Essentials of Data Science with R Software – 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture No. 33
Binomial Distribution in R

Hello friends, welcome to the course of Essentials of Data Science with R Software – 1; in which we are trying to handle the topics of probability theory and statistical inference. So, you can recall that in the last lecture we had considered the Bernoulli and binomial distributions. And we also understood that in the binomial distribution, if you try to take n equal to 1, that will become the Bernoulli distribution.

So, now in this lecture I am going to show you that how you can handle binomial distribution; or equivalently the Bernoulli distribution in R software. So, whenever you are trying to employ these probability mass function or or say any probability density function also, in R software there are couple of objectives that we want to fulfill. The first thing is this how to generate a random number from that distribution. That means how to generate a number which will look like as if I have conducted this experiment in real life.

And suppose if I try to repeat that experiment 100 times that means, if I try to generate here 100 such observations. They will look like as if I have conducted the experiment in real life 100 times. So, for that how to generate an appropriate random number; then how to compute the cumulative density function or cumulative distribution function, why? Once you can compute the cumulative distribution function that is CDF; you can compute various types of probabilities very easily.

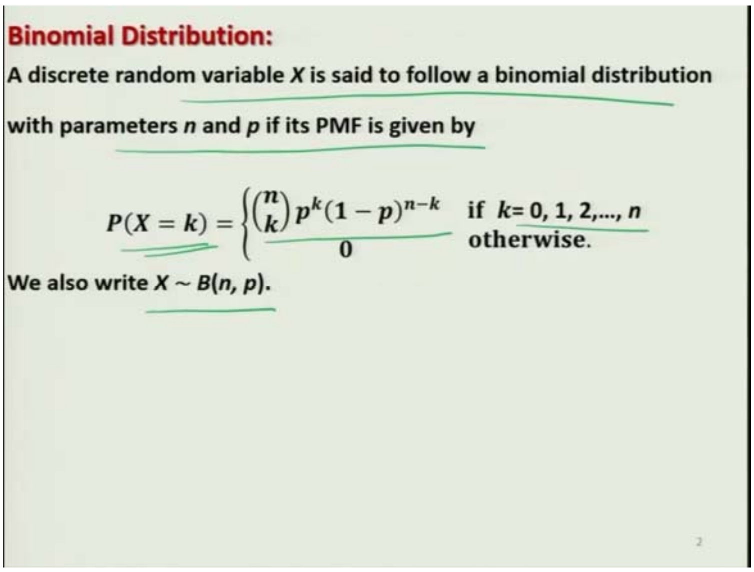
And similarly, if you want to compute the probability at any point; that is the value of the density at any point. You have to be careful that when we are trying to use the R software; then it will always call it as a density. Whereas, this concept is valid, when we are trying to consider the probability density function in the case of probability mass function. We call it as a probability mass function, but you do not get confused with those things. So, these are some important properties that we always try to find through the R software.

So, what I am going to do that I will try to show you all these commands, what are the important parameters, what are their interpretation, how are you going to execute it, how are you going to interpret it. But, definitely one request before I move forward; whatever commands I am going to give you here, they are in much detail. But, surely it discussing all the possibilities, all the options; it is not possible. So, whenever you want to understand these things my very honest suggestion is that look into the help menu of that option.

For example, if you want to generate the binomial distribution; try to look into the help menu. And it will give you entire information, complete details and that will make you complete. My objective here is very simple, now at this moment, you have to ask yourself and tell, do you know anything about that how to do such computation of binomial distribution R software. Answer is yes, I am sure that after half an hour when I am completed the lecture; then you will see that what was there, it was a very simple thing; but that is my objective.

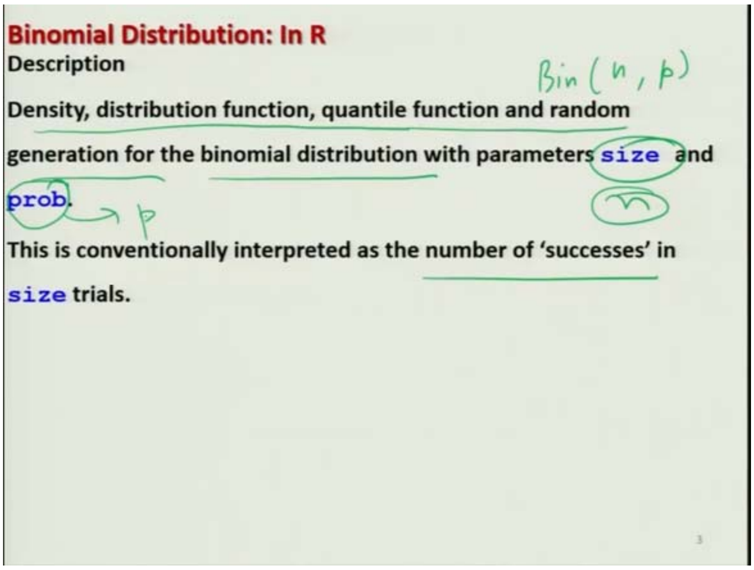
I want to make you say that that was a pretty simple thing; but, it does not mean that it is complete. Why? My job here is to give you confidence that you can learn it. Once you can learn these basic fundamentals; I assume that you are on the track and then after that you can travel and sky is the limit. So, let us begin our lecture and try to see how we can handle the binomial distribution in the R software.

(Refer Slide Time: 04:25)



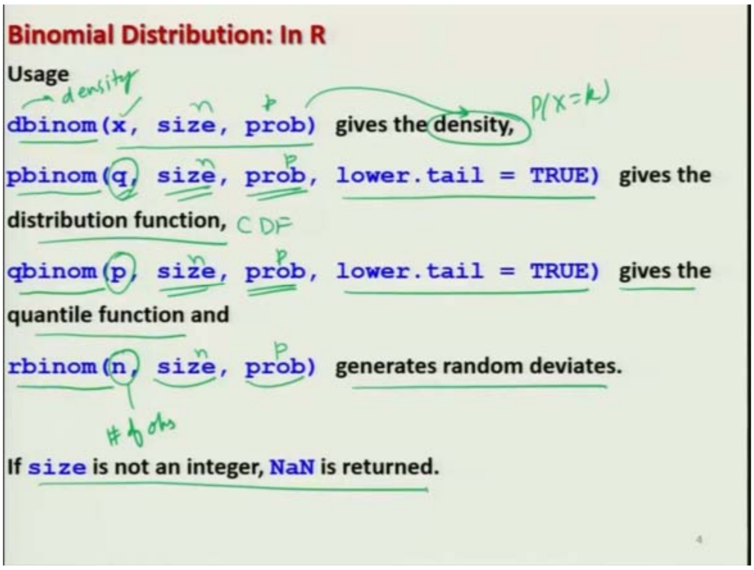
So, a discrete random variable x is set to follow a binomial distribution with parameters n and p, if its probability mass function is given by, probability of X equal to k is equal to $\binom{n}{k}$ $p^k(1-p)^{n-k}$; if k is equal to 0, 1 …, n. And this is indicated by X follows a binomial n, p; this is what we have learnt in the last lecture.

(Refer Slide Time: 04:49)



Now, I am coming to the say R software how are you going to handle it. So, in the R software, you can compute the density distribution function that is CDF. You can compute different types of quantile function, and you can generate the random numbers from the binomial distribution. And here as you have done that it is indicated by binomial n, p; there are two parameters n and p. So, n is going to be indicated by their size and p; and p is going to be indicated by another parameter, say prob. And there is one understanding in R that very conveniently we interpret the the value of n as the number of successes. How do you define successes, that is up to you.

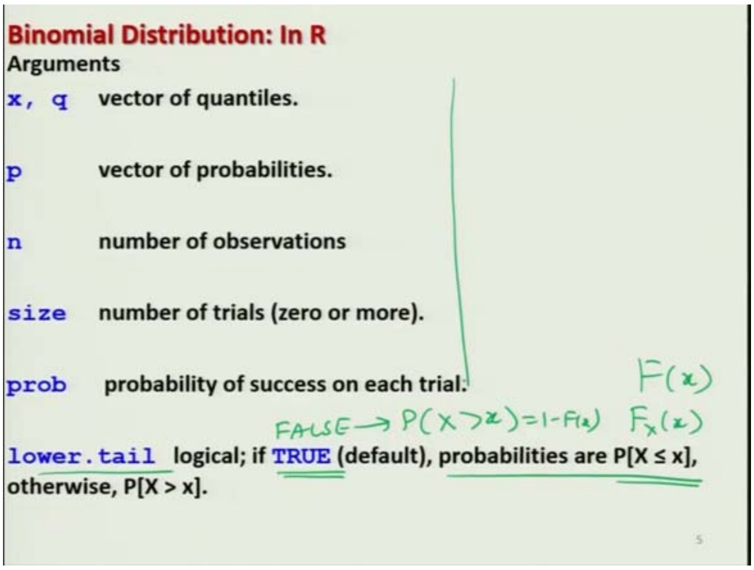
(Refer Slide Time: 05:45)



So, now in case if you try to look into the commands; there means I will be discussing here the four possible commands. Say, one here is dbinom, so dbinom that is the abbreviation for binomial; and d means density. So, the command dbinom, and in case if you write down the parameters x, size and prob inside the parenthesis. This will give you the value of the density for a given x, n and here p; so, you can compute any type of probability, the value of probability of X equal to k.

And similarly, if you try to use the command here pbinom p b i n o m; and then this is going to give you the CDF, the distribution function. And inside the parenthesis you have to give a couple of parameters; one here is q, then here is size, which is here n, prob is prob which is p. And lower dot tail equal to f is a logical variable that can take a value, say TRUE or FALSE. So, I will try to show you that how it makes the difference.

(Refer Slide Time: 7:28)



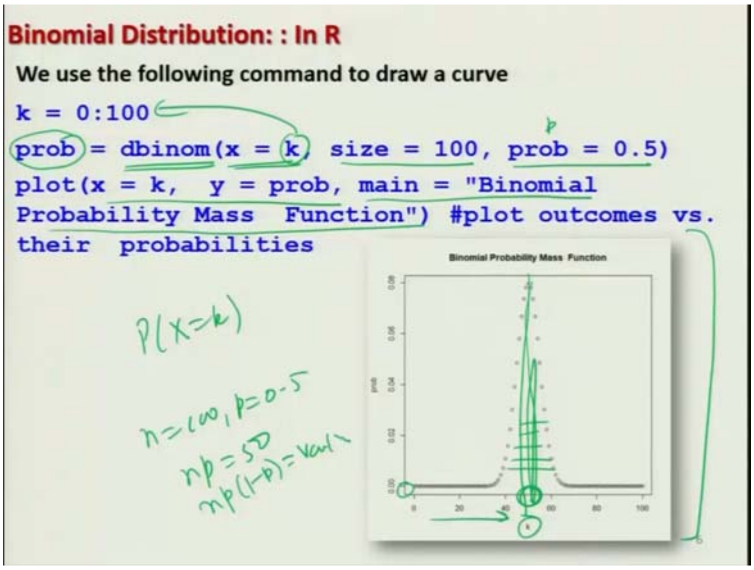
For example, if you try to see that in case if you try to take lower dot tail is equal to TRUE, which is actually the default. The you compute the probabilities, probability X less than equal to x, which is your actually $F_X(x)$ or you had written here as a F(x). And in case if you make it FALSE, if it is FALSE, then it is going to compute the probability X greater than x; which is 1 minus F of x. So, actually this will be true for all the probability mass functions and probability density function, which we are going to understand in the forthcoming lecture.

So, I have explained you here, after this I will assume that the same story will continue there itself. So, this is the meaning of this lower tail and q is the data vector; where you have to give that what is the value of q, for which you want to find out the CDF. Similarly, this q b i n o m qbinom; this will provide the value of the quantile function. Do you remember we had done the quartiles, deciles, percentiles, et-cetera; they were called in general as quantiles. So, this qbinom will provide as the quantile function, for a given value of p. And this then you have to specify the size n, probability p and lower dot tail equal to TRUE; that is the same thing, as you have done in the case of pbinom.

And similarly there is another command here r b i n o m rbinom; this generates the random number. So, this is the number of observation n; be careful this is not the n of binomial n, p. But, this is the total number of observation that you want to generate, and size here is n, n and p r o b prob here is p. So, in case if you try to use this rbinom, you will get here the random numbers; which will appear as if you have conducted the experiment in real life.

So, this is very useful when you are trying to do different type of simulation experiment. And remember one thing if this this size is not an integer, then we get the outcome NaN. What is NaN? Means you have to learn from your R course. So, now here are the details of the symbols that I just explained you.

(Refer Slide Time: 10:21)



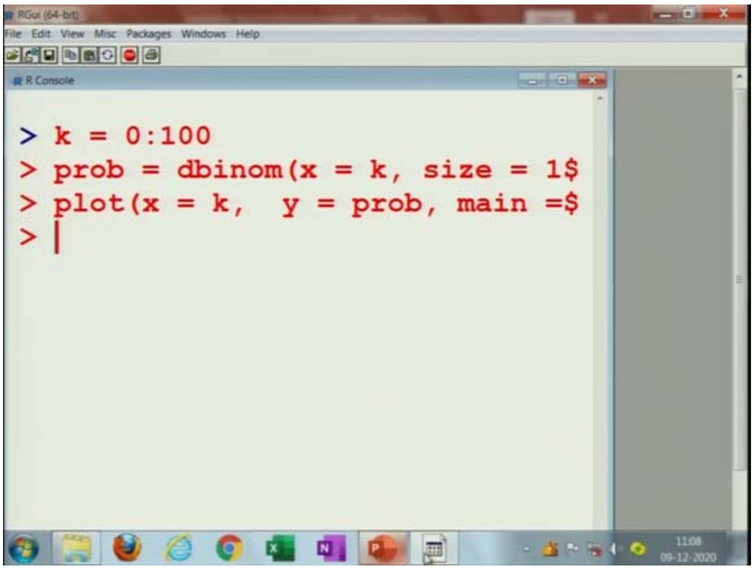
So, now let me try to take different types of of exercises and try to illustrate you many many things, in the given time frame. Suppose I want to plot this density of the binomial mass function probability, binomial probability mass function. Means I just want to know how the observations will look like, means if I try to really conduct this experiment in a real life. So, what I try to do here that I try to use the command dbinom to produce the probabilities.

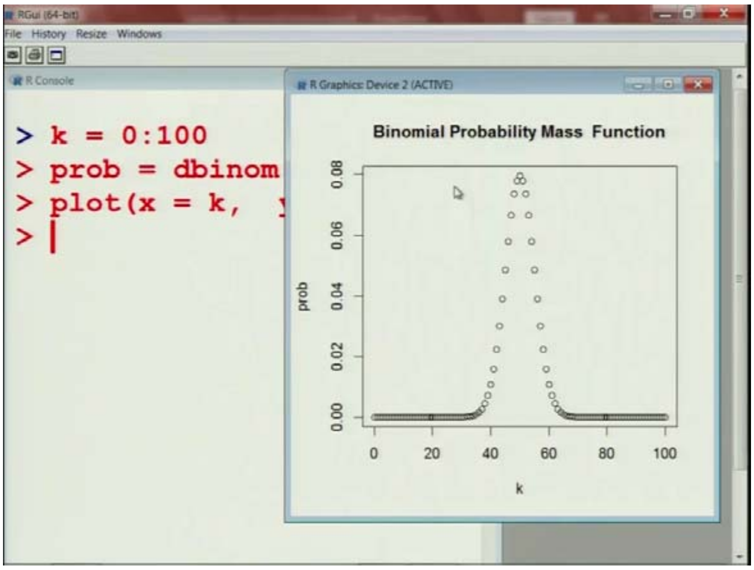
And I try to compute the density at x equal to k and I try to vary the k between say 0 to 100. So, k will be going from 0, 1, 2, 3, 4 up to 100 you can take any value; but I am just taking here 100. So, that I can show you clearly and the value of n, I am taking to be 100; and suppose I am taking the value of p to be here 0.5; means equi-probable case like as head or tail. My idea is just to give you an illustration that how the different things or different types of information can be retrieved from such commands; which are really going to help you in the data science.

And after that I try to plot these probabilities which I have stored in the this function prob. So, and then I try to plot this say this k and prob like a like this; and means this main will give the title that binomial probability mass function etc. If try to now execute this expression on the R console, you will get a graph like this one. You can see at these are the value of k this at 0, the probability is 0; and as the value of k is increasing, means probability of X equal to k that is increasing; you can see here the curve will look like.

It is not actually a curve, these are the individual points; because this is a probability mass function. So, you can see this will look like this, you can see here at this moment somewhere; this is the average value. You can see here that n is equal to 100 and p is equal to 0.5; so the mean is np which is close to 50. So, you can see at this around around this point, the values are concentrated equally on the both the sides. And if you can see there is a variation in these values; so that is going to give you the idea of np into 1 minus p, which is the variance of x. So, now first let me try to show you this computation on the R console, and then I will move forward.

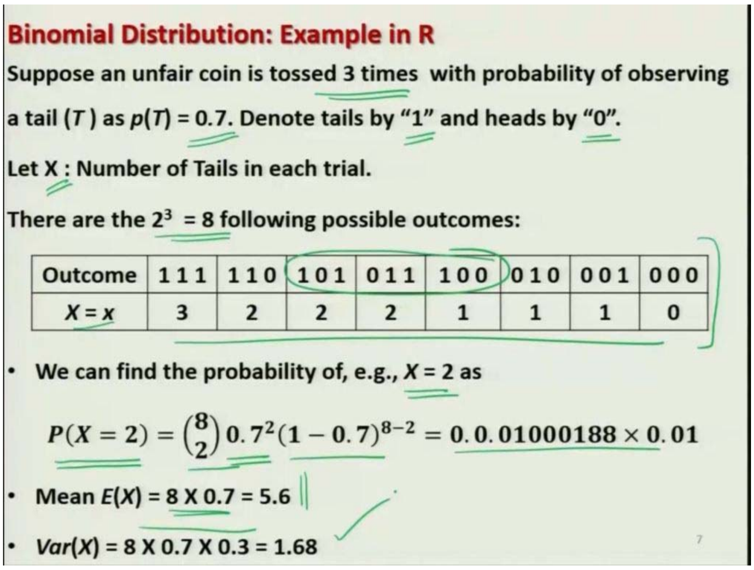
(Refer Slide Time: 13:39)





And who can see here that this type of information is really going to help you in real life. Because I always felt as a student that what these values are trying to give us, and how this probability will look like. How the picture graphics will look like, so that is what I am trying to do here. And if you try to plot such a curve here, you simply just copy and paste those commands over here. You can see we are obtaining the same curve that we have obtained here; so means how to plot and how to use the plot function. Now, that you have to see yourself, because I believe that you have done these things in the R software.

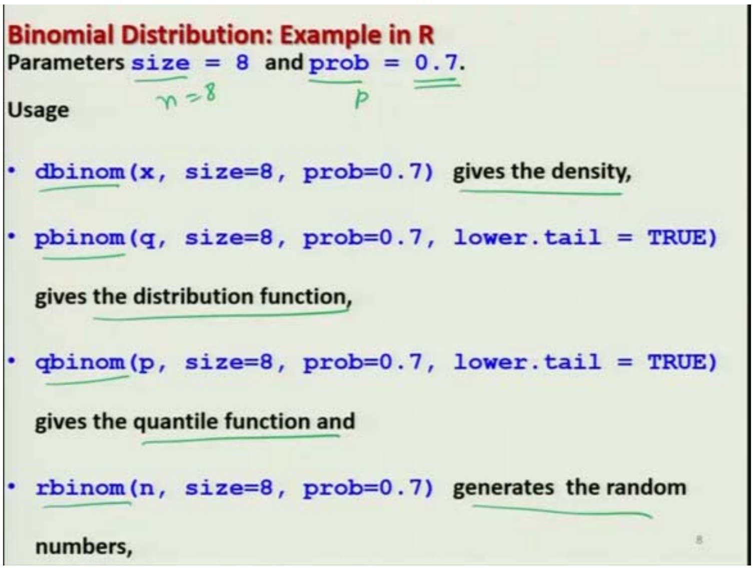
(Refer Slide Time: 14:20)



So, now let me try to take the example that I consider in the last lecture, in which there was a coin that was tossed 3 times. And the probability of observing that tail was 0.7, it was not 0.5; and the occurrence of tail is indicated by 1 and the occurrence of head is indicated by 0. So, and this x is the random variable indicating the number of tails in each trial. So, there are total two cube, 8 possible outcome, which are listed here; and the value of x are also given here. And this experiment can be imitated using the concept of binomial distribution.

And suppose if you want to compute the probability of X equal to 2; that means any of these three events are occurring. Then this probability can be computed by just using the binomial distribution, probability X equal to 2; which is 8 choose 2, p raised to the power of 2 into 1 minus p raise to power of n minus x. That is 8 minus 2 and this will come out to be some value here. So, now in case if you try to see the mean value, this will come out to be n into p; and the value of variance that is n p q. That can be obtained by here 8 into 0.7 into 0.3, which will come out to be 1.68.

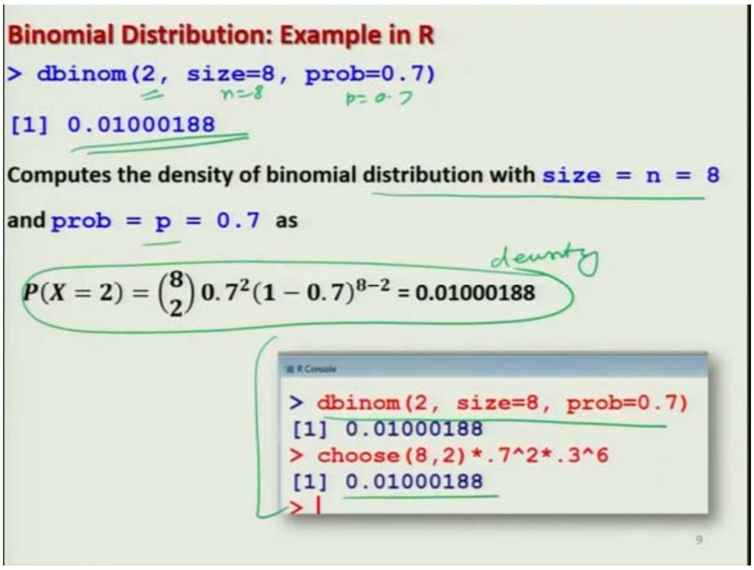
(Refer Slide Time: 15:43)



But, my objective is that I want to conduct the same experiment inside the R console that looks strange. But, let me try to show you, how you can imitate the same experiment inside the R console; and and it will look like as if you yourself have conducted the experiment. So, now I use the parameter size and prob; so size here is n is equal to 8, and prob here is p which is given as

0.7. And we know that this the command dbinom is going to give us the density, pbinom is going to give us the CDF, qbinom is going to give us the quantile function; and rbinom is going to give us the random numbers, that can be generated from this distribution.

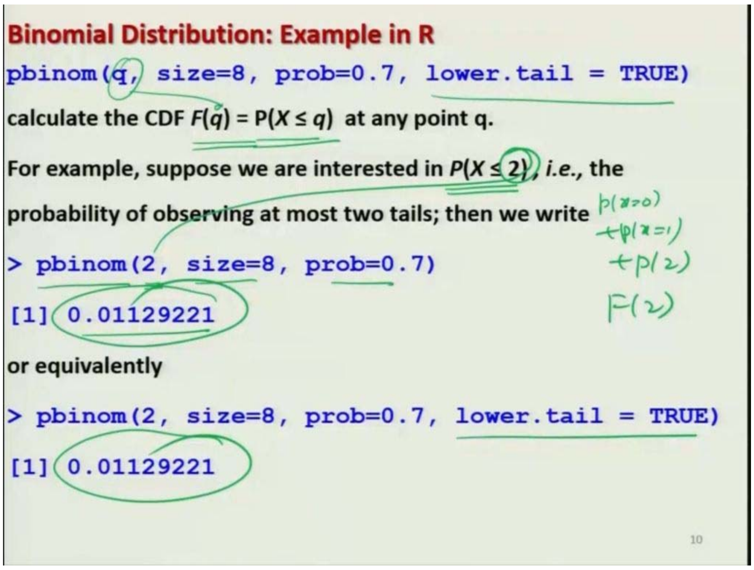
(Refer Slide Time: 16:32)



So, now let me try to show you the applications, and I would try to show you that whatever you wanted to compute manually that can be computed R. So, now if you see we have computed this value; so now if you want to compute the same value in the R software. Now, you have to understand which of the command is going to use; this is what this is density that means the probability. So, you have to simply say here dbinom 2, n is equal to 8 and p is equal to 0.7; and this will give you this probability, which is mentioned here.

So, you can see here that this compute the density of the binomial distribution with size equal to n and prob equal to p. And you can see here that both the values are going to be the same, and if you try to compute these values on the R console also you can see that this you have obtained from the R software, and this you have obtained from the direct computation. So, now you can see the probabilities that you are trying to compute manually, they can be computed directly on the R console also.

(Refer Slide Time: 17:57)

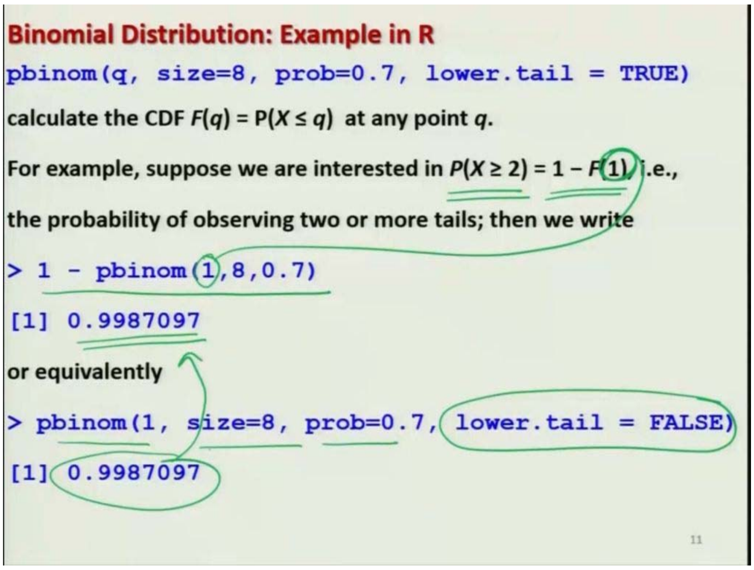


So, now I try to illustrate the use of pbinom. So, this pbinom computes the CDF and CDF was something like F(q) which is probability that X less than q; or less or X less than or equal to q, so this q is given here. So, now you can have an option that lower dot tail is equal to TRUE or FALSE, depending on your need. So, for example suppose we are interested in the same example in probability that X less than or equal to 2. Means so that means that X takes the value 0, 1 and 2 and this is going to be probability at x equal to 0, plus probability of x equal to 1, plus probability that x equal to 2. So, now this can be obtained by F(2).

So, now for that I simply have to write down here pbinom, then here 2; this 2 is coming from where? Here. And then n and p and then you will get here the value like this. And the same thing you if you want to obtain, you can also write down here lower dot tail is equal to TRUE. And you can see you are getting the same value, because as we discussed lower dot tail is equal to TRUE is the default. If you want to change it that can be converted into FALSE, depending on the need what you have.

So, you can see here computation of CDF is very easy, but my question is that what is this computing that will come only from the theory. That what was the CDF of the binomial distribution.

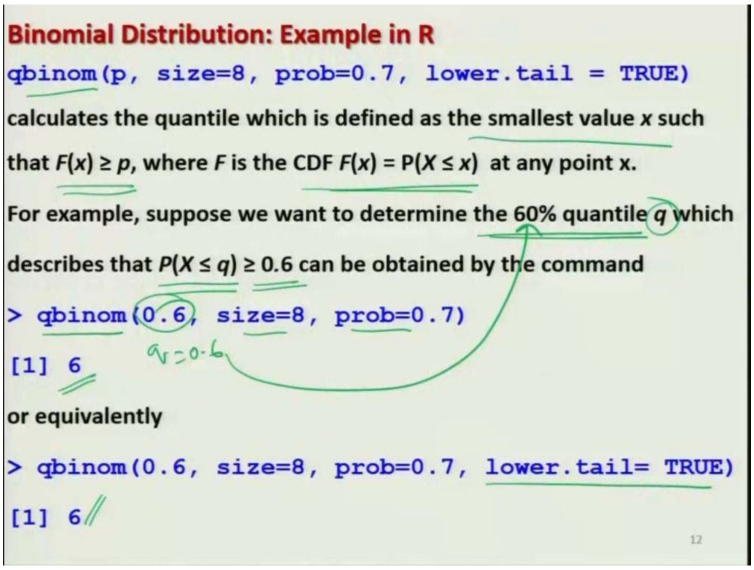
(Refer Slide Time: 19:36)



And that is why you need a strong theoretical background to understand this data science. And similarly suppose if you want to know the probability that X greater than or equal to 2; so, that is going to be simply here 1 minus F(1). So, now if you want to do this thing, then I have here two options; that I can write down here 1 minus pbinom, and this here 1. This 1 is coming from here this 1, and np; so this will give me the value of the CDF. Or, equivalently I can also use here the option lower dot tail is equal to FALSE. And in this case I do not have to subtract like 1 minus F(1); but I can simply say pbinom 1, size and prob values; and lower dot tail is equal to FALSE.

So, this will give me the same value what you have obtained. So, you can see whatever way you want to compute that you can compute according to the need of the experiment. But, definitely if you do not know the value of PDF, CDF, EMF etc.; you cannot understand what they are trying to do.

(Refer Slide Time: 20:48)

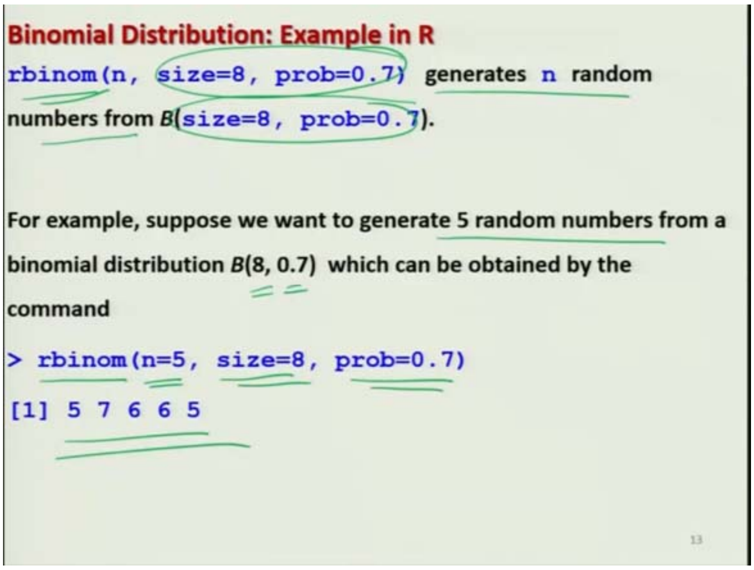


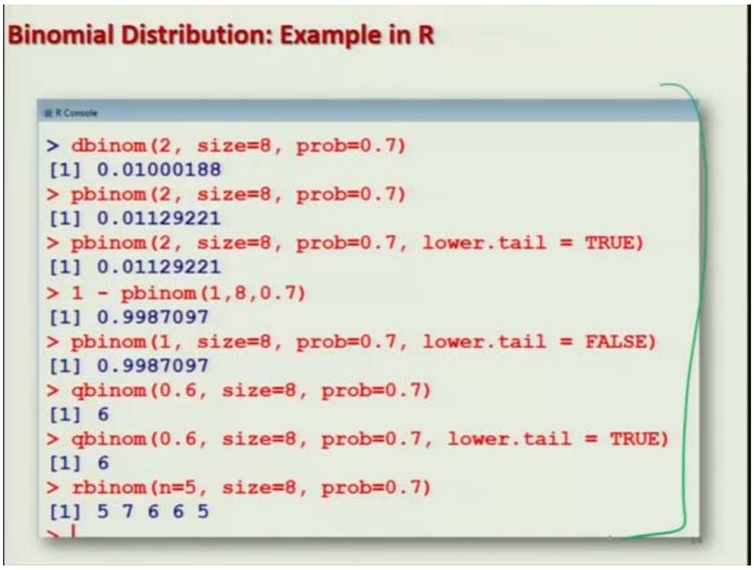
Now, let me try to take the option qbinom which is going to compute the different types of quantiles. So, the quantiles are what? They are defined as the smallest values, small x such that F of x is greater than or equal to p; where F is the CDF of x, at any given point x. For example, suppose you want to determine the 60 percent quantile; suppose this quantile is given by 6, which is described as probability X less than equal to q is greater than or equal to 0.6. So, this can be directly obtained by using the command qbinom. So, and then you have to give here q is equal to 0.6; so this is coming from here this 60 percent quantile. And then you have to give the value of n and p, and then you can get here 6.

And equivalently if you want to use the lower dot tail is equal to TRUE, you can also use it; but you will get the same value here. But, now you have to understand one thing that earlier we had obtained these quantiles using the function quantile. But, now here in this case you are obtaining the quantiles from a binomial distribution. That means your random variable is going to follow the probability laws, where the probabilities are going to be controlled by the mathematical form of the binomial distribution.

In the other case you had simply computed the the probability without considering any probability distribution. And you can see that every probability distribution will have a different type of feature; so this quantiles are going to be change. And this function helps you in finding out the quantiles from a binomial distribution. And similarly, when we try to consider different probability function, we will have a different commands.

(Refer Slide Time: 22:42)

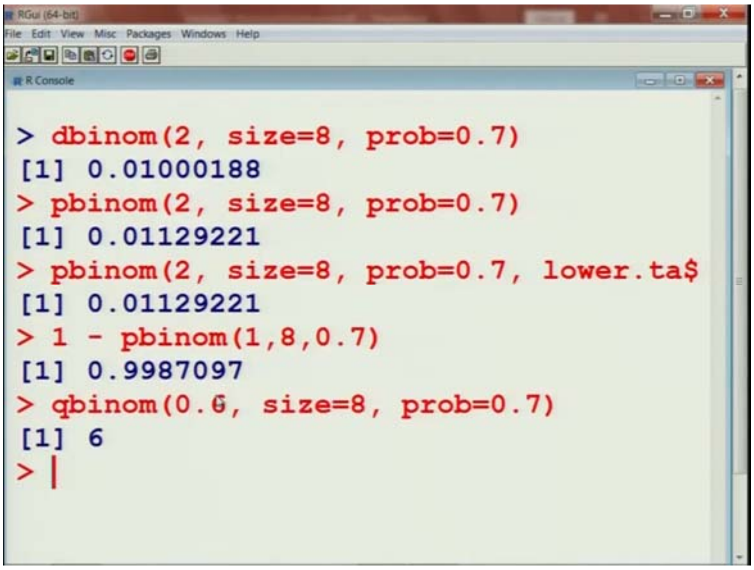


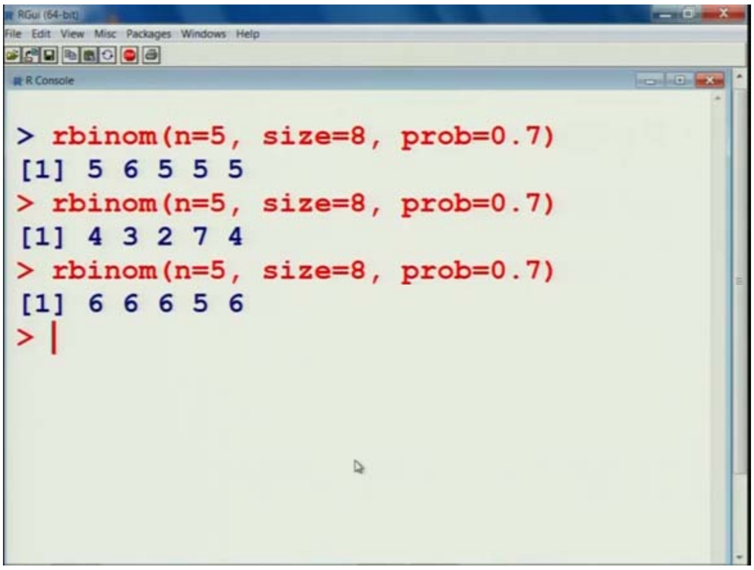


Now, we come to the command rbinom; so this rbinom is going to generate the small number of random numbers from the binomial distribution; with the size and prob parameter which are given here. For example, if you want to generate 5 random numbers from a binomial distribution, with n equal to 8 and p equal to 0.7. According to your definition this can be obtained by writing rbinom n equal to 5, size equal to 8, prob equal to 0.7.

And you can see that these are the 5 values and they are resembling as, if you have conducted the experiment and got these values. You can see here whatever outcomes I have shown you this is the screenshot. I will try to show you now these things on the R console, so that you get more convinced.

(Refer Slide Time: 23:42)





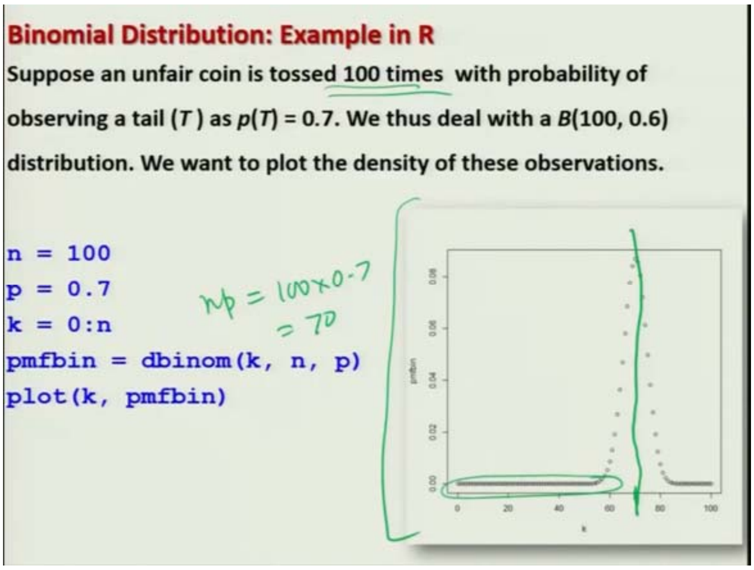
So, let me try to find out here directly from dbinom, so this is going to give you the probability for x equal to 2. So, let me try to clear the screen, you can see here this is coming out to be like this. And similarly, if you try to compute this CDF; you can see here this is coming out to be.

You simply have to write this command here, like this you can see this is the same value; and if you try to give the option of lower dot tail is equal to TRUE. You can see this is going to give you the same value, you can see here this is the same value as here like this.

And similarly, if you want to compute the the probability of X greater than or equal to 2; you can compute it by writing 1 minus pbinom. I am showing you here quickly, so that you can be confident that whatever you are going to get; that is the same thing what I am trying to show you. So, this is giving you the 60 percent quantile, which is here 6.

And similarly if you want to generate the rbinom that is you want to generate the random numbers. Definitely, now this is not going to be the same, when you try to do it yourself; because these random numbers are going to be change. Change every time if I try to repeat it 3 times, you are getting 3 sets of random numbers. So, here you have to be careful and now we try to come to one more part.

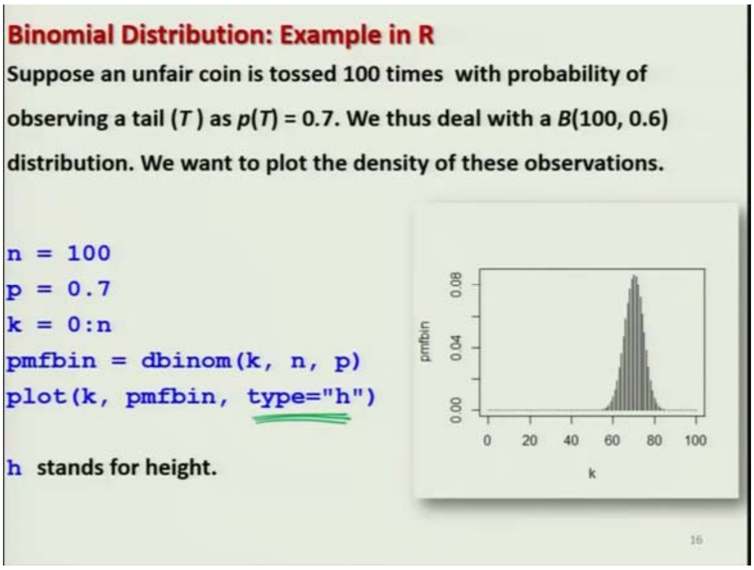
(Refer Slide Time: 25:20)



So, this is just for your information, you can do the same thing with the whatever, I had shown you earlier that I created this curve. Suppose an unfair coin is tossed 100 times, then definitely if you want to compute the probability at a k equal to 0, 1, 2 up to here 100. Then, you can plot these types of thing and you can see here this curve will look like this. Since, I have already have done this job, so I will not show you; but my main part here is interpretation. You can see this np is mean is a 100 into 0.7, which is close to 70; and you can see this mean value is close to 70.

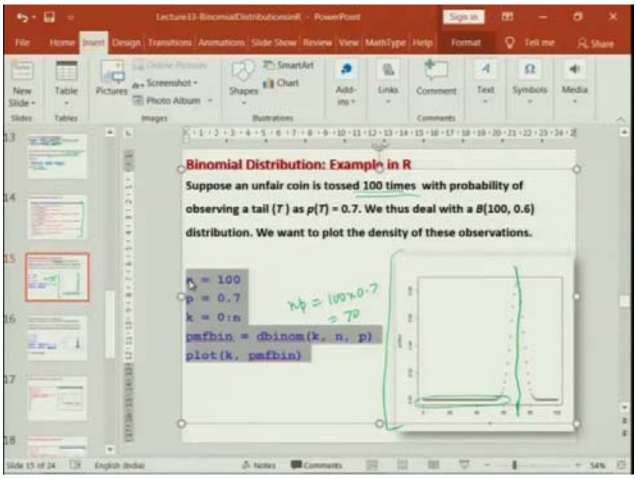
So, by looking at such curves, you can have a fair idea that in case if k is very low what is the probability? And if the k is increasing, then how the probability is increasing and after decreasing?

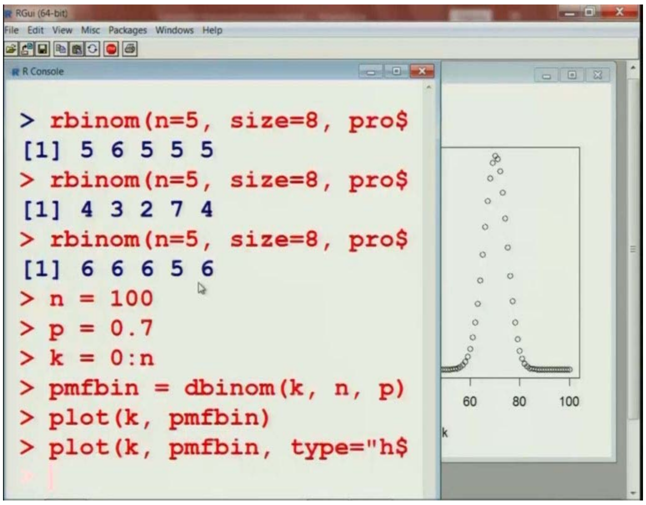
(Refer Slide Time: 26:15)

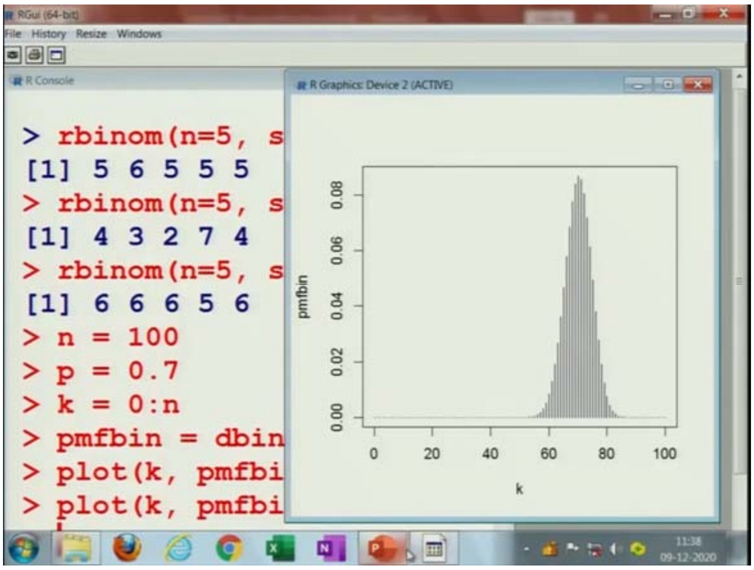


And similarly, if you try to do the same thing just by giving the different type of plot, which was a high density plot, the same thing what you have obtained here that can be obtained in the height format.

(Refer Slide Time: 26:34)

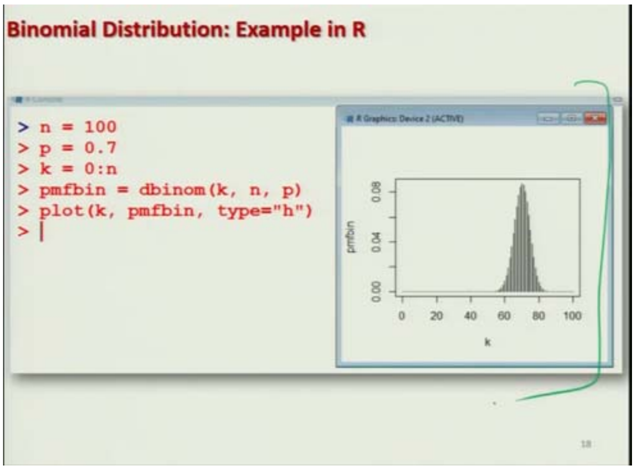

Binomial Distribution: Example in R
Suppose an unfair coin is tossed 100 times with probability of observing a tail (T) as p(T) = 0.7. We thus deal with a B(100, 0.6) distribution. We want to plot the density of these observations.
n = 100
p = 0.7
k = 0:n
pmfbin = dbinom(k, n, p)
plot(k, pmfbin)


> rbinom(n=5, size=8, pro$
[1] 5 6 5 5 5
> rbinom(n=5, size=8, pro$
[1] 4 3 2 7 4
> rbinom(n=5, size=8, pro$
[1] 6 6 6 5 6
> n = 100
> p = 0.7
> k = 0:n
> pmfbin = dbinom(k, n, p)
> plot(k, pmfbin)
> plot(k, pmfbin, type="h$
60
80
100
k


R Graphics: Device 2 (ACTIVE)
> rbinom(n=5, s
[1] 5 6 5 5 5
> rbinom(n=5, s
[1] 4 3 2 7 4
> rbinom(n=5, s
[1] 6 6 6 5 6
> n = 100
> p = 0.7
> k = 0:n
> pmfbin = dbin
> plot(k, pmfbi
> plot(k, pmfbi
pmfbin
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
k

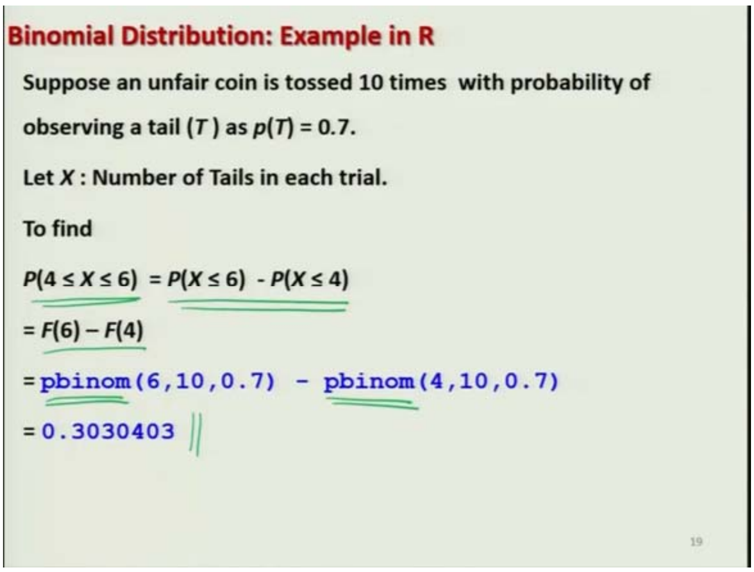
So, I think let me show you these things so that you get more confident; it will take a couple of minutes. So, if you try to see let me try to see here, if you try to plot here; this will give you a curve like this one. And if you try to plot with type equal to say h; that means high density, you can get this type of curve. So, so these types of curves will give you different types of information that we will try to learn in the forthcoming lectures.

(Refer Slide Time: 27:10)



So, these are the very simple thing and these are the screenshot of the same thing, which I shown you here; so that you can be confident that I am showing the correct outcome.

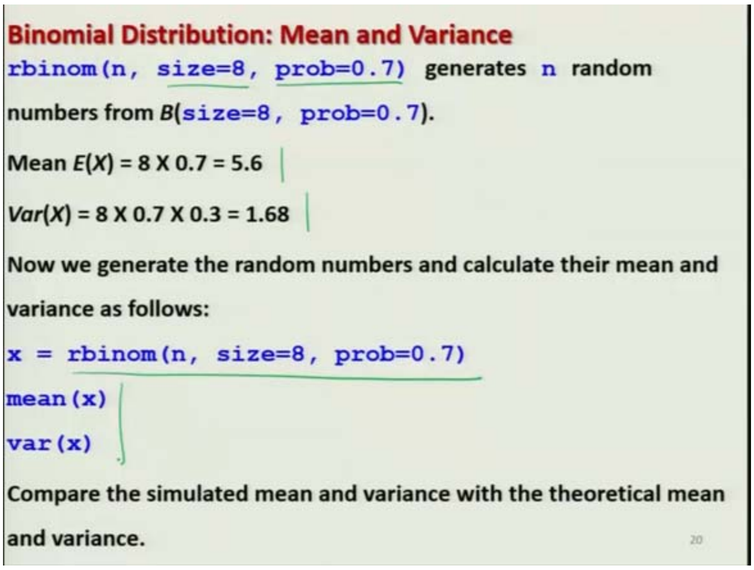
(Refer Slide Time: 27:19)



Now, let me try to show you that how you can compute a different type of probability using the R software. Suppose we consider the same example and we want to compute the probability that X is lying between 4 and 6. So, by the rules of CDF, we know that we can write it as probability X less than equal to 6 minus probability of X less than equal to 4; which is equivalent to here F(6) minus F(4). So, now you know how to compute this F(6) and F(4) using the pbinom command; so, you simply try to give these values and you will get the probability.

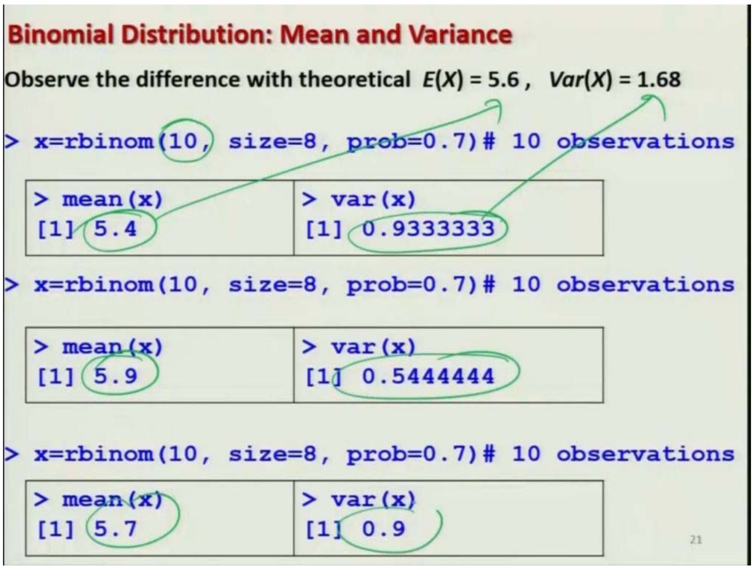
So, you can see that whatever you have learned during the CDF that how to compute different types of probability; then can be computed here directly.

(Refer Slide Time: 28:04)



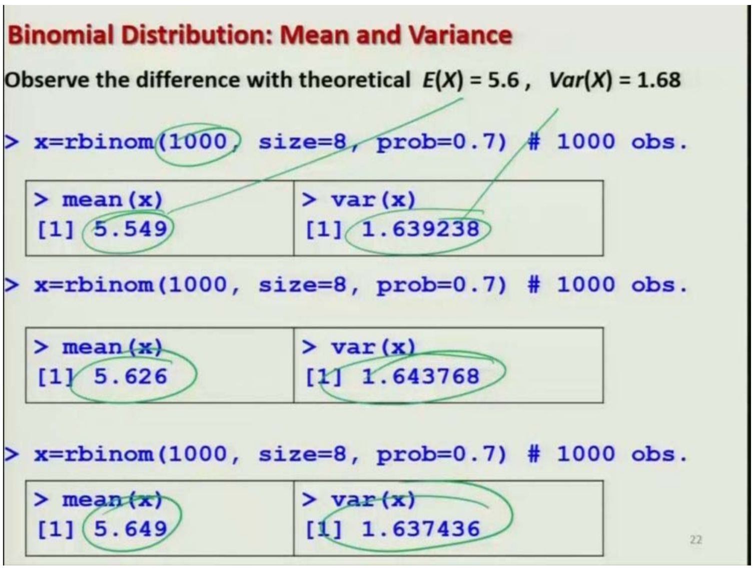
Now, suppose if you try to take the example from say that you try to generate R, say here random numbers and with the parameter n equal to 8 and p equal to 0.7. And here in this case the theoretical mean will come out to be 5.6, and variance will come out to be 1.68. And now I want to show you that how these things can be viewed, from the simulation. So, I try to generate the random number from this from the same distribution; and I try to compute their mean and variance.

(Refer Slide Time: 28:41)



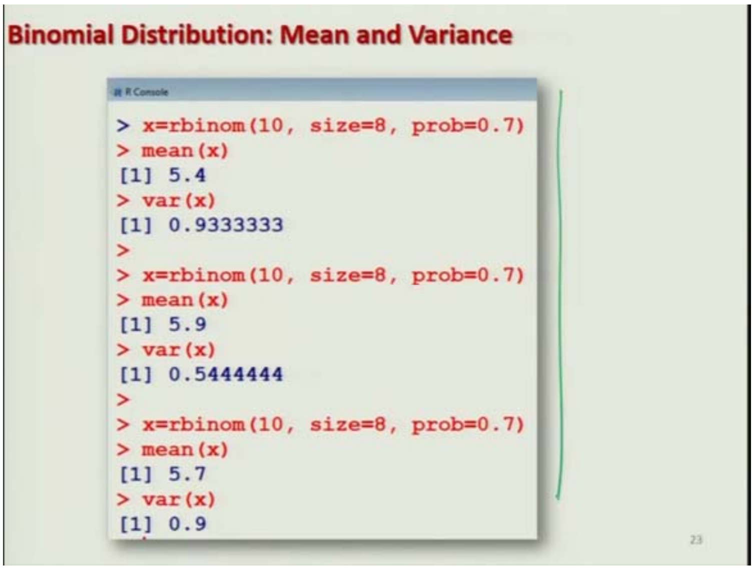
I already had done this thing in the case of discrete uniform. But, I want to repeat it so that I can show you. So, if you if you try to compute the mean and variance just based on that 10 number of observations. You can see this value is coming out to be 5.4 and variance is 0.933; which are you have to compare with the theoretical mean. And if you try to repeat, you can see that the mean is varying and the variance is also varying a lot.

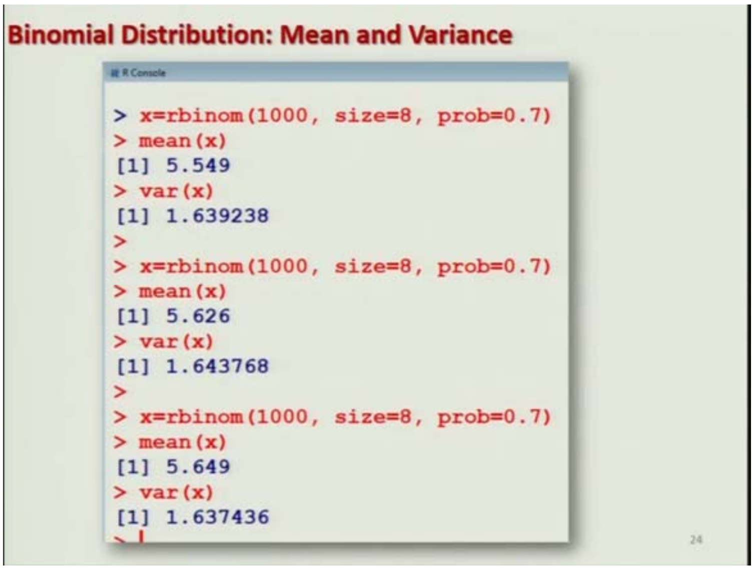
(Refer Slide Time: 29:09)



But, in case if you try to increase the number of observations, suppose I make it 1000 observations; and then I try to compute the mean and variance. You can see that these values are coming out to be more closer than with 10 observations. And even the the values that you are trying to compute in every repetition, they are also very close. You can see these are the value of mean and these are the value of variance. So, I would say that why do not you try and try to conduct it; you will get a different value.

(Refer Slide Time: 29:40)





And these are the screenshot of the same result which I have just shown you; so you can be confident that at least you will get the same value. At least I am getting the same value what I have reported and then you can try it yourself. So, now we come to an end to this lecture, and now I have given you in detail that how you can execute this binomial distribution in this R software.

So, now this is your turn try to take simple example from your assignments, books; try to solve them manually. And try to solve the same problem using the R software; that will give you more confidence. And then you can understand what type of probabilities is binomial distribution is trying to give you. And how are you going to use it; well this lecture was in more detail. After that I am going to repeat the similar things for for other types of probability distribution. So, it is very important for you that you will revise and learn this lecture very carefully. So, you try to practice it and I will see you in the next lecture; till then good bye.