

Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M P Gururanjan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 39
Bernoulli trials and binomial distributions

Welcome to Dealing with Materials Data. In this course we are trying to understand the collection, analysis and the interpretation of a data from materials science and engineering. We are in the third module which is on probability distributions and we are going to begin with some discrete probability distributions, specifically we are going to talk about Bernoulli trials and the Binomial distributions.

(Refer Slide Time: 0:38)

Module: Probability distributions

Bernoulli trials and binomial distributions

IITB 2/8

Bernoulli trial

- Consider an ideal solution of A and B;
- Pick a random atom. Is it B? Answer is "Yes" or "No"; There are only two outcomes.
- What is the probability that the answer is "Yes"? x_B where x_B is the composition of the alloy. The probability is the same for each trial.
- Outcome from different trials are independent
- Bernoulli trial
- PMF of Bernoulli trial: $f(x_B, k) = x_B^k(1 - x_B)^{1-k}$ where $k \in \{0, 1\}$ (0 is No and 1 is Yes)



Let us consider an ideal solution of A and B or random solid solution of A and B. If you pick a random atom, is it B? If you ask the question then the answer is either yes or no. In other words there are only two outcomes and if you are looking for B atoms and if you find a B atom then you can say that the outcome is a success and if you do not find a B atom you can say that the outcome is a failure.

So, but success and failure are, within quote marks, because suppose if your interest is not in B atom but in A atoms, then finding B atom will be considered as a failure and finding A atom will be considered as a success. Because there are only two outcomes, success is compliment to failure and failure is compliment to success.

And what is a probability that the answer is 'yes'. So, like I said it is a random solid solution and you pick random atom. What is a probability that it will be a B atom? That is basically given by the composition of the alloy. Suppose, if it is a 50 atomic percent alloy then the probability of finding the atom to be B will be 50 percent because we have assumed the random solid solution or we have assume that it is like a ideal mixture.

So, any random atom that you have pick, the probability that it will be of type B will be given by the alloy composition itself. So, x_B will be the probability and this probability is the same for any of trials. In other words, implicitly we are assuming that it is a large number of atoms from which we are picking some and so this process is not going to change the composition. Or we are

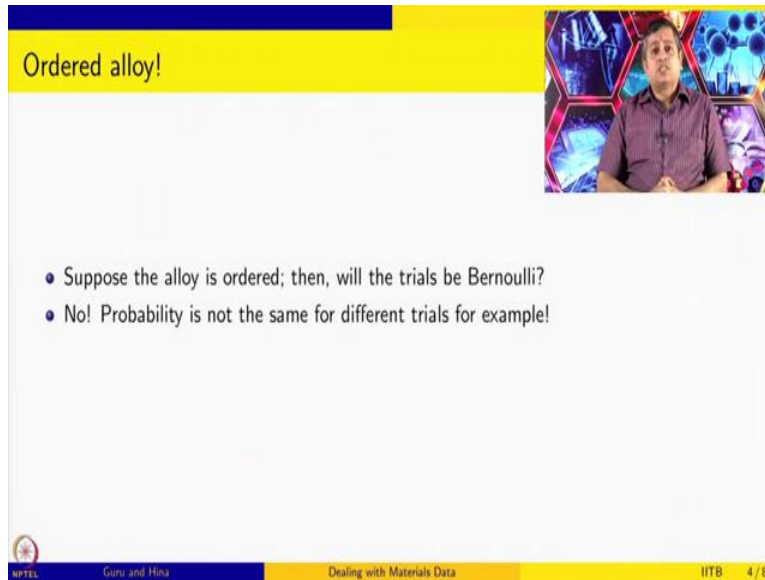
assuming that we are just probing and finding that it is either B or not and then we are going to leave that atom there.

So, it is not going to change our probability so any number of times you do this experiment. So, these are the or the assumptions that have gone in and we are also going to assume that the outcome from different trials are independent that if you have made one measurement it is not like the second measurement is going to be affected by your first measurement. So under these conditions so that there are only two outcomes and the probability of the success is some p and it remains same for all trials and different trials are independent.

$$f(x_b, k) = x_b^k (1 - x_b)^{1-k} \text{ where } k \in \{0,1\} \text{ (0 is No and 1 is Yes)}$$

Describes the result of every Bernoulli trial. If it is x_B , it is a B atom then the probability is x_B and k is 1 so $1 - x_B$ will have $1 - k$ in the exponent so because k is 1 that is 0, so that will give you 1 so you will get x_B . If it is not a B atom then because k is 0 x_B to the power 0 will become 1 so $1 - x_B$ to the power $1 - 0$, so it will become $1 - x_B$ so that is a probability of finding a non B atom in an alloy of composition x_B . So this is a Bernoulli trial.

(Refer Slide Time: 4:52)



Ordered alloy!

- Suppose the alloy is ordered; then, will the trials be Bernoulli?
- No! Probability is not the same for different trials for example!

NPTEL Guru and Hina Dealing with Materials Data IITB 4/8

Based on Bernoulli trial you can now ask the question suppose the alloy was not random, it is an ordered alloy say, let us say that there are specific sites which are occupied by A atom and there are specific sites which are occupied by B atom. Now, will the trials be Bernoulli? They will not be, because if you do different trials if you pick random atom depending on where the atom was picked from, depending on the site the probability of finding a B atom will be different for different sites. So, this cannot be considered as a Bernoulli trial if it is a random solution or if it is an ideal solution you can consider this as a Bernoulli trial.

(Refer Slide Time: 5:40)

Binomial distribution

- Suppose n independent Bernoulli trials are conducted, of which k trials are "successful" – that is, the atom was of type B: The result is a binomial distribution
- PMF of binomial distribution:

$$f(k; n) = \frac{n!}{k!(n-k)!} x_B^k (1-x_B)^{n-k} \quad (1)$$

where the ! is the factorial symbol.

- Mean value of binomial is nx_B
- Variance of binomial distribution: $nx_B(1-x_B)$
- Variance is proportional to n ; so, relative standard uncertainty goes as $\frac{1}{\sqrt{n}}$
- Large sample: more accurate

MPYEL Guru and Hina Dealing with Materials Data IITB 5/8

If you, now conduct such Bernoulli trials, suppose you conduct n independent Bernoulli trials of which k are successful what does that mean? You say try to look at some 100 atoms and you find that about 55 atoms are B., so the result of such an exercise of conducting n independent Bernoulli trials of which k are successful is known as a Binomial distribution. And the probability mass function of Binomial distribution, it depends on k how many successful trials you got and how many total trials you conducted which is n .

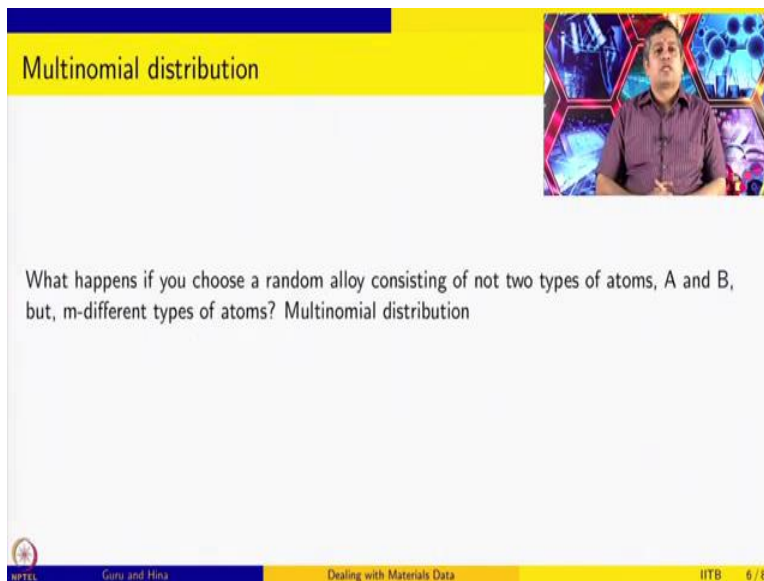
$$f(k; n) = \frac{n!}{k!(n-k)!} x_B^k (1-x_B)^{n-k}$$

The mean value of the Binomial distribution is n times x_B , whatever is the probability that you are assuming and variance of the Binomial distribution is n times x_B times 1 minus x_B . You can notice that the variance is proportional to n because of which if you take relative uncertainty that will go as 1 by square root of n because sigma will go as root n and you are dividing by n because that is the number of trials so you will get $\frac{1}{\sqrt{n}}$.

And this a very nice information to have because this shows you that if you do larger and larger number of experiments your uncertainty goes down, your relative uncertainty goes down because your variance is proportional to n . So, this basically gives you the assurance that you can improve your accuracy by doing large number of experiments. This is not surprising because if you take 10 atoms and find that some three of them are B.

And if you conclude that the composition is 0.3 because you had the probability as x_B and you did 10 and the mean value that you got is 3 so 3 divided by 10 will be 0.3 that will be the mean that you will decide, but if you do probably 100 maybe you will get 35, so you can get better accuracy maybe you do 1000 and then you get some 367 or something. So, you can improve the accuracy by doing larger number of experiments so that is what this gives this information on variance gives.

(Refer Slide Time: 8:37)



The slide features a yellow header with the text "Multinomial distribution". In the top right corner, there is a video inset showing a man in a purple shirt speaking. The main body of the slide is white and contains the text: "What happens if you choose a random alloy consisting of not two types of atoms, A and B, but, m-different types of atoms? Multinomial distribution". At the bottom, there is a blue footer bar with the IITB logo on the left, the text "Guru and Hina" in the center, "Dealing with Materials Data" on the right, and "IITB 6 / 8" on the far right.

Suppose, if you choose a random alloy and let us say that consists of not two types of atoms but m different types of atoms such alloys are known and they are equi-molar multicomponent alloys, they are also sometimes known as high entropy alloys. In such cases what happens, again you can describe using a similar distribution function which is known as a multinomial distribution function, Binomial is for two and Multinomial is for more than 2. So, if you have m different types of atoms and sometimes these types of materials are made for example 20 percent of 5 different components you make an alloy, so in that case you will get the multinomial distribution.

(Refer Slide Time: 9:31)

Binomial distributions

- Common whenever we have n independent events, each with one of two outcomes – "Success" with probability p or "Failure" with probability $(1 - p)$
- Microstructure analysis and geometric probabilities: Let us say we pick random points from the microstructure that consists of two phases; the point is either in Phase 1 or in Phase 2 with the probability given by their relative area fractions
- Suppose you pick n components and check if they are in working condition
- ...



Now, Binomial distributions are common whenever we have n independent events and each one has two outcomes and the success is with probability p and failure is with probability $1 - p$ and these events are independent and the probability is also the same for every trial. Whenever this happens you will see that the result will be a Binomial distribution. In microstructure analysis and in the case of geometric probabilities, for example, this will again be Binomial because they are like Bernoulli trials.

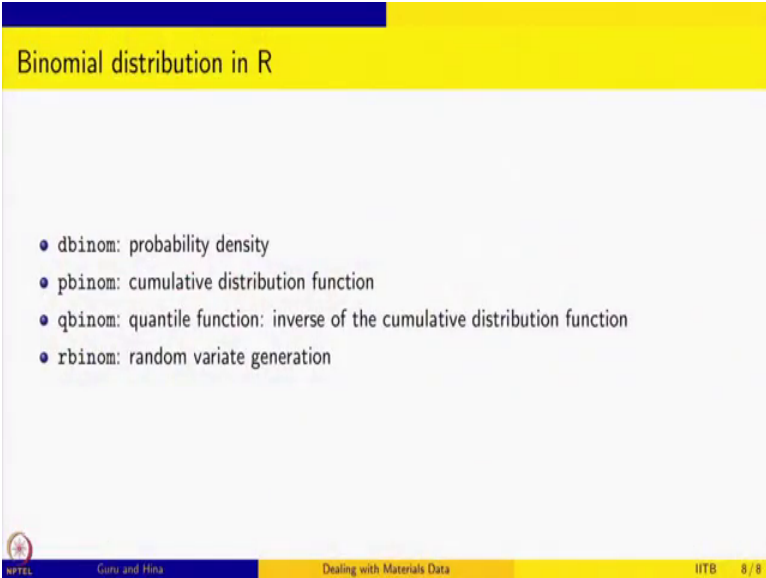
Suppose, you picked random points from a microstructure, let us say that, that consist of two phases, like the steel that we considered earlier which had phase 1 or phase 2, so if you pick random points and if those points happens to be either from phase 1 or phase 2, then you can think of the process as I pick a point is it phase 1 or phase 2, so the answer is 'yes' it is phase 1 or no it is not phase 1 in which case it is phase 2 and what would be the relative probability of this that would depend on the area fraction.

Because if phase 2 has larger area fraction and the random picking will be most of the times end up with phase 2. And if phase 1 has more area fraction you will. So, some amount of quantitative information about the microstructure you can get by doing exercise of this task. This also happens suppose if you have n components, identical components and if you pick and see if it passes quality test for example or if it is in working condition for example.

The answer again is 'yes' or 'no' and depending on the component that you are choosing and its process or property, it might or might not pass with a probability p and if you keep doing n such experiments and each one is independent and the probability does not change as you are doing your experiment for success then that will also be a Binomial distribution. So, in many, many different cases one comes across binomial distributions and this two are just two cases the random alloy picking and geometric probability in microstructure analysis are just two examples.

You might think that the random alloy example is a bit farfetched but we will look at a case and understand that it is not so, it is very relevant and we are going to look at one such microscopic technique where this has relevance or in any case. So, how do we deal with the binomial distributions using odd, this is a set of common theme that you will see that runs through the rest of this module.

(Refer Slide Time: 12:43)



Binomial distribution in R

- dbinom: probability density
- pbinom: cumulative distribution function
- qbinom: quantile function: inverse of the cumulative distribution function
- rbinom: random variate generation

NPTEL Guni and Hina Dealing with Materials Data IITB 8/8

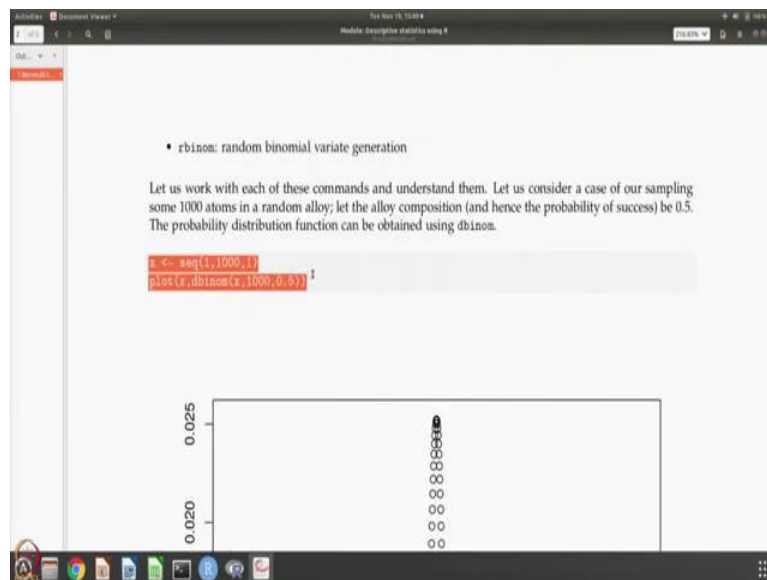
There are always a 4 commands in this case it is a binomial distribution, so binom is common and you have dbinom which is for probability density, pbinom which is for the cumulative distribution function, qbinom which is a quantile function which is a inverse of the cumulative distribution function, rbinom which is for random variate generation.

We have seen some of these commands earlier are rnorm for example we used to get a random variates from a normal distribution or the lnorm we used to get random variates from a log normal

distribution. So, we have seen some of these commands earlier and so we are going to do this, so it is for binomial so have dnom, pbinom, qbinom and rbinom.

And similarly for other distributions you will have this d, p, q or along with the distribution so that is how this commands work. So, let us now go to R and work with some of these commands. What do the probability distribution function and cumulative distribution function, the quantile function look like for binomial distribution and how do we generate random variates from the binomial distributions so that is what we will do now.

(Refer Slide Time: 14:00)



Let us go to R, so here is the first set of commands to use dbinom. So x is a sequence, so let us say that it goes from 1 to 1000 in steps of 1 and we want to plot x and we are going to get the binomial distribution density. And so there are 1000 is the n and 0.5 is basically the probability p.

(Refer Slide Time: 14:54)

The image displays two screenshots of the RStudio environment. The top screenshot shows the R console with the following text and code:

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"  
Copyright (C) 2019 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> x <- seq(1,1000,1)  
> plot(x,dbinom(x,1000,0.5))  
>
```

The right-hand pane shows the 'Values' window with the following content:

```
X      num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
```

The bottom plot shows a binomial distribution with a sharp peak at $x = 500$. The x-axis ranges from 0 to 1000, and the y-axis ranges from 0.0000 to 0.0200.

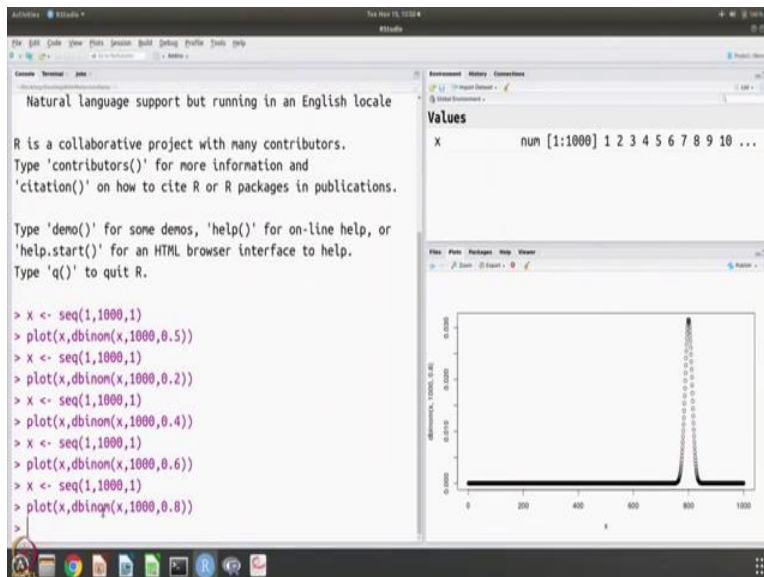
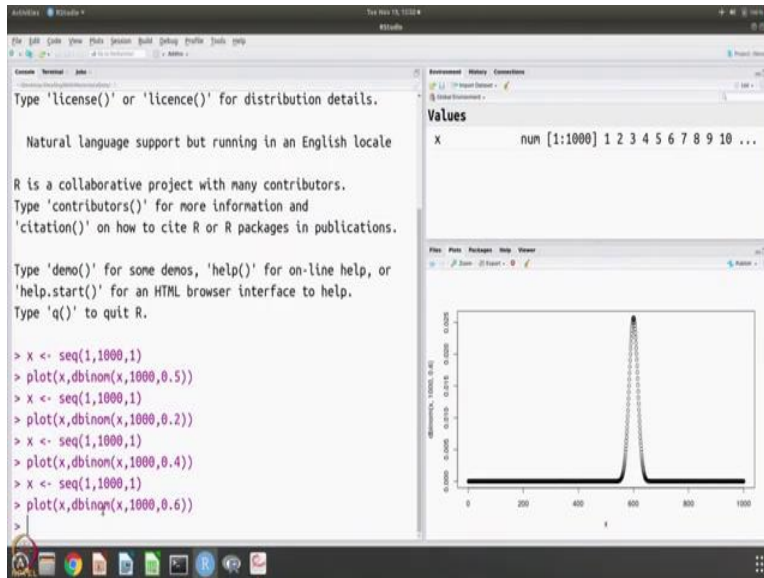
The bottom screenshot shows the R console with the following text and code:

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> x <- seq(1,1000,1)  
> plot(x,dbinom(x,1000,0.5))  
> x <- seq(1,1000,1)  
> plot(x,dbinom(x,1000,0.2))  
> x <- seq(1,1000,1)  
> plot(x,dbinom(x,1000,0.4))
```

The right-hand pane shows the 'Values' window with the same content as the top screenshot:

```
X      num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
```

The bottom plot shows a binomial distribution with a sharp peak at $x = 200$. The x-axis ranges from 0 to 1000, and the y-axis ranges from 0.0000 to 0.0200.



So, this is how the distribution function looks of course you can change the probability and you can look at how it changes so you can so it keep shifting from where the peak is going to be so this is the dbinom.

(Refer Slide Time: 15:22)

Document Viewer

Tue Nov 16, 10:24 AM

Module: Descriptive statistics using R

1

The cdf can be obtained using pbinom:

```
x <- seq(1,1000,1)
plot(x, pbinom(x, 1000, 0.5))
```

2

RStudio

Tue Nov 16, 10:24 AM

RStudio

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

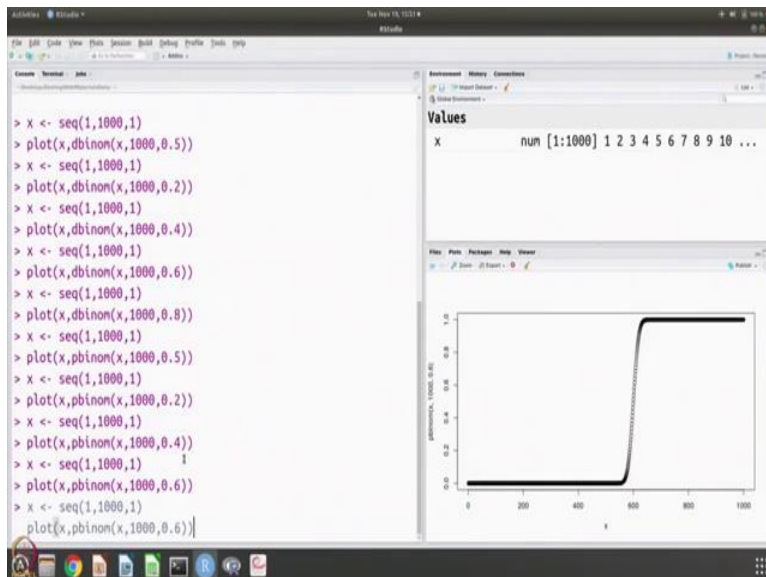
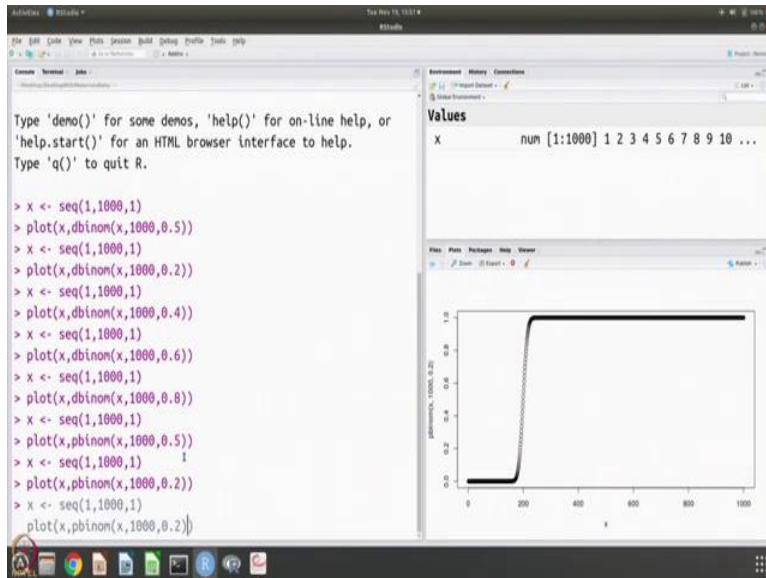
```
> x <- seq(1,1000,1)
> plot(x, dbinom(x,1000,0.5))
> x <- seq(1,1000,1)
> plot(x, dbinom(x,1000,0.2))
> x <- seq(1,1000,1)
> plot(x, dbinom(x,1000,0.4))
> x <- seq(1,1000,1)
> plot(x, dbinom(x,1000,0.6))
> x <- seq(1,1000,1)
> plot(x, dbinom(x,1000,0.8))
> x <- seq(1,1000,1)
> plot(x, pbinom(x,1000,0.5))
>
```

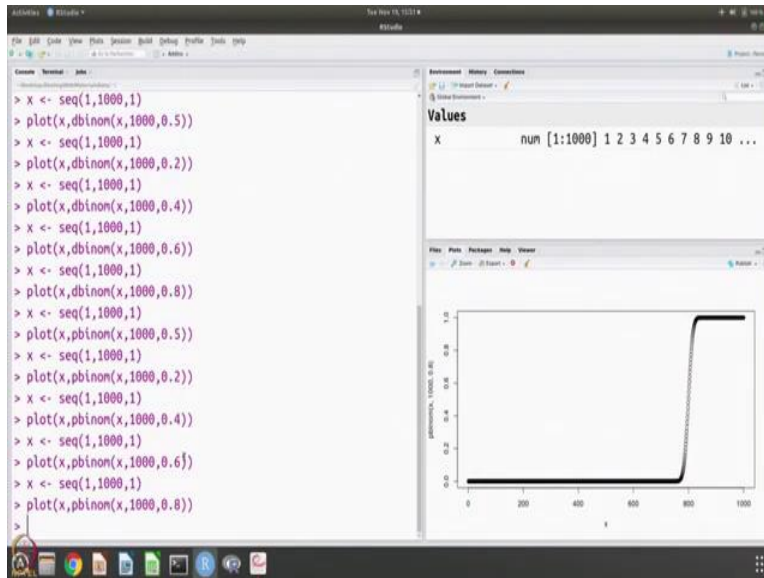
Environment History Connections

Values

x num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...

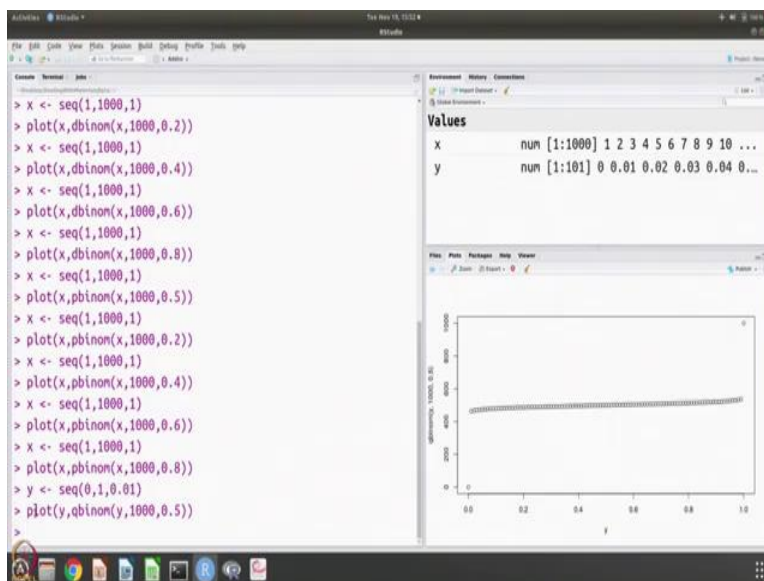
x	pbinom(x, 1000, 0.5)
1	0.0000000
2	0.0000000
3	0.0000000
4	0.0000000
5	0.0000000
6	0.0000000
7	0.0000000
8	0.0000000
9	0.0000000
10	0.0000000
...	...
495	0.0000000
496	0.0000000
497	0.0000000
498	0.0000000
499	0.0000000
500	0.5000000
501	0.5000000
502	0.5000000
503	0.5000000
504	0.5000000
505	0.5000000
...	...
995	1.0000000
996	1.0000000
997	1.0000000
998	1.0000000
999	1.0000000
1000	1.0000000

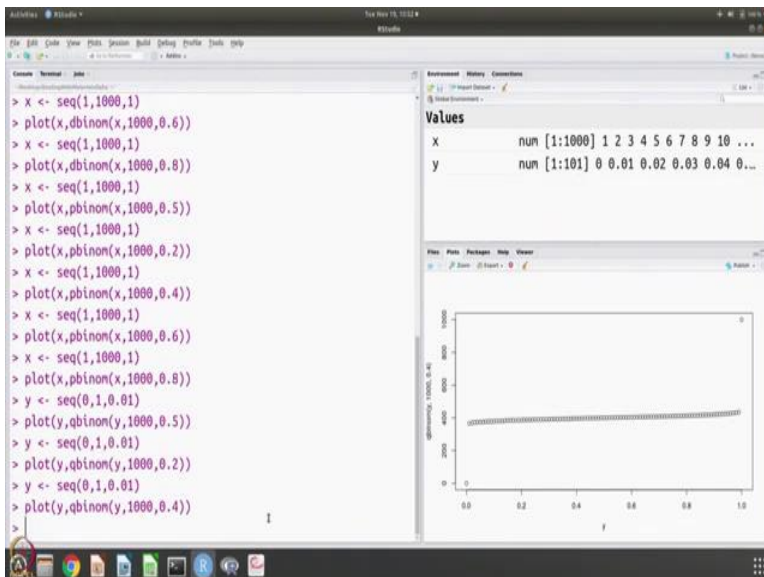
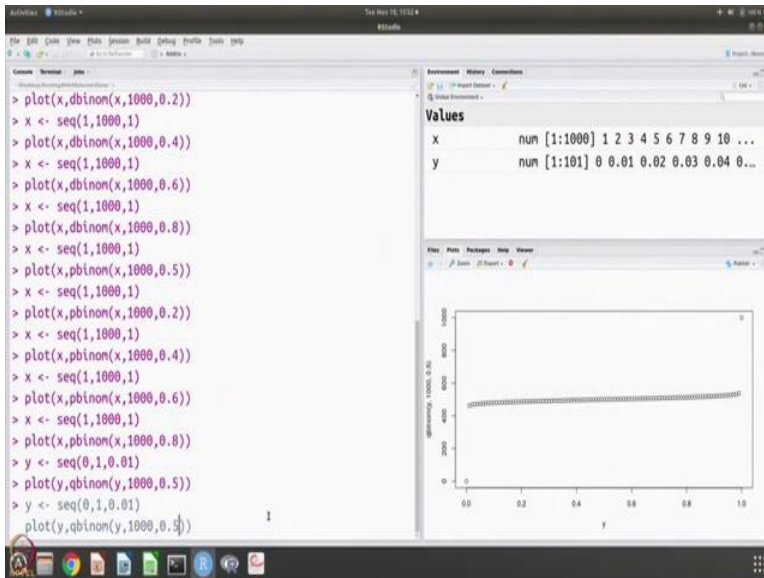


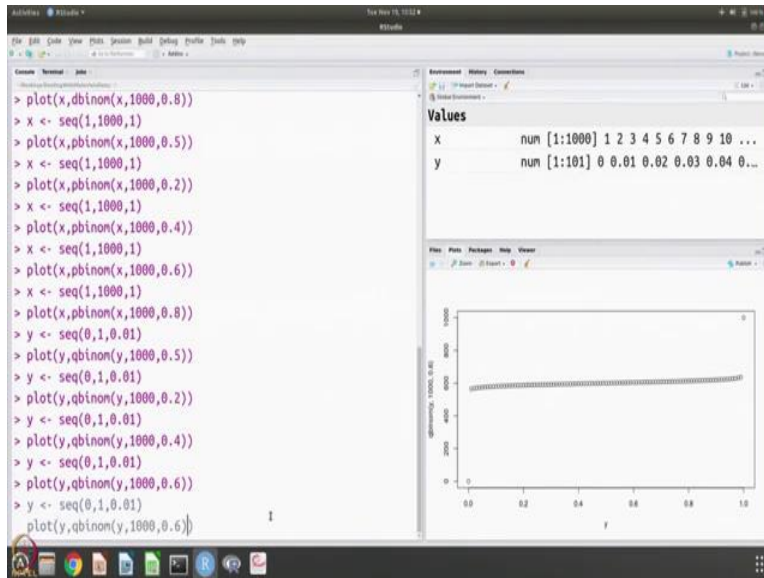


And so we can do the next one which is the pbinom. Again it is a same sequence so x is from 1 to 1000 in steps of 1 and now we want to plot x with the cumulative distribution function pbinom for when the probability of success is 0.5 so this is how it looks and of course if you shift then the probability distribution the cumulative distribution function shifts and so this is also expected. Now we want to get the quantiles but remember for quantiles the value should run from 0 to 1 because it is the inverse for this what is the x is? What we are trying to get?

(Refer Slide Time: 16:40)

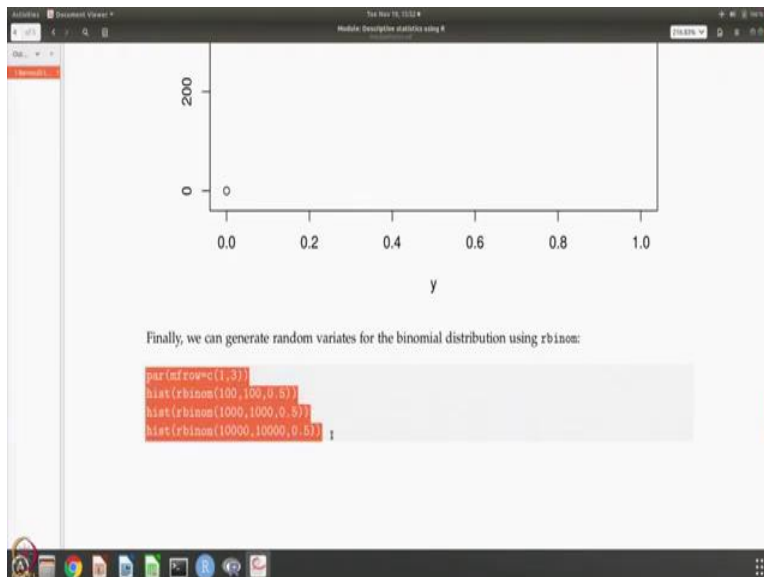


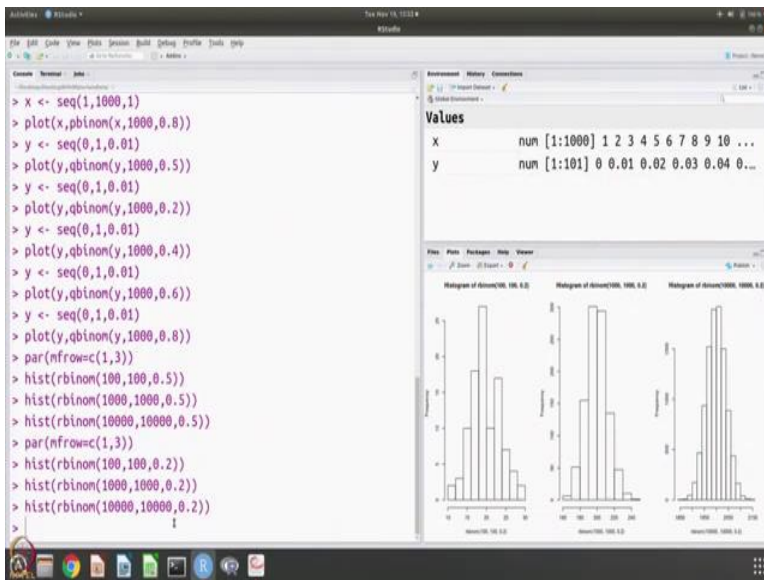
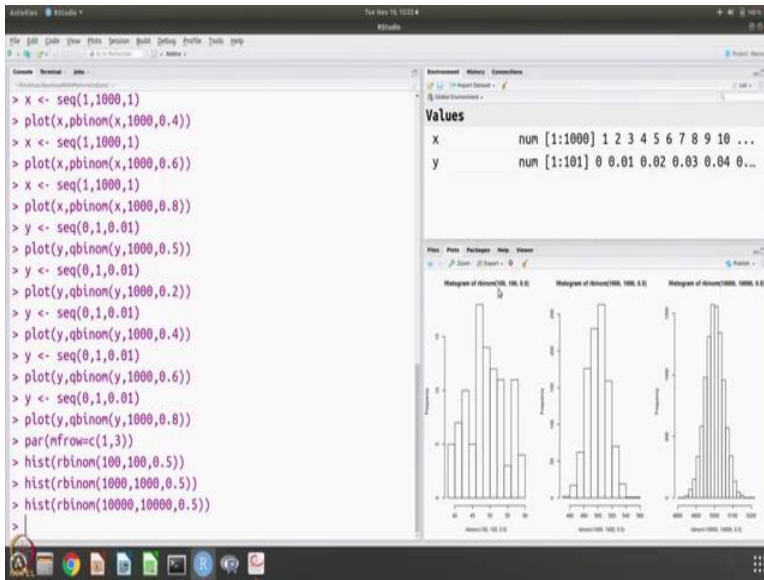


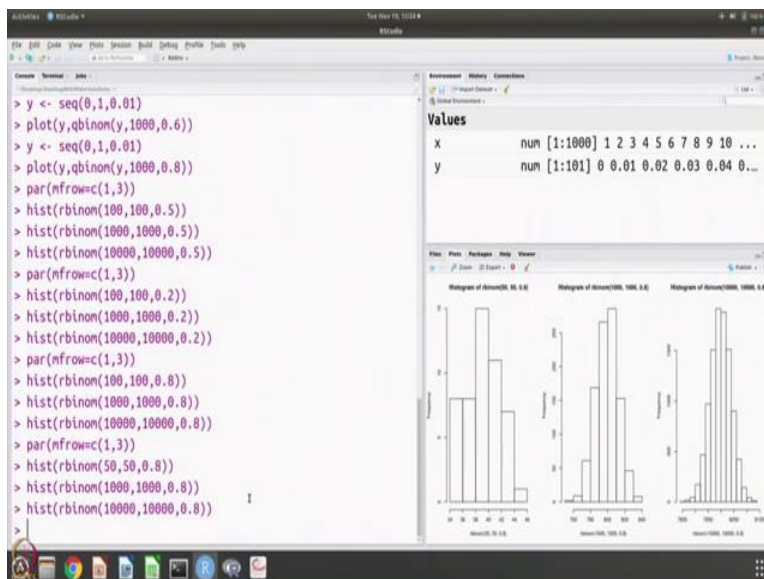
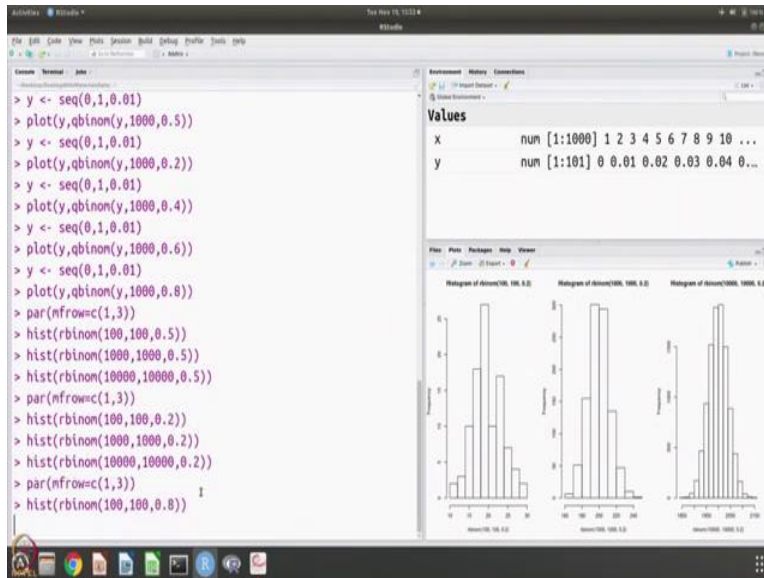


So, the x sequence that we should get should go only from 0 to 1, 0 to 1 in 0.01. So let us call that as y you should go from 0 to 1 in terms of 0.1 and we want to get y and this is qbinom and let us say for 0.5 so this is the probability the quantile function which is the inverse of the distribution function and of course you can get it for 2, you can get it for 4, can get it for 0.6, get it for 0.8 and so on. So, this is basically the inverse of the function that you earlier.

(Refer Slide Time: 17:16)







Finally if you want to generate random variates of course you can use that using rbinom and let us do that. So, I am going to make 3 plots and I am going to pick random variates 1, 100 or 1000 or 10,000, with 0.5 as the probability value and plot the histograms, so that is what you see this is the case when I picked 100, this is case where I picked 1000, this is the case where I picked 10000. So, this is with 0.5 then what happens if I do with 0.2? So this is what you get for 0.2. What happens for 0.8?

So, you can sort of see that for example 0.8 for small values skews like this and then for 1000 it is like this and when it comes to 10000 that is not making much of difference. So probably if you do maybe this as some 50 or something we can clearly see the difference let us see so you can see the

distribution how it changes with the more and more of random variates that you are generating or more and more number of times you are repeating this exercise.

So, we will come back to this there is an interesting theorem which we are going to look at later but this is just to show you how to deal with these distributions and work with them using R so we have just looked at Binomial and Bernoulli trials. We are going to go through each one of this distributions so it is a whole zoo of distributions that we have of which we have chosen only few.

There are very large number of distributions that are available and so we are going to deal with few of them as we go along which are of importance and relevance along with the information as to where they are useful or where they should be used and why? If possible. So that is what we are going to look at for the rest of this module on dealing with probability distributions using R thank you.