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

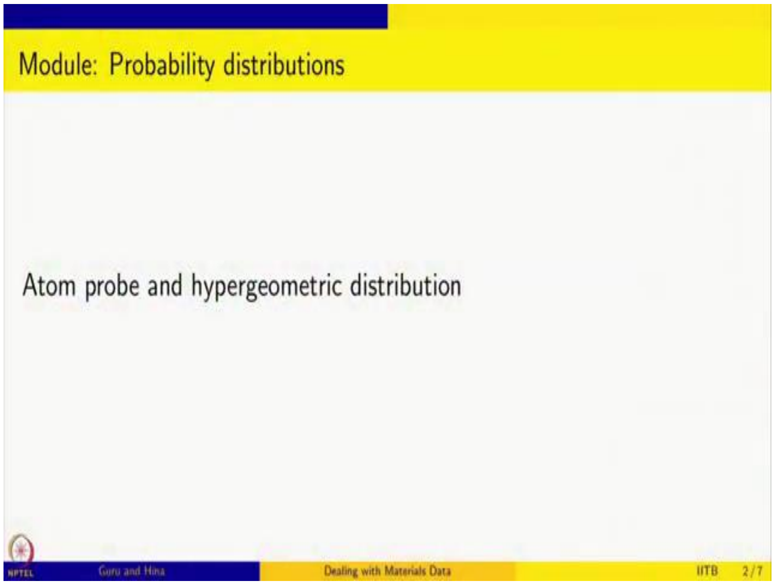
Lecture 41
Atom probe and hypergeometric distribution

Welcome to Dealing with Materials Data. In this course, we will learn about the collection analysis and interpretation of data from material science and engineering. We are in Module 3 and we are looking at probability distributions, specifically we are looking at discrete probability distributions. So far and we have looked at the Bernoulli trials, binomial and negative binomial distributions and we are going to continue with more discrete distributions and we are using a practical example where these distributions are important.

Specifically, it is a characterization tool it is a microscopic technique to find out the volumes at very small length scales of samples and this is known as atom probe and we are looking at the process of detecting atoms in atom probe technique and how we can calculate the error bars basically the uncertainties that come about knowing the process knowing what happens in the atom probe bar technique in terms of selection and detection of atoms.

Can we say something about the actual composition of the sample based on what we measure and say how much is there? So, there is a error in the process of both selection and detection and if we know them, can we say how much is the error in the actual sample in the composition that we give for the actual sample? So, that is the question that we are trying to answer.

(Refer Slide Time: 1:55)



Module: Probability distributions

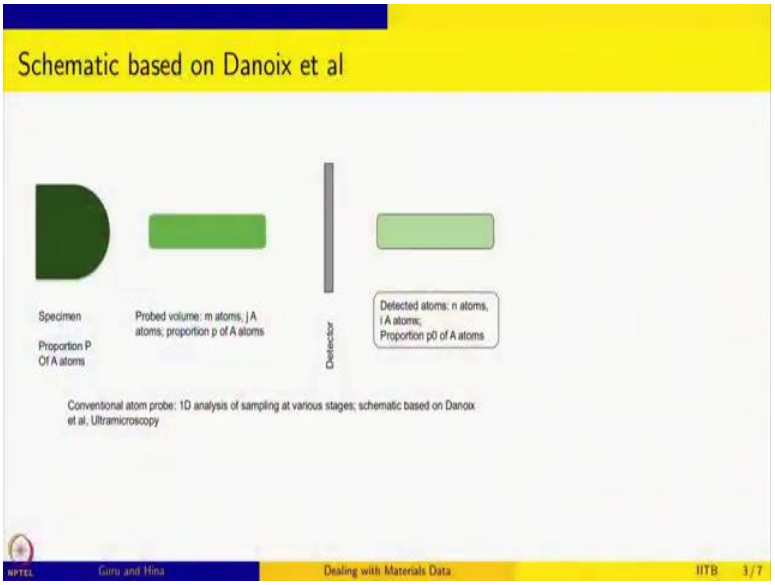
Atom probe and hypergeometric distribution

NPTEL Guru and Hina Dealing with Materials Data IITB 2/7

So, we will continue with atom probe and we will see that, we looked at the selection process and we came to the conclusion that it is binomial specifically it is negative binomial. So we said that, ok. So, suppose if you have to detect 100 atoms, how many failures will happen before you reach the target of 100s. So that is what the negative binomial distribution was and we saw that for a detected efficiency of 0.6, obviously, you will fail about anywhere between some 50 to 70 times before you actually reach 100.

So, that is what the negative binomial distribution showed and we, of course, calculated the cumulative distribution function, quantile function and so on and we also learned how to pick random variants from the distribution. So, if you want to simulate this process, for example, then it is important to pick random variants when you are simulating this process for the first stage to use the negative binomial distribution and calculate the values.

(Refer Slide Time: 3:10)



Schematic based on Danoix et al

Specimen

Proportion P
Of A atoms

Probed volume: m atoms, j A atoms; proportion p of A atoms

Detector

Detected atoms: n atoms, i A atoms; Proportion p0 of A atoms

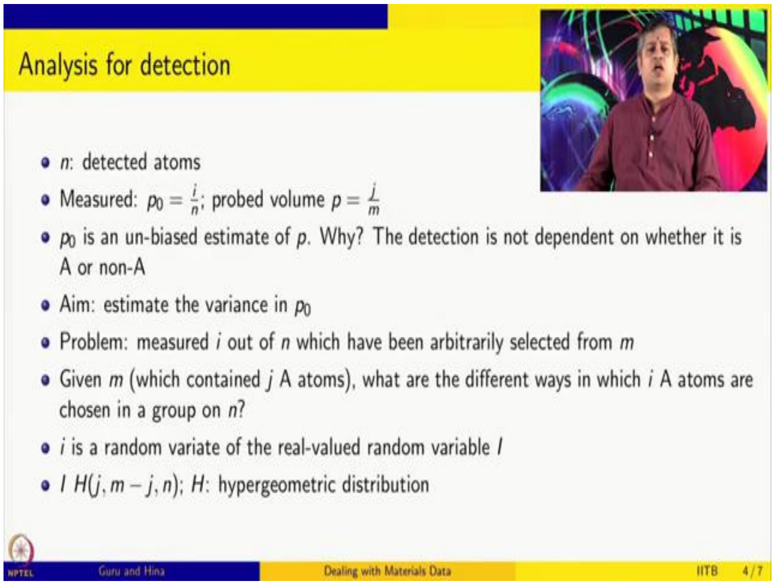
Conventional atom probe: 1D analysis of sampling at various stages; schematic based on Danoix et al, Ultramicroscopy

So, now we are going to look at the hypergeometric distribution and we will find out why it is relevant for the atom probe technique and so this is again just to remind you the schematics from Danoix et al. So, we have a specimen and the specimen has proportion of the atoms and we take a volume V of this specimen and we pull out the atoms from that volume. So, that is m atoms from volume V are pulled out of which j of them are of type A and so the proportion p of the atoms in this is j by m.

Now, these atoms are expected to fall on the detector and out of the atoms that fall on the detector n of them get detected out of which i of them are A atoms and we hence we able to calculate the proportion p_0 of the A atom. So, this is the actual measurements. You will say, ok detection happened for 100 atoms and the third 33 of them are of type A. So 33 by 100 is the composition that you would measure at this point and it will have its own error. And what we are saying is how much is the error in the actual sample composition that we give based on this value.

So, that is what we are trying to answer. So, we looked at this process of pulling out m atoms and we realized that these atoms when they fall on the detector, either they get detected or they are not detected and based on this binary process, given the detector efficiency, one has to think of negative binomial distributions to know how many failures will happen before you reach a success of a given number of atoms. So that is the negative binomial and now we are going to see how hypergeometric is relevant.

(Refer Slide Time: 4:52)



Analysis for detection

- n: detected atoms
- Measured: $p_0 = \frac{i}{n}$; probed volume $p = \frac{j}{m}$
- p_0 is an un-biased estimate of p. Why? The detection is not dependent on whether it is A or non-A
- Aim: estimate the variance in p_0
- Problem: measured i out of n which have been arbitrarily selected from m
- Given m (which contained j A atoms), what are the different ways in which i A atoms are chosen in a group on n?
- i is a random variate of the real-valued random variable I
- I $H(j, m-j, n)$; H: hypergeometric distribution

So let us say that the n is the number of detected atoms. So, the measured $p_o = \frac{i}{n}$ and that is but in the probed volume there are $\frac{j}{m}$. So, this is p and this is p_0 that we measure. So, what can we say about a p_0 and p, we can say that p naught is an unbiased estimate of p. Why because the detection is independent of whether it is an A atom or not an A atom irrespective of what is the type of atom the detector always detects that and just that it does not detect all atoms that fall on it.

It detects only a fraction of them, but that fraction is not a biased towards detecting for example, it is not like if 20 A atoms fall it will detect 18 and 20 B atoms fall it will detect only 10, if that happens then this proportion that we measure has no relevance to the probed volume or unless you know how much is the differences you cannot say anything about the composition in the probed volume.

However, in this case, because the detector is independent of whether the A or non A that is falling on it, it is going to detect with the given efficiency that efficiency remains the same. So, we can think of p_0 as an unbiased estimate of p. So, our interest is to estimate the variance in p_0, how much is the variance in p_0. So, then we know how much is the variance going to be in the value that you are giving for composition for this specimen.

So, it involves two variances, we have already seen there is a negative binomial and there is a variance associated with it and because m is itself as an estimate, it is not just a single number. So there is an estimate there and then we there is an estimate for p_0. So that is what we are trying to find out, what is the variance in p_0. So, the problem is you measured i out of n, and this n have been arbitrarily selected from m; arbitrarily selected because the detector does not have any favorites, it detects both A and non A with the same efficiency.

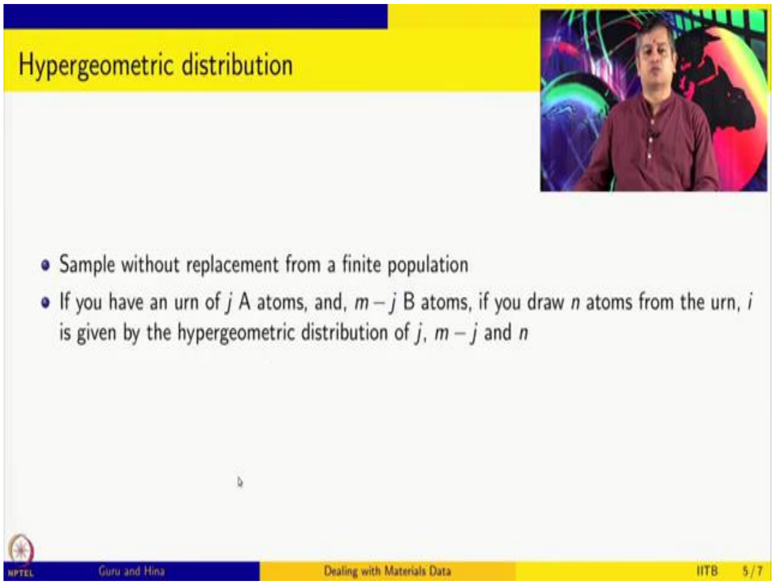
Now, given m which contained j A atoms, what are the different ways in which i A atoms are chosen in a group of the n. So, that is the question we are asking. So, we know that there were m atoms out of which n got detected in that n what are the different ways in which i of them happened to be A atom?

So, this is the question we are asking i is a random variate of the real valued random variable I, let us say capital I, then the I is a distribution and that distribution is called hypergeometric distribution

$$I\ H(j, m - j, n)$$

So, what is this hypergeometric distribution? Basically it answers this question given m which contained j A atoms, what is the difference? What are the different ways in which i atoms are chosen in a group of n. So that is what hyper geometric distribution is.

(Refer Slide Time: 8:13)



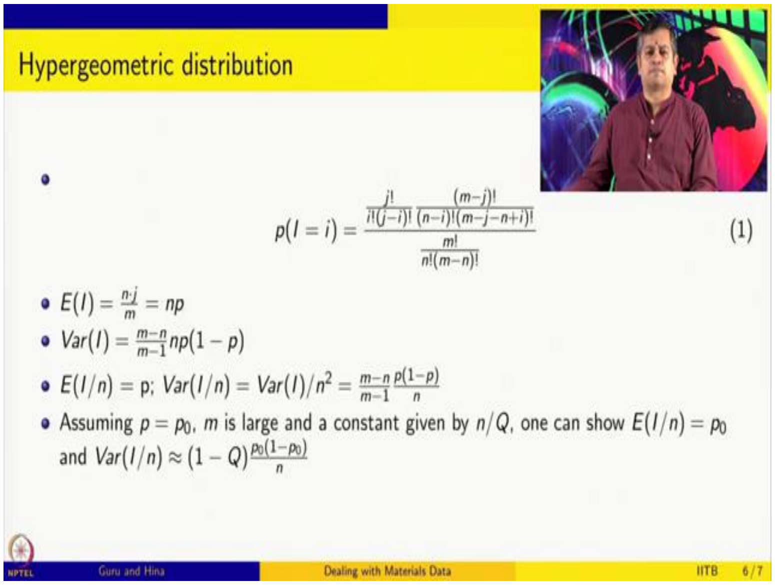
Hypergeometric distribution

- Sample without replacement from a finite population
- If you have an urn of j A atoms, and, $m-j$ B atoms, if you draw n atoms from the urn, i is given by the hypergeometric distribution of j, $m-j$ and n

It is called sampling without replacement from a finite population, ok. if you have an urn of j atoms and the m minus j B atoms, let us say or m minus j non A atoms and if you draw n atoms from this urn, then i of them or A, that is given by the hypergeometric distribution and the parameters are j, m minus j and n.

So, this is the hypergeometric distribution by definition, so it is a sampling without replacement, and it is a finite population. So because it is a finite population, every time you pull out an atom, if it happens to be A or B, then depending on that the further probability of picking another A atom will change. So it is a finite population.

(Refer Slide Time: 9:04)



Hypergeometric distribution

- $$p(I=i) = \frac{\frac{j!}{i!(j-i)!}\frac{(m-j)!}{(n-i)!(m-j-n+i)!}}{\frac{m!}{n!(m-n)!}} \qquad (1)$$
- $E(I) = \frac{n \cdot j}{m} = np$
- $Var(I) = \frac{m-n}{m-1}np(1-p)$
- $E(I/n) = \text{p}$; $Var(I/n) = Var(I)/n^2 = \frac{m-n}{m-1}\frac{p(1-p)}{n}$
- Assuming $p = p_0$, m is large and a constant given by n/Q, one can show $E(I/n) = p_0$ and $Var(I/n) \approx (1-Q)\frac{p_0(1-p_0)}{n}$

Guru and Hina | Dealing with Materials Data | IITB 6/7

$$p(I=i) = \frac{\frac{j! \qquad (m-j)!}{i!(j-i)!(n-i)!(m-j-n+i)!}}{\frac{m!}{n!(m-n)!}}$$

Remember this is the number of total number of atoms out of which n of them we are detecting and here we have j of them in that population to be A out of which i of them we are detecting and the m minus j of them are non A out of which we are detecting the n minus i have non A yet so, that is what this quantity is and so the expectation value for the hypergeometric

$$E(I) = \frac{n.j}{m} = np$$

$$Var(I) = \frac{m-n}{m-1}np(1-p)$$

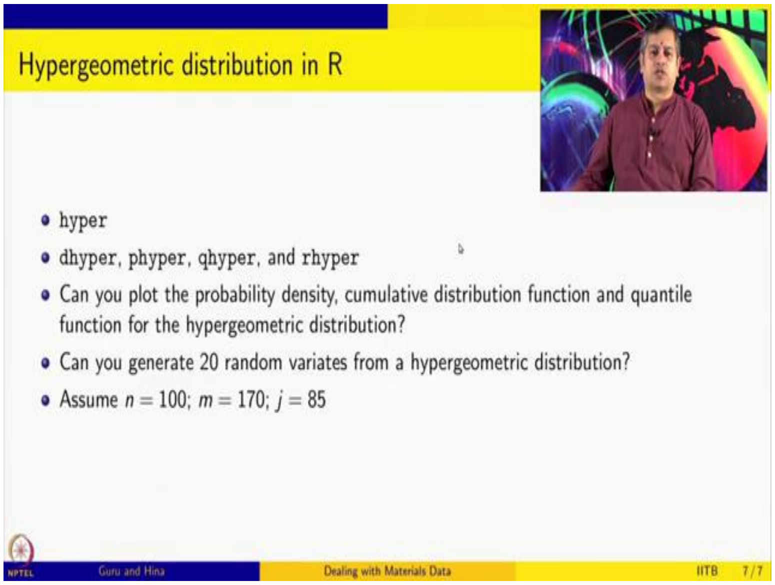
$$E\left(\frac{I}{N}\right) = p; Var\left(\frac{I}{n}\right) = \frac{Var(I)}{n^2} = \frac{(m-1)p(1-p)}{(m-1)n}$$

$$Assuming\ p = p_0, m\ is\ large\ and\ a\ constant\ given\ by \frac{n}{Q}, one\ can\ show\ E\left(\frac{I}{n}\right) = p_0$$

$$Var\left(\frac{I}{n}\right) \approx \frac{(1-Q)p(1-p_0)}{n}$$

Remember, we are making several assumptions and approximations we are our first assuming that the m is a constant, which is not true, it is only an estimate it cannot be constant number and we are assuming m is large which might be ok. So, you can do an experiment by pulling out large number of atoms and we are assuming p is equal to p_0 which is also a good assumption or approximation. If so, then the variance is one minus q p_0 into one minus p_0 by n. So, this is the hypergeometric distribution.

(Refer Slide Time: 11:46)



Hypergeometric distribution in R

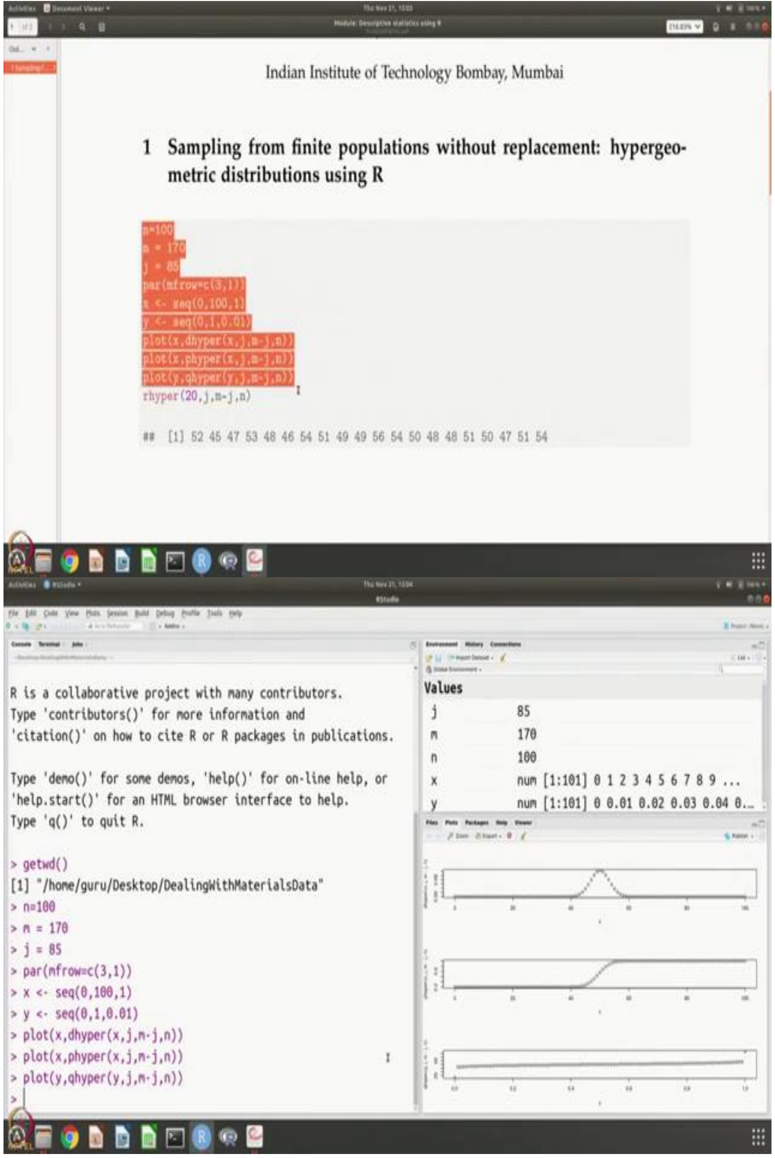
- hyper
- dhyper, phyper, qhyper, and rhyper
- Can you plot the probability density, cumulative distribution function and quantile function for the hypergeometric distribution?
- Can you generate 20 random variates from a hypergeometric distribution?
- Assume $n = 100$; $m = 170$; $j = 85$

NPTEL Guru and Hina Dealing with Materials Data IITB 7 / 7

So, let us take a look at hypergeometric distribution a little bit it how to deal with it in R. So, it is given by this command hyper. So, you have dhyper, phyper, qhyper and rhyper and these are mean for probability density, cumulative distribution, quantile function and this is for generating random variants and hypergeometric distribution has 3 parameters we saw.

So, let us say that we pulled out 170 atoms and let us say that 85 of them are actually of type A and we want to know if you detect 100 of them and then what is going to be the number of A atoms that they are going to have in this 100 i that is the value that is interest to us. So, that is what the hypergeometric distribution is going to give us. So, we will do as usual using R.

(Refer Slide Time: 13:13)



Indian Institute of Technology Bombay, Mumbai

1 Sampling from finite populations without replacement: hypergeometric distributions using R

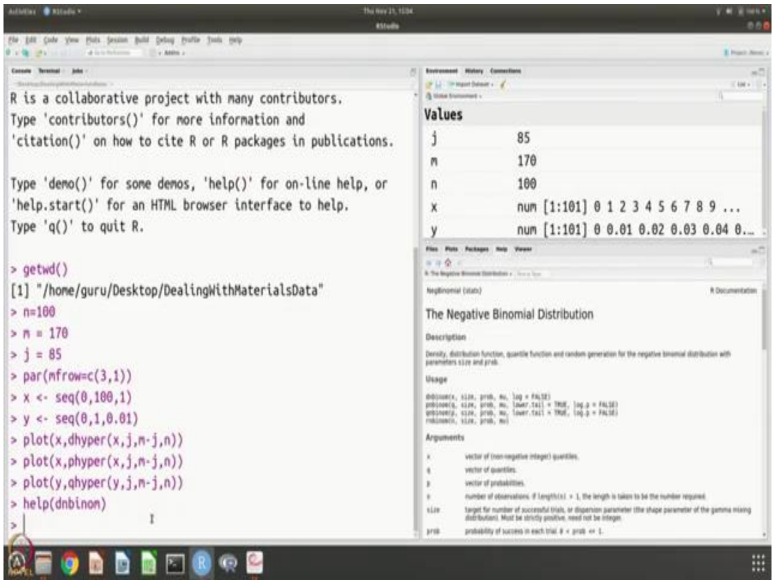
```
n=100
m = 170
j = 85
par(mfrow=c(3,1))
x <- seq(0,100,1)
y <- seq(0,1,0.01)
plot(x,dhyper(x,j,m-j,n))
plot(x,phyper(x,j,m-j,n))
plot(y,qhyper(y,j,m-j,n))
rhyper(20,j,m-j,n)

## [1] 52 45 47 53 48 46 54 51 49 49 56 54 50 48 48 51 50 47 51 54
```

```
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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
>
```

Values

j	85
m	170
n	100
x	num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y	num [1:101] 0 0.01 0.02 0.03 0.04 0...

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Jobs
Environment History Connections
Files Plots Packages Help Viewer



```
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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
>
```

```
Values
j          85
m          170
n          100
x          num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y          num [1:101] 0 0.01 0.02 0.03 0.04 0...
```

NegBinomial {stats} R Documentation

The Negative Binomial Distribution

Description

Density, distribution function, quantile function and random generation for the negative binomial distribution with parameters size and prob.

Usage

```
dnbinom(x, size, prob, mu, log = FALSE)
pnbinom(q, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
qnbinom(p, size, prob, mu, lower.tail = TRUE, log.p = FALSE)
rnbinom(n, size, prob, mu)
```

Arguments

x vector of (non-negative integer) quantiles.

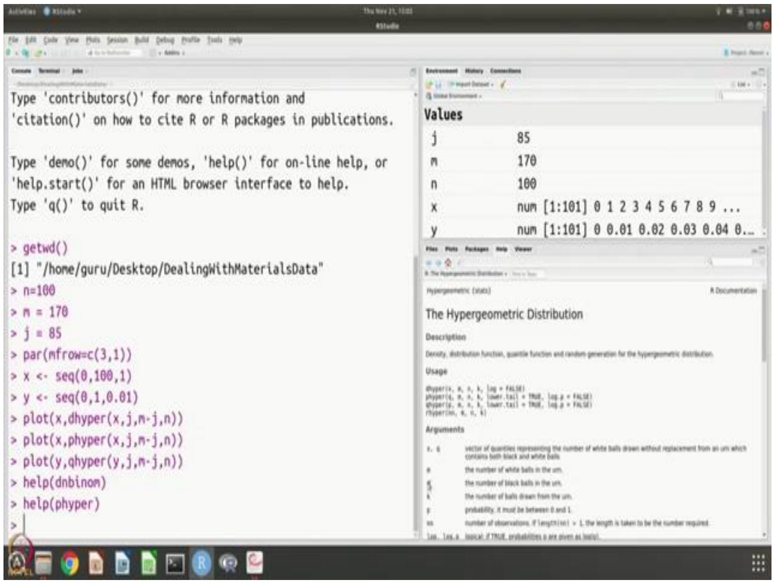
q vector of quantiles.

p vector of probabilities.

n number of observations. If length(n) > 1, the length is taken to be the number required.

size target for number of successful trials, or dispersion parameter (the shape parameter of the gamma mixing distribution). Must be strictly positive, need not be integer.

prob probability of success in each trial. 0 < prob <= 1.



```
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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
> help(phyper)
>
```

```
Values
j          85
m          170
n          100
x          num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y          num [1:101] 0 0.01 0.02 0.03 0.04 0...
```

Hypergeometric {stats} R Documentation

The Hypergeometric Distribution

Description

Density, distribution function, quantile function and random generation for the hypergeometric distribution.

Usage

```
dhyper(x, m, n, k, log = FALSE)
phyper(q, m, n, k, lower.tail = TRUE, log.p = FALSE)
qhyper(p, m, n, k, lower.tail = TRUE, log.p = FALSE)
rhyper(nn, m, n, k)
```

Arguments

x, q vector of quantiles representing the number of white balls drawn without replacement from an urn which contains both black and white balls.

m the number of white balls in the urn.

n the number of black balls in the urn.

k the number of balls drawn from the urn.

p probability, it must be between 0 and 1.

nn number of observations. If length(nn) > 1, the length is taken to be the number required.

log, log.p logical; if TRUE, probabilities p are given as log(p).

So, let us open R and the version is 3.6.1. So, we get the working directory. So, we are in the right directory. So, we can deal with up. So, this is a repetition sort the earlier plot that we made for the negative binomial distribution. So, n is equal to 100, m is equal to 170, j is equal 85 that is given to us. So, we are again going to make a column of row of plots, 3 rows are there and so, 3 plots we are going to make and for x the sequences 0 to 100 in steps of 1, for y it is 0 to 1 in steps of 0.01.

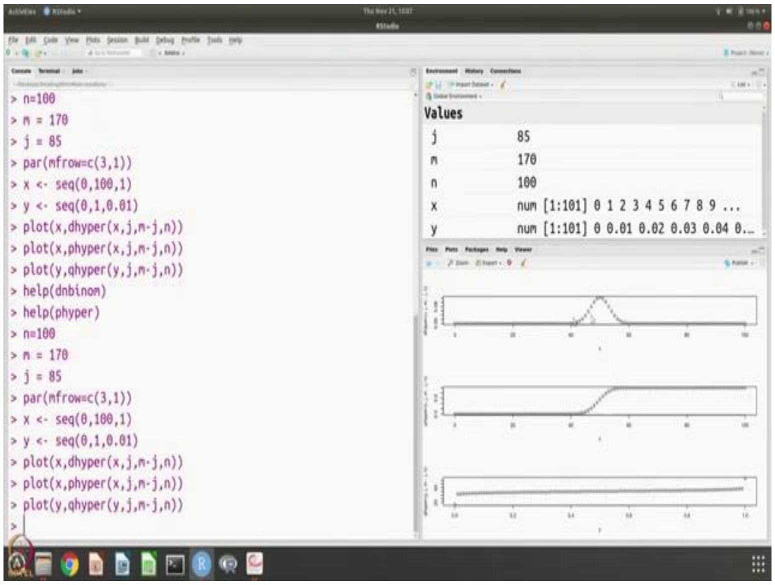
And so we are going to plot the probability density, probability mass function and cumulative distribution function and the quantile function and remember the dhyper, you have to give these parameters. So, you can see the distribution function, the cumulative distribution function that is the probability mass function and the quantile function. So, all 3 of them are plotted and of course, if you want to get help in any of these cases, so, for example, if you want to get negative binomial, so, you can say dnbinom.

So, it gives you and it tells you what are the things that you have to give to this function a vector of non-negative integer quantities you have to give that is why we give this x and q is a vector of quantiles. So, for calculating the quantile plot for example, we need to give the quantiles So, it has to be obviously between 0 and 1. So, that is why we have chosen and p is the vector of probabilities and so, that is the in given binom and n is the number of observations. So, that is something that we have we have not. So, we will use the n, but we want the random period how many of them we need right.

So, here we have to give the size probability and that there is also alternative characterization in terms of mean but we did not use for the dnbinom. Similarly, for hyper let us say phyper. So, we have to give m and k, x or q or p or things that you give here and what are these m and k. So, here it is described in terms of white and black balls in an urn, you can think of them as a and b for example, are you A and non A for example, number of white balls in the urn and a number of black balls in the urn, that is why we give m and the j and m minus j and k is the number of balls down from the urn. So, that is the n for us.

So, it is, you can think of this as the white or A; black or non A out of which how many of them we are pulling out of that is the number of balls drawn from this sample that is n and then this distribution gives how many of them are going to be the atoms of type A, hypergeometric distribution is used for sampling without replacement and these are the parameters and so, it gives you the information about what this function is giving.

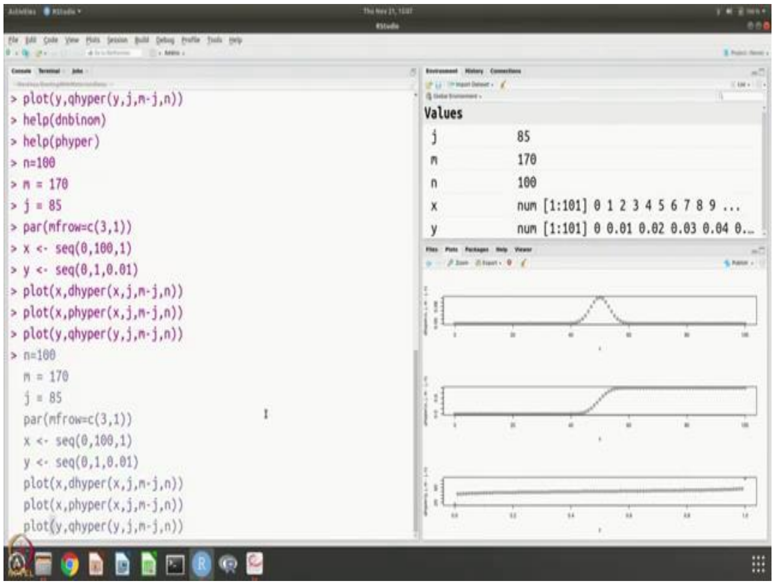
(Refer Slide Time: 17:11)



```
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
> help(phyper)
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> 
>
```

Values

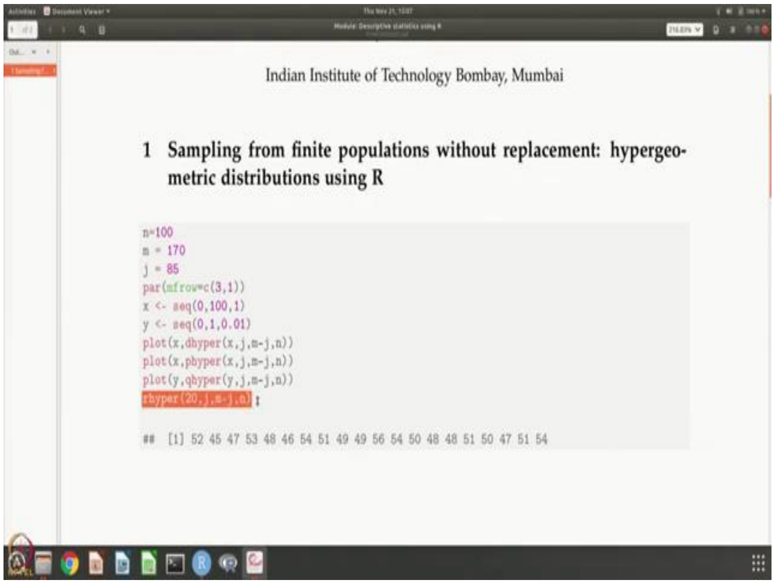
```
j         85
m         170
n         100
x         num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y         num [1:101] 0 0.01 0.02 0.03 0.04 0...
```



```
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
> help(phyper)
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> n=100
  m = 170
  j = 85
  par(mfrow=c(3,1))
  x <- seq(0,100,1)
  y <- seq(0,1,0.01)
  plot(x,dhyper(x,j,m-j,n))
  plot(x,phyper(x,j,m-j,n))
  plot(y,qhyper(y,j,m-j,n))
```

Values

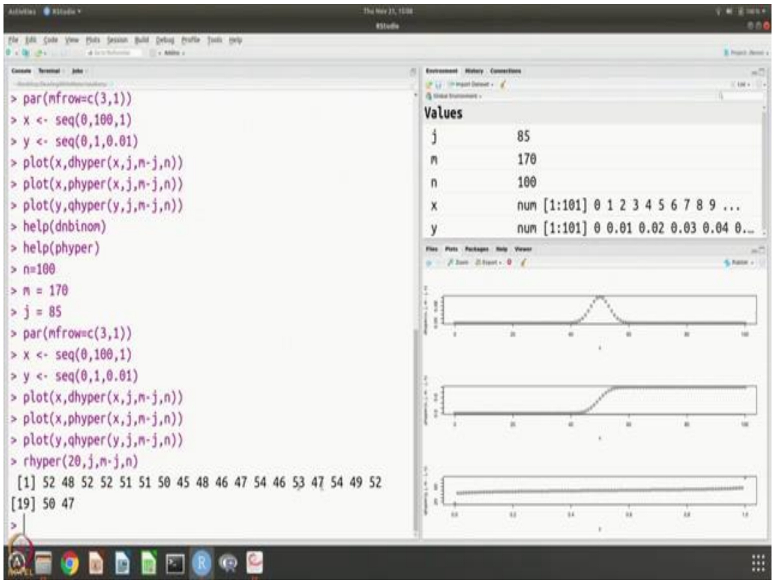
```
j         85
m         170
n         100
x         num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y         num [1:101] 0 0.01 0.02 0.03 0.04 0...
```



Indian Institute of Technology Bombay, Mumbai

1 Sampling from finite populations without replacement: hypergeometric distributions using R

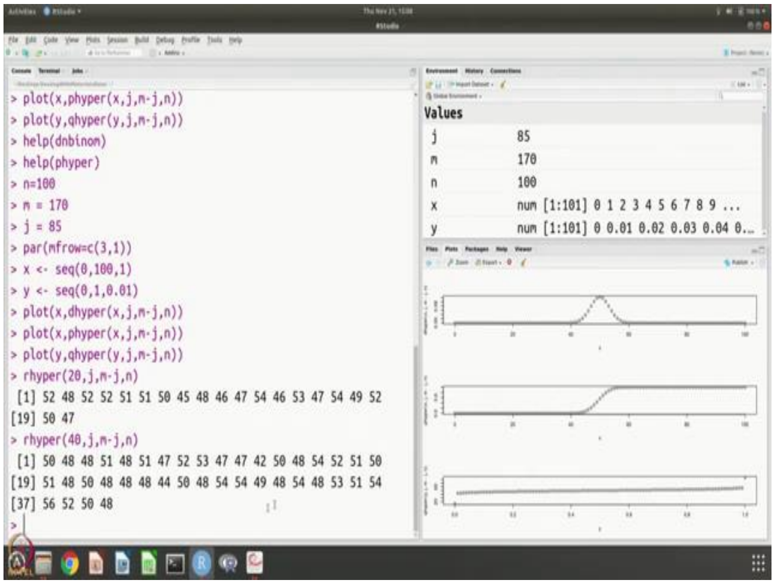
```
n=100
m = 170
j = 85
par(mfrow=c(3,1))
x <- seq(0,100,1)
y <- seq(0,1,0.01)
plot(x,dhyper(x,j,m-j,n))
plot(x,phyper(x,j,m-j,n))
plot(y,qhyper(y,j,m-j,n))
rhyper(20,j,m-j,n)

## [1] 52 45 47 53 48 46 54 51 49 49 56 54 50 48 48 51 50 47 51 54
```



```
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
> help(phyper)
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> rhyper(20,j,m-j,n)
 [1] 52 48 52 52 51 51 50 45 48 46 47 54 46 53 47 54 49 52
[19] 50 47
> 
```

Values

```
j     85
m     170
n     100
x     num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y     num [1:101] 0 0.01 0.02 0.03 0.04 0...
```



```
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> help(dnbinom)
> help(phyper)
> n=100
> m = 170
> j = 85
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dhyper(x,j,m-j,n))
> plot(x,phyper(x,j,m-j,n))
> plot(y,qhyper(y,j,m-j,n))
> rhyper(20,j,m-j,n)
 [1] 52 48 52 52 51 51 50 45 48 46 47 54 46 53 47 54 49 52
[19] 50 47
> rhyper(40,j,m-j,n)
 [1] 50 48 48 51 48 51 47 52 53 47 47 42 50 48 54 52 51 50
[19] 51 48 50 48 48 48 44 50 48 54 54 49 48 54 48 53 51 54
[37] 56 52 50 48
>
```

```
Values
j      85
m      170
n      100
x      num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
y      num [1:101] 0 0.01 0.02 0.03 0.04 0...
```

And so, this is what we found. So, when we do this plots tell you that for example, if you pull 100 out of a sample which had 85 A and 85 non A, then you are willing to measure approximately about 50 and there is a distribution. So, and this is a cumulative distribution function, and this is a quantiles. So, we can also of course, get the random variants and from this distribution so, you can get random variants from the hypergeometric distribution.

So, let us do that. So, we are saying that, ok, let us generate 20 random variants from this distribution which has this parameter so we are saying that 85, 170 total, so 85 non A and then let us say 100 of them from this sample we take out, then what will be the random variants that you would get in such a distribution. So you get numbers like 52, 48, 52 so that is what we saw. So it is about 50.

It is a distributed and so it is closer to 52, 48, 52, 51, 50 and 45 and 46, 47 and so occasionally you have some 40, yeah. So, the lowest that you will get is probably about 46 here and you get to 54. So, you can generate more random variates and see what happens. So you can then let us say 40 and then you will see that so it is distributed and occasionally you will see numbers going off like 56 in this case for example, or 44, in this case for example.

So, you will have the distribution about 50 and it will show you. So, this is expected because remember the composition that we are saying is 50 percent and so about 50 percent is what is going to return and it is going to return values about 0.5. So, that is what we expect and that is what we see. So, we have now looked at the atom probe technique and there were two stages, selection stage and detection stage and we have looked at the statistics and at each stage one is negative binomial, the other one is hypergeometric distribution.

And we have found out because once you know what distribution they are, you know, what their variances are going to be, they can be represented in terms of the parameters that represent this distribution and knowing this variances of course, we have learned how to do the error propagation in the previous module. So, we can use this information we have now and tried to do the error propagation and we did discuss during our propagation last time.

For example, that how to deal with if it is independent how to deal with it is not independent and so, we will continue with a similar discussion for the atom probe technique, analyzing the errors are random variations that you see in your measurement and how does it contribute to the error in your measurement of the composition of the sample.

So, that is what we will do. So, we have given two examples of for two discrete distributions, the negative binomial and hypergeometric they are all based on Bernoulli process and binomial distribution. So, we will continue, in the next session we will try to calculate the variance in capital p in the sample the proportion of atoms knowing what is the proportion of A atoms in your detected number of atoms if you detected n atoms and you profound the proportion to be p_0 what can you say about the error in the composition that you will get for this sample?

So, that is the question that we will answer in the following sessions. Thank you.