

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 40

Atom probe technique and negative binomial distribution

Welcome to Dealing with Materials Data. This is a course on collection, analysis and interpretation of materials data. We have done two modules so far, introduction to R and descriptive statistics using R. We are in the third module, this is the module on random variables. We are looking at some special random variables and random variables are of two types discrete and continuous and some like uniform distribution could be both discrete and continuous.

So, we are looking at some discrete random variables. We looked at Bernoulli trials and Binomial distributions. We are going to continue with the discrete random variables and we are going to also see some practical applications. In the last session when we discussed Bernoulli trials and Binomial distributions we were talking about random alloy, it is an A-B alloy, it is a binary alloy and randomly picking atoms from the alloy and deciding whether it is B or not and how that can give you some information about the alloy composition is what we were discussing.

And there are actually microscopic techniques which do similar things and that is what we are going to discuss in this session. So it is a continuation of discrete random variables and Bernoulli trials and Binomial distributions.

(Refer Slide Time: 1:45)

Module: Probability distributions

Atom probe technique and negative binomial distribution

NPTEL Gauri and Hina Dealing with Materials Data IITB 2/8

Atom probe technique

- Atom probe: measure compositions at sub-nanometric length scales;
- Detector efficiency: less than unity. Hence, compositions are estimates and not actual values
- How to calculate the variance in the estimation?
- Danoix et al, Ultramicroscopy, 2007. Vol. 107, pp. 734 and 739. Two papers!
- First paper: 1D conventional atom probe

NPTEL Gauri and Hina Dealing with Materials Data IITB 3/8

So, we are going to talk about technique which is known as atom probe technique and how that leads to negative binomial distribution is what we want to discuss in this session. So, atom probe technique is a technique to measure compositions in sub-nanometric length scales, so it is really fine measurement of composition. And the composition measurement depends on detector efficiency and detector efficiency is less than 1.

That means that if there are some 10 atoms that you pull out from your sample and even if all 10 of them fall on the detector the detector does not recognize all 10 of 10, so it has its own efficiency,

so it detects only some fraction of the atoms that actually reach the detector because of the which the compositions that we measure are actually estimates and they are not.

If you pull out say 10 atoms and detector actually detects all the 10 and then if you know whether they are of A type, B type etc. then you can actually give the exact composition for that number of atoms that you are taking out from the sample because that does not happens so its estimate. And what we are interested, so we are going to look at one more thing that we discussed in when we were discussing descriptive statistics which is error analysis.

We are always interested in its very difficult to do any experiments without errors so it is a given that there will be errors and standard deviations. As long as we have control over the standard deviation we are ok, so in the last session also we discussed how the accuracy can be improved and knowing that if you do more experiments for example you can get better accuracy is a good thing to know.

In a similar fashion its ok, even if your measurements are only estimates but if you have an idea as to how much is the error, then you are ok with the measurement or the measurement is more useful to us. So, in this context we want to calculate the variance or the standard deviation in the estimation of the composition from the atom probe experiments so that is our interest and this session and one more session probably following this is based on the paper written by Danoix et al in in Ultramicroscopy in 2007.

There are two papers the first one is what I am going to discuss in greater detail but the second one is also equally interesting and make some interesting point and it has all the statistical analysis which is done very nicely, so I strongly recommend this paper for you to take a look at and we are going to discuss some aspects of the paper. The paper also shows you how simple things that we are learning or simple ideas that we are learning are really of great use in doing actual analysis of this type.

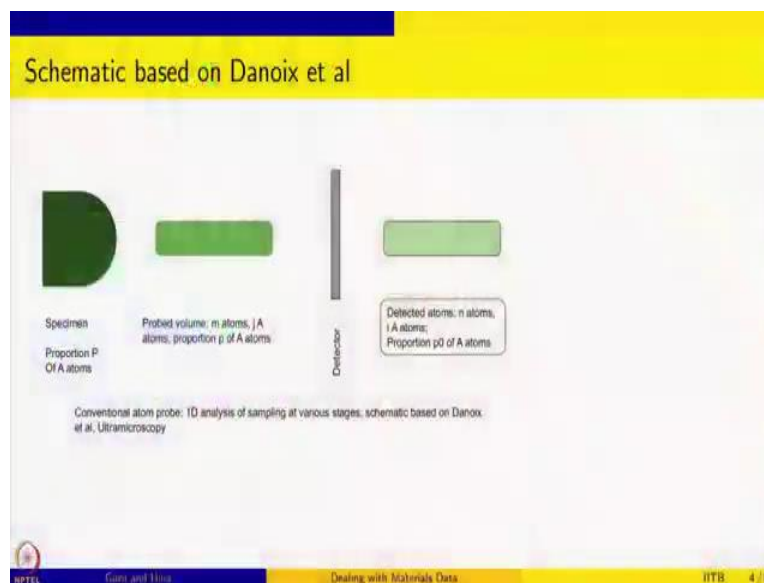
So, this is measurement of composition at sub-nanometric length scales so how accurately you can measure or how much is the error or what you can say about the error or how to improve your experiment and so this kind of things, so it is a very practical example which shows why we need to understand distributions because we started the session on probability distributions saying that

it is very important to understand experimental data because our understanding is that every experiment is actually probing this distribution and every result is actually a random variates from a distribution.

So, knowing the distribution is very-very important to understand how the error is or what the data is telling and so on and so forth. So, here is a nice example which uses some probability distributions, which also uses the ideas of error analysis and shows in a very practical scenario how this things are important and it has a very surprise ending, so I really like this example so I recommend that you actually take a look at this paper and try to read it on your own.

So, we are going to discuss the first paper which discusses what is known as 1D conventional atom probe. It describes the process by which in this experiment the atoms are selected and detected and we are going to translate this understanding into a statistical language and we are going to understand then the statistics that comes out of this and based on that we are going to make some calculations or analysis about the variance in this experiments.

(Refer Slide Time: 6:28)



So, what is the atom probe experiments? So, this is the schematic based on Danoix et al., they have a very nice schematic. So here you can see that there are 3 regions I have marked, so this is a sample from which we are going to pull out a portion and then that is going to these, these atoms

are going to fall on the detector or they are expected to fall on the detector out of which some of them are detected by the detector.

So, the specimen from which you are trying to pull out the atoms or which you are trying to study has proportion p of A atoms so we are interested in knowing the composition of A atoms in the specimen let us say and so we are interested in knowing A and not A, these are the only two cases we are interested in and if it is A then we are going to keep counting. So how many atoms are there for the total number of atoms that we pull out from this sample.

So, this is the specimen and the proportion of A atoms in the specimen is p . Now, the probe volume is consisting of m atoms, so from this sample we are going to pull out m atoms and out of which j atoms will be of type A. So, the proportion of A atoms in this m atoms that we pull out is p so this is the notation that we are using.

And then these m atoms that you take from the specimen are expected to fall on the detector but only n of them is detected by the detector and so i is the A atoms that is there in the detected sample or detected atoms. And the proportion is p_0 of A atoms in the detection stage. So, I have marked them in different shades of green to show that if you think of the proportion of A atoms, they need not be the same in all three stages.

So, it is important but we will later see in which cases it is the same and in which cases it is not the same or where we are making the approximation or assumption that it is the same, how reliable it is, is it ok and things like that. So, this is the detector that is shown schematically. So, this is from the paper of Danoix et al so we have p proportion of A atoms in the specimen out of which we pull out m atoms and j of them supposed to be A atoms so p is basically j by m .

So, that is the proportion of A atoms out of which n atoms get detected and i of them are A atoms, so i by n happens to be the p not is a proportion of A atoms. At this stage after the atoms are detected. Our interest is obviously in knowing the composition in the specimen and we are doing two things one is that we are pulling out A atoms and so this is called the selection process.

And we are going to assume that this is a random solid solution that is very important if it is not then the statistics has to be different and this we realized even last time, so you can you have to assume Bernoulli trial which means that you have to assume some independents of events and that

the probability does not change so on and so forth, those assumptions are not valid if it is not random solid solution.

And we also have to assume that the specimen that we are looking at is actually a representative volume of the material which we are looking at, if not also you will get wrong results. Suppose, if you pulled out a region which is extremely rich in the atoms and then you made this experiment, you will reach wrong conclusions about the proportion of the atoms in the material. So we are also assuming in addition that this is a random solid solution that the sample that we have pulled out is actually representative of the specimen that we are looking at.

So, from that we are taking out m atoms and we are trying to get and our interest, so that is the selection process and then there is a detection process and detection depends on the detector and its efficiency and so on. So, our interest is after measuring this so all that you will get at the end of the measurement is ok, we detected n atoms out of which i of them happen to be A atoms and then so you can calculate the proportion of A atoms in detection stage.

Based on that can we say something about the composition of the specimen and if we give the composition of the specimen, what is the error? That is the more important question that we are interested in answering.

(Refer Slide Time: 11:35)

The slide is titled "Analysis of 1D atom probe" and contains the following text:

- Random solid solution: assume
- Specimen consists of A and non-A atoms; the number of atoms in specimen is infinite; proportion of A atoms in the specimen is P
- Probed volume V contains m atoms; j atoms belong to species A; proportion of A atoms is p ; $j = mp$
- Out of m , n atoms will be detected by the detector; i belong to the species A; this proportion of A atoms $p_0 = \frac{i}{n}$
- Detector efficiency $Q = \frac{n}{m}$
- Detector efficiency $\approx 60\%$

The footer of the slide includes the IPTCL logo, the text "Guru and Hira", "Dealing with Materials Data", and "IITB 5/8".

So, let us do the analysis so how do we do that? Like I said we are assuming that it is a random solid solution and we are assuming that the specimen consisting of A atoms and non A atoms. We are also assuming that the number of atoms in the specimen is infinite, so this assumption of infinite will keep coming. What we mean is that compared to the number of atoms that we are taking out the total number of atoms that are there in the specimen is very large.

That is one way of understanding this infinity or this removal of m atoms that we are doing from the sample is not going to change the composition of the specimen too much so that will happen if you have to large number of atoms and if you are pulling out a small number of atoms from that. So, this is also an important assumption or approximation that we are making.

So as long as this is a good assumption that the specimen actually has more number of atoms compared to the number of atoms that you are taking out for your analysis, it will still be a good experiment and you will get reliable results. So, the proportion of A atoms in the specimen is p and probed volume is V that is the volume of the m atoms that we pulled out and j atoms actually belong to species A out of this m atom taken from this volume V , so proportion of A atom is p .

So, j is nothing but m times p . Out of m , n atoms will be detected by the detector, i belongs to the species A so the proportion of A atoms is p not which is i by n or i is n times p not like j is m times p . The detector efficiency is n by m , if there are m atoms that fall on the detector only n of them are detected, the efficiency of the detector is n by m .

Detector efficiency is approximately 60 percent and detector efficiency is not exact because how many of them are falling on m we have no idea, so the detector is also a binary process, it either detects an atom that is falling on it or it not detecting. If it is not detecting we do not know the atom actually fell on the detector and the detector fail to detect it or if it did not even fall on detector.

So we do not know this information and that is why the detector efficiency is approximate and this if we know these two numbers exactly this is exact efficiency but we do not know the numbers so that is one of problems and that is one of the reasons for uncertainty or the variance and we want to understand that and we want to exactly calculate what it is and that is the analysis of the only atom probe experiment that is done in this paper.

(Refer Slide Time: 14:37)

Analysis: conventional 1D atom probe

- Results of conventional atom probe – time ordered sequence of detected ions
- Evaporate atom; it hits the detector; if it is detected, it is counted; if not, we do not know if the atom has hit the detector or not! Detection probability is 60%
- Composition of the sampled volume: based on detecting n A atoms
- Given the detector efficiency, an estimate for m is $\frac{n}{Q}$
- Why m is an estimate? Detection process is binary; there is a finite probability for detecting all incident atoms to no incident atom
- Analogy: pick one atom from a random binary alloy; it is either B or not; if you pick large number of them, the result of all such experiments will give you the x_B
- m is a random variate of the (real valued) random variable M



So, you can think of the result of a conventional atom probe as a time ordered sequence of the detected ions. It is so the atoms are getting pulled out of the sample and there is a sequence in that, so there is a selection process and there is a detection process and we are keeping track of how many are getting detected and out of which how many are of type A. So, the process is to evaporate an atom and it hits the detector, if it is detected it is counted if not we do not know if the atom has hit the detector or not.

The probability however we know that the 60 percent of the atom that fall on the detector it detects, so approximately, so this is the number that we have. Now the composition of the sample volume is based on n , A atoms that are getting detected. So, that the composition of the sample volume, sample volume is decided based on detecting n atoms out of which i of them are of type A.

So, given the detector efficiency, so an estimate for m is n by Q because n by m is Q so m is nothing but n by Q , but m is only an estimate, it is not the exact number, this is because like I mentioned, detection process, detection process is binary so there is a finite probability for detecting all incident atoms and there is also a finite probability for detecting no incident atom. So, how do we understand this idea that it is an estimate?

Think of an analogy, so let us say that we pick one random atom from a random binary alloy, it is either B or not now if you pick a large number of them then the fraction of B atoms that you would pick would correspond to actually the composition. But there will always be an error so it is not

never going to be exactly, suppose if your alloy composition is 0.5 you can never assure that by picking let us say n atoms you will have B type n by 2 cases.

So, you might pick 10, may be 3 or 4 of them will be of type B., if you pick 100 maybe about 45 or 54 of them are B and maybe if you picked 1000 then you will get some 492 or 512 or something like that. So, as you go to larger and larger number of sampling that you do, you will see that value that you calculate goes closer to the actual value it is supposed to have, but there is always going to be an error in the process.

So, it is only an estimate it is not the exact number and that is what is being told here also so if you know exactly m atoms are falling and n of them are detected you can calculate the efficiency exactly but if you do not know how many atoms actually fell and you detected only n atoms then it is very difficult to say, exactly what it is but it will be a distribution, it will be and what is that distribution that is what we are going to find out.

So, m is a random variate because it is not one number so it has a variation, and its real valued because it is after all the number of atoms and let us say that the random variable that describes this so the distribution from which this is supposed to be random variates, let us call it as capital M . Remember our idea is that experiments actually give you sampling of a probability distribution, so the probability distribution is M here and out of which we are trying to pull out or we are trying to look at the realization of M or that random variable takes a value M .

(Refer Slide Time: 18:57)

The slide is titled "Negative binomial distribution" and contains the following content:

- A binary variable: detected or not detected
- Success with a given probability, (Q)
- Question: how many atoms should impinge on the detector for n atoms to be detected?
- Answer: Negative binomial with parameters (n, Q)
- Negative binomial: number of failures in a sequence of Bernoulli trials before a target number of successes is achieved
- $M \sim B^-(n, Q)$

At the bottom of the slide, there is a footer with the following text: NPTEL, Guru and Hina, Dealing with Materials Data, IITB 7/8.

So, it is a binary variable atoms are either detected or not detected and there is a probability of success, quote unquote, “success,” so in this case the success is that the atom is detected. Now, how many atoms are should impinge on the detector for n atoms to be detected? That is a question you can ask. And the answer is that, it is a negative binomial with parameters n that is how many successful results that you are getting and the probability of success.

And so negative binomial is the number of failure in a sequence of Bernoulli trials because it is a Bernoulli trial, remember that its either detected or not detected and the success of detection is always Q , it is not changing and different atoms getting detected or not detected is independent, so it is a Bernoulli trial and as the process goes on this is not going to change either so the detector is going to detect with the same efficiency.

Under that assumption you can see that it i a Bernoulli trial and what we are asking it that ok, “We know that number of success is known, n atoms are detected but we are asking how many failures happened before this success is achieved?”

$$M \sim B^-(n, Q)$$

In this case the number of atoms that get detected by the detector and Q is the probability, in this case it is the efficiency of the detector. So, this is the idea behind the negative binomial, so

remember we had discussed Bernoulli and Binomial now we are moving to negative binomial distribution.

(Refer Slide Time: 20:54)

The slide is titled "Negative binomial distribution in R". It contains the following content:

- $f(M = m) = \frac{(m-1)!}{(m-n)!(n-1)!} Q^n (1-Q)^{m-n}$
- Expectation of m : $\frac{n}{Q}$ (as expected)
- Variance $\frac{n(1-Q)}{Q^2}$
- `nbinom`
- `dnbinom`, `pnbinom`, `qnbinom`, and `rnbinom`
- Can you plot the probability density, cumulative distribution function and quantile function for the negative binomial distribution?
- Can you generate 20 random variates from a negative binomial distribution?
- Assume $n = 100$; $Q = 0.6$

At the bottom of the slide, there is a footer with the IITB logo and the text "Guru and Hina Dealing with Materials Data IITB 8/8".

$$f(M = m) = \frac{(m-1)!}{(m-n)!(n-1)!} Q^n (1-Q)^{m-n}$$

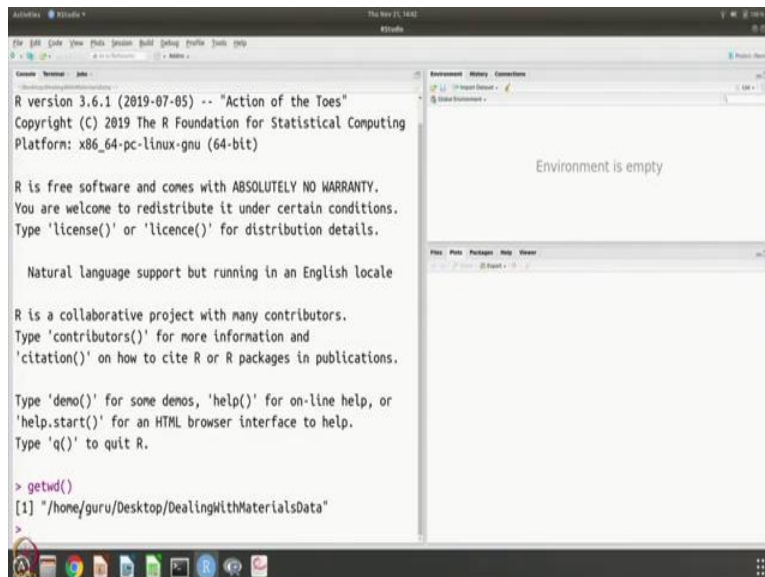
And as we discussed last time so you remember that it was `binom` for the binomial distribution and it is `nbinom` for negative binomial distribution. So, `dnbinom`, `pnbinom`, `qnbinom`, `rnbinom` are the commands and we know what they stand for, they stand for the probability mass function and this stands for the cumulative distribution function, this stands for the quantile function and this is to generate random variates from this distribution.

So, that is why the next question – Can we plot the probability density, cumulative distribution function and quantile function for the negative binomial distribution? Remember `pnbinom` and `qnbinom` are sort of inverses, `qn` is actually inverse to `pn` and the inverses are important to know the confidence intervals so we will come back to confidence intervals and discuss them later. But these are the things that we are interested in any distribution that we take.

And can we also generate 20 random variates from the negative binomial distribution so that is the question and obviously to calculate this values we need the parameters and for negative binomial distribution the parameters are n and Q , so assume n is equal to 100, Q is equal to 0.6 because we know that the probability of detection of the success rate is 0.6, so the probability is 60 percent of successfully detected.

So, the success in our case is given with the probability 0.6 and let us assume that we want 100 atoms to be detected so the question we are asking is, how many failures should happen before you detect 100? So that is given by the negative binomial distribution and so let us do this calculation using R.

(Refer Slide Time: 24:04)



```
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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
>
```

M P Gururajan and Hina A Gokhale
Indian Institute of Technology Bombay, Mumbai

1 Bernoulli trials and negative binomial distributions using R

```

n=100
Q = 0.6
par(mfrow=c(3,1))
x <- seq(0,100,1)
y <- seq(0,1,0.01)
plot(x,dnbinom(x,n,Q))
plot(x,pnbinom(x,n,Q))
plot(y,qnbinom(y,n,Q))
rnbinom(20,n,Q)

## [1] 61 55 55 80 54 43 68 62 48 81 87 60 55 65 53 87 78 66 61 76

```

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.

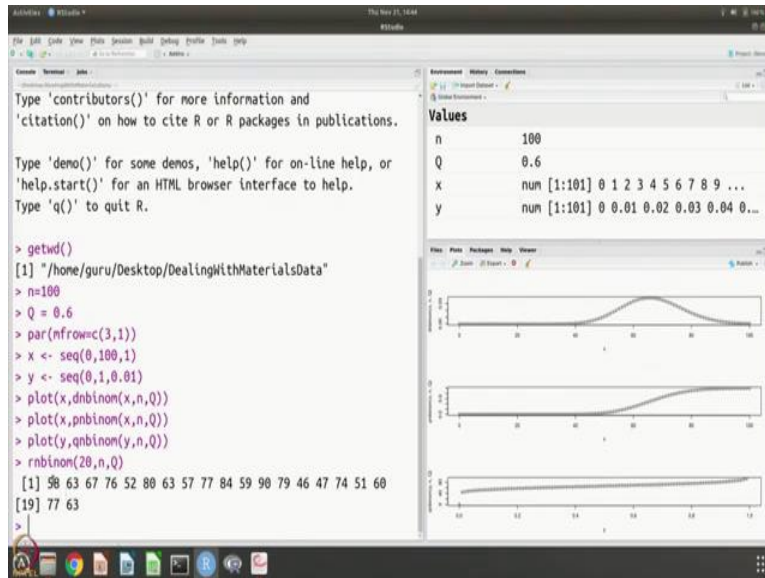
```

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> n=100
> Q = 0.6
> par(mfrow=c(3,1))
> x <- seq(0,100,1)
> y <- seq(0,1,0.01)
> plot(x,dnbinom(x,n,Q))
> plot(x,pnbinom(x,n,Q))
> plot(y,qnbinom(y,n,Q))
>

```

Values

| | |
|---|--|
| n | 100 |
| Q | 0.6 |
| 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... |



So, as usual we first have to check that we have the right version of R and it is a good to know what is the working directory so that is a working directory so now we want to so let us do this lets go through this script line by line. N is equal to 100 because that is the value for which we are calculating this negative binomial distribution and the Q which is the probability is 0.6 and by you know that this means that we are going to have a set of three maps, 3 plots and that is why it is 3 by 1.

So, we are going to have a three rows of plots and x is from 0 to 100 and y is from 0 to 1 in steps of 0.1 and first one is plotting the probability density or mass function and second one is to plot the cumulative distribution function and third one is of course to plot the quantile function. So, let us do that and you can see so this is the probability distribution function and so if you want to detect 100 atoms with 0.6 probability so you would expect about 65 atoms to be detected.

And so if you want to detect and this is the cumulative probability distribution so sometimes so this is gives you the added probability so at any x value so what is the probability that you are not exceeding x and the survival is 1 minus this and this is the quantile function, so we have done and how do we get the random variate of course that is also a simple command so we say.

So, we want to have random variates from the negative binomial distribution we want 20 of them and we know that n is 100 Q is 0.6 and you can see so 58, 63, 67, 73 and so on up to 63. So, this is the negative binomial distribution and so this is relevant for the atom probe and we can work

with R using the nbinom to deal with this function. So, we will come back and we will continue with the atom probe and the analysis of variance thank you.