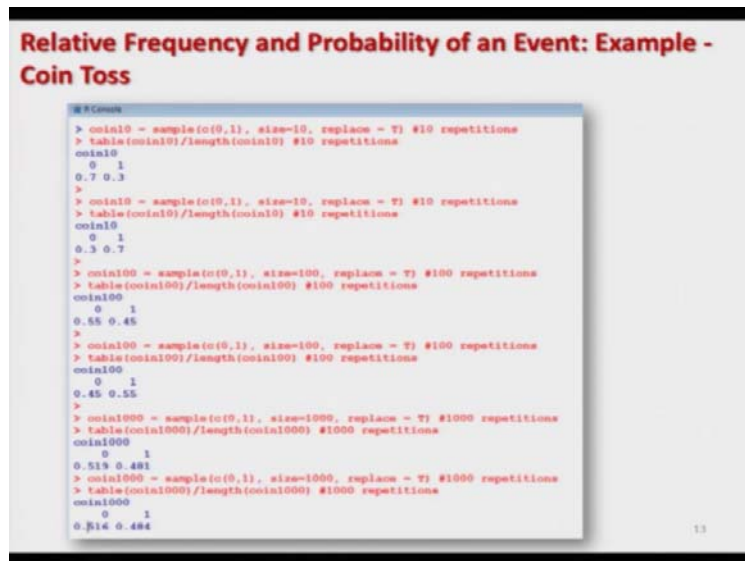


```
Relative Frequency and Probability of an Event: Example - Coin Toss
1000 repetitions
> coin1000=sample(c(0,1), size=1000, replace = T)
> table(coin1000)/length(coin1000)
 0      1
0.519 0.481

> coin1000=sample(c(0,1), size=1000, replace = T)
> table(coin1000)/length(coin1000)
 0      1
0.516 0.484
```

And now in case if you try to repeat this experiment say 1000 time. So, here I am trying to get here 1000 values of head and tails in terms of zeros and ones. And if I try to compute the relative frequency this is coming out to be 0.519 and 0.481. And similarly, if I try to repeat this. This is again coming out to be 0.516 and 0.484. But now you can see that as you are trying to increase the number of times of the repetitions of the experiment, these value are converging towards 0.5.

(Refer Slide Time: 22:45)

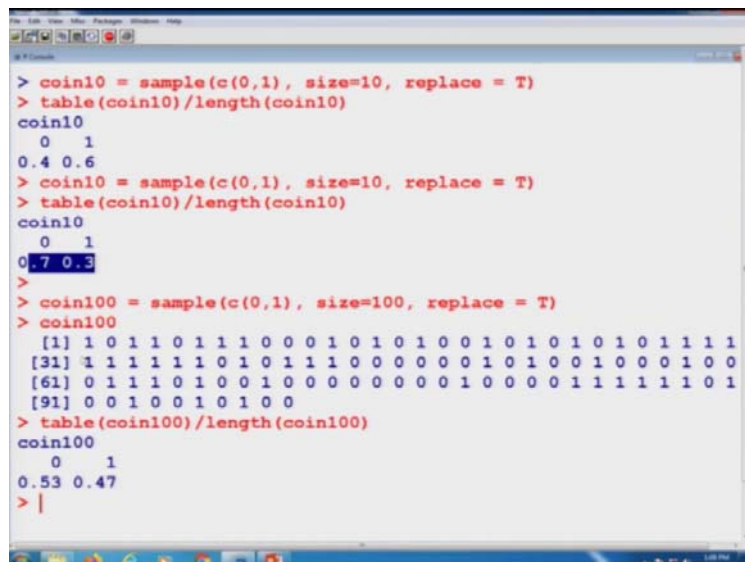


Relative Frequency and Probability of an Event: Example - Coin Toss

```
R Console
> coin10 = sample(c(0,1), size=10, replace = T) #10 repetitions
> table(coin10)/length(coin10) #10 repetitions
coin10
 0  1
0.7 0.3
>
> coin10 = sample(c(0,1), size=10, replace = T) #10 repetitions
> table(coin10)/length(coin10) #10 repetitions
coin10
 0  1
0.3 0.7
>
> coin100 = sample(c(0,1), size=100, replace = T) #100 repetitions
> table(coin100)/length(coin100) #100 repetitions
coin100
 0  1
0.55 0.45
>
> coin100 = sample(c(0,1), size=100, replace = T) #100 repetitions
> table(coin100)/length(coin100) #100 repetitions
coin100
 0  1
0.45 0.55
>
> coin1000 = sample(c(0,1), size=1000, replace = T) #1000 repetitions
> table(coin1000)/length(coin1000) #1000 repetitions
coin1000
 0  1
0.519 0.481
>
> coin1000 = sample(c(0,1), size=1000, replace = T) #1000 repetitions
> table(coin1000)/length(coin1000) #1000 repetitions
coin1000
 0  1
0.514 0.484
```

And this is here the screenshot of the same outcome which I have shown you but now I will try to show you these things on the R console also. So, let us try to come to the R console and we try to do the same thing over here. So, let me try to just copy this command on the R console over here.

(Refer Slide Time: 23:06)

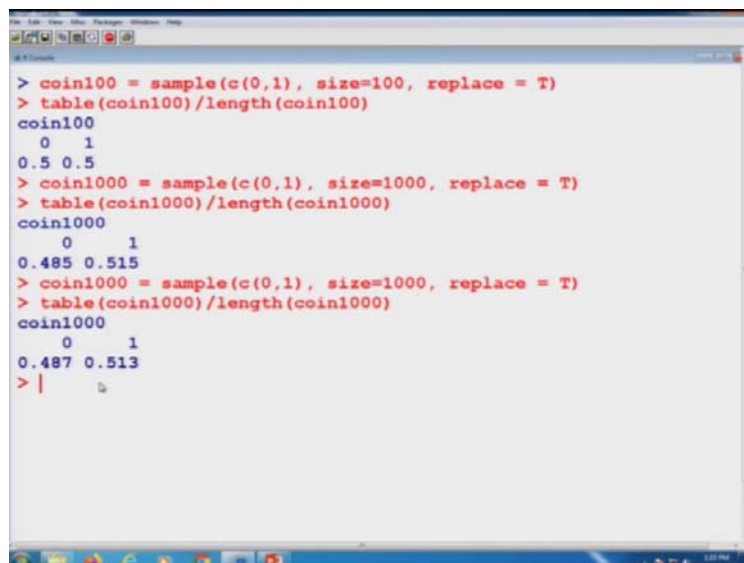


```
R Console
> coin10 = sample(c(0,1), size=10, replace = T)
> table(coin10)/length(coin10)
coin10
 0  1
0.4 0.6
> coin10 = sample(c(0,1), size=10, replace = T)
> table(coin10)/length(coin10)
coin10
 0  1
0.7 0.3
>
> coin100 = sample(c(0,1), size=100, replace = T)
> coin100
 [1] 1 0 1 1 0 1 1 1 0 0 0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 1 1
 [31] 1 1 1 1 1 1 0 1 0 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0
 [61] 0 1 1 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 1 1 1 0 1
 [91] 0 0 1 0 0 1 0 1 0 0
> table(coin100)/length(coin100)
coin100
 0  1
0.53 0.47
> |
```

And I try to execute it so that we save some time also and you can see here that if you try to do it. That it is coming out to be 0.4 and 0.6. That means if you are getting 4 heads and 6 tails. And if you try to repeat this experiment you can see here that this time it is coming out to be 0.7 and

0.3. But now let me try to increase the number of observations tools suppose here 100. And let me try to put it here say coin 100. So, you can actually see here this is how you will get the value here 0, 1, 0, 1 and so on. But if try to compute here this relative frequency of this coin 100 data vector. You can see here that it is coming out to be 0.53 and 0.047 which is close at 0.5 compared to 0.7 and 0.3.

(Refer Slide Time: 23:58)



```
> coin100 = sample(c(0,1), size=100, replace = T)
> table(coin100)/length(coin100)
coin100
 0  1
0.5 0.5
> coin1000 = sample(c(0,1), size=1000, replace = T)
> table(coin1000)/length(coin1000)
coin1000
 0  1
0.485 0.515
> coin10000 = sample(c(0,1), size=10000, replace = T)
> table(coin10000)/length(coin10000)
coin10000
 0  1
0.487 0.513
> |
```

And if you try to repeat this experiment, so value will change and it is coming out to be 0.5 and 0.5 in the case of 100. But this is only a matter of chance actually that it is exactly coming out to be 0.5 and 0.5. But now in case if I try to make it here the sample size to be 1000 that means we are trying draw 1000 values of heads and tails that means we are repeating it 1000 time, then this relative frequency will come out to be here like this of the relative frequency of coin 1000 data vector. And you can see here this is coming out to be closer to 0.485 and 0.515 and if you try to repeat this experiment then you will see here that this probability is approaching towards 0.5.

So, now if you try to repeat this experiment for say 10000 time 1 million times you will see that their relative frequency is approaching towards 0.5. And this is what I mean when we say that the probability of getting a head or a tail is $1/2$. So, now we come to an end to this lecture and I have taken a very simple example to convince you that what is the definition. What you have done earlier? And what is the interpretation? What is the correct meaning of this thing?

Now I would say you try to take some more example try to compute it the probability by hand manually. And try to execute it under R software and try to see what happens. The more you practice this will give you a better insight that how the probability is interpreted and computed. So, you try to practice it and I will see you in the next lecture with one more example, till then goodbye.