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

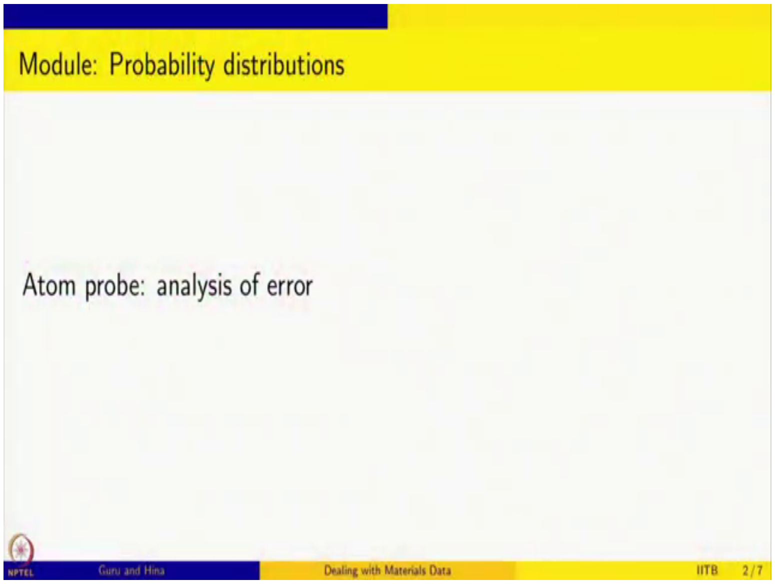
Lecture 42

Atom Probe: Analysis of error

Welcome to Dealing with Materials Data, we are looking at collection analysis and interpretation of data from material science and engineering. We are in the third module probability distributions using R and in this module we have discussed several discrete probability distributions and we have taken one practical example or a case study for the importance of these distributions where they occurred and which is the atom probe technique and we are also using this as an example of error analysis.

Because our aim is first to understand the process that takes place in atom probe and then find out the right statistics and knowing the information on the statistics in terms of variances, what can we say about the composition determination of the sample and its error, based on the measurements that we make in the atom probe experiments, so that is the question that we are trying to answer and so, in this session, we are willing to do the error analysis.

(Refer Slide Time: 1:25)



Module: Probability distributions

Atom probe: analysis of error

Guru and Hina
Dealing with Materials Data
IITB 2/7

So, in the previous two sessions, we showed that the selection process and detection process have negative binomial and hypergeometric distributions and knowing that those distributions, then it is easy to say what is the expected value and what is the variance. So, that is what we had done.

And in this we are going to put together all the analysis and we are going to talk about the error analysis, which is a follow-up on something that we had done earlier in the descriptive statistics session. So, we did talk about error propagation and how to do analysis and so on and so forth. So, we will continue in this session now to look at error propagation in the atom probe experiment.

(Refer Slide Time: 2:14)

Atom probe: conventional 1D

Specimen

Proportion P
Of A atoms

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

Detector

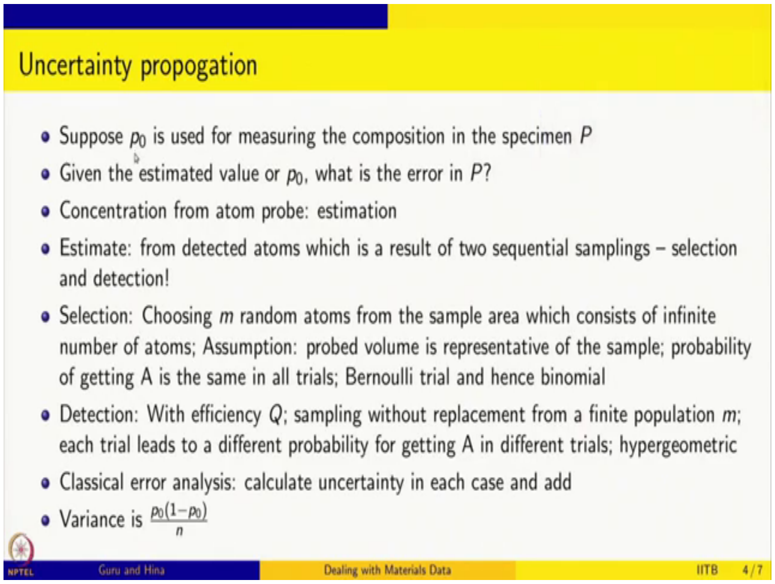
Detected atoms: n atoms,
i A atoms;
Proportion p0 of A atoms

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

Just to remind, we have a specimen and it has a proportion P of the atoms and we have probed a volume and we have pulled out the A atoms from a volume V out of which j of them are A atoms and so the proportion is P of A atoms and these m are expected to fall on the detector out of which n are detected and in the detected atoms i of them are A atoms and the proportion happens to be P_0.

And so there is a selection process and there is a detection process and in these two processes you get errors and you get the total error or total uncertainty in the quantity that you are trying to determine by adding up these errors. How do you add is the crucial question and that is what we are going to discuss.

(Refer Slide Time: 3:01)



Uncertainty propogation

- Suppose p_0 is used for measuring the composition in the specimen P
- Given the estimated value or p_0, what is the error in P?
- Concentration from atom probe: estimation
- Estimate: from detected atoms which is a result of two sequential samplings – selection and detection!
- Selection: Choosing m random atoms from the sample area which consists of infinite number of atoms; Assumption: probed volume is representative of the sample; probability of getting A is the same in all trials; Bernoulli trial and hence binomial
- Detection: With efficiency Q; sampling without replacement from a finite population m; each trial leads to a different probability for getting A in different trials; hypergeometric
- Classical error analysis: calculate uncertainty in each case and add
- Variance is $\frac{p_0(1-p_0)}{n}$

Guru and Hina Dealing with Materials Data IITB 4 / 7

So, p_0 if you say that is the composition in the specimen P. So, suppose we say that let us use this p_0 to measure the composition of the specimen, which is P. Now, given the estimated value of p_0 what is the error in P. So, because remember this p_0 as connected to this P to two steps. So, there is a error at every stage so you have to take that into account.

So, concentration from atom probe is only an estimation. It is an estimate because from detected atoms we are trying to say something about the composition of the specimen. But this detection is based on two sequential samplings, it is not single sampling. We are not just pulling out atoms and finding out what is a fraction, we pull out some atoms out of which only a fraction is detected.

So, there is a selection process and there is a detection process. Like we discussed earlier selection is choosing m random atoms from the sample area and the sample area is assumed to consist of infinite number of atoms. In other words, the m atoms that we pull out is very small compared to the total number of atoms in the sample area and the assumption is that probed volume is representative of the sample.

Again if this is not so, we will get wrong results if you do the analysis, but we are assuming that the probed volume is representative of the sample. So, whatever results we can get we can say is actually the composition of the sample or is reflect, reflects the composition or it estimates the composition of the sample. The probability of getting A is the same in all trials. So, it is Bernoulli trial.

So, we are pulling out random atoms from the given infinite atoms that are there and because the alloy is random solid solution, the probability of picking an A atom will depend on its composition. It is directly proportional to its composition, and so it is a Bernoulli trial, so it is a binomial. So this we have already seen that it is happens to be a negative binomial, but it is a binomial.

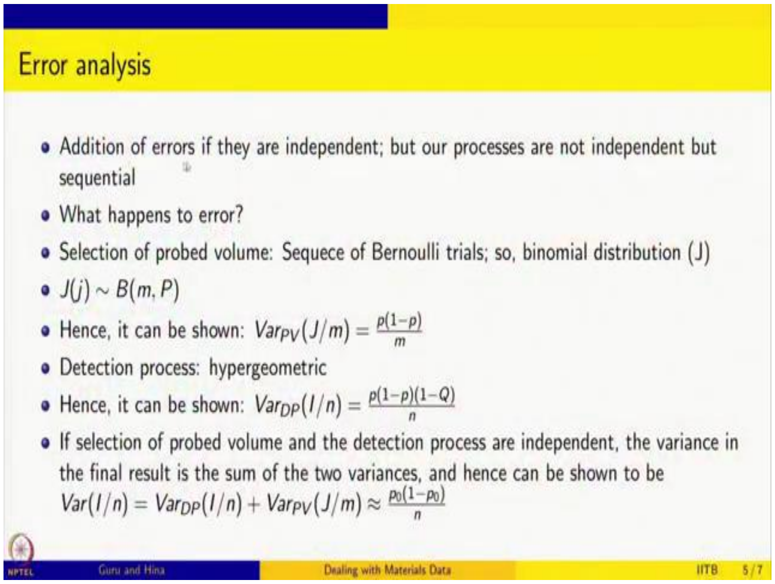
Now, the detection is done with an efficiency Q. But the detection process is sampling without replacement from a finite population m, and so out of m, each trial leads to different probability for getting A because we are having a finite population. Because of which this distribution again, it is based on Bernoulli it is detected or not detected and so on, except that it is a brief sampling without replacement from a finite population, so it happens to be hypergeometric.

So, we have binomial, we have hypergeometric and we know the variance in this case, we, we derived it in one of the earlier sessions, we know what is the variance and this we have derived in the previous session. So, classical error, error analysis is to calculate uncertainty in each case that is the variance in each case and add them up.

This was what was done and the variance was reported to be p_0 into 1 minus p_0 by n, but there is a problem because like we discussed in the error analysis sessions, you cannot add the errors unless you assume that the processes are independent and we can clearly see that in this case the processes are not independent.

Because you have to select the atoms only from the selected atoms you are going to make the detection. So, the selection is, the detection is conditional on selection. So, you cannot assume that these two are independent and so it is wrong to do this kind of classical error analysis where uncertainties are just added from Danoix paper, you understand that this what was were, this was what was done before they wrote their paper and tried to correct this. They said that look, this is not independent process, it is wrong to add them. So, let us do the proper analysis and what is the proper analysis?

(Refer Slide Time: 7:38)



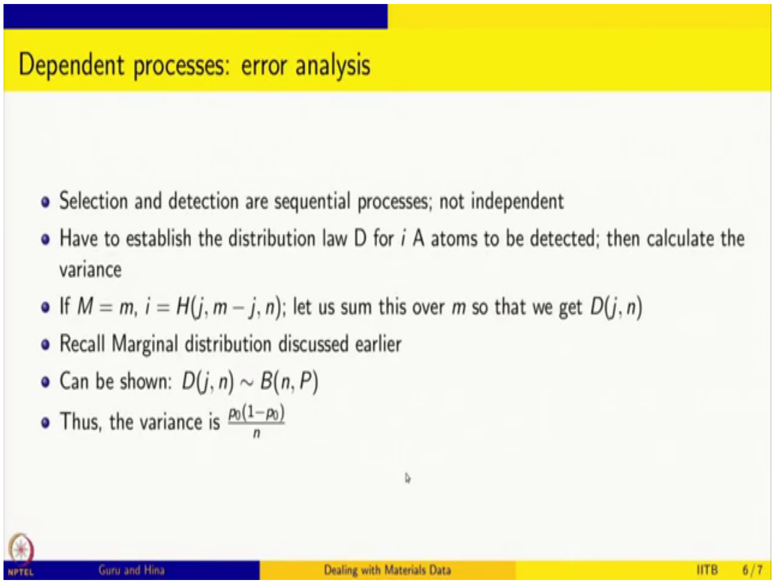
Error analysis

- Addition of errors if they are independent; but our processes are not independent but sequential
- What happens to error?
- Selection of probed volume: Sequece of Bernoulli trials; so, binomial distribution (J)
- $J(j) \sim B(m, P)$
- Hence, it can be shown: $Var_{PV}(J/m) = \frac{p(1-p)}{m}$
- Detection process: hypergeometric
- Hence, it can be shown: $Var_{DP}(I/n) = \frac{p(1-p)(1-Q)}{n}$
- If selection of probed volume and the detection process are independent, the variance in the final result is the sum of the two variances, and hence can be shown to be $Var(I/n) = Var_{DP}(I/n) + Var_{PV}(J/m) \approx \frac{p_0(1-p_0)}{n}$

Because the errors of not the processes are not independent the addition of errors is not correct. Ours is a sequential process. So, we should look at what happens to error. See, selection of probed volume is a sequence of Bernoulli trial, so it is a binomial distribution and it goes as binomial with some probability. So, the variance for this process is p into one minus p by m. Now, detection processes hypergeometric. So, you can show that the variance on detection processes P into one minus P into one minus q by n.

So, we know that two variances, what we should, if the selection of probed volume and the detection process are independent variance in the final result is some of two variances and so we can just add them up and if you add them up approximately you can show it to be p_0 into 1 minus p_0 by n and but because this is not correct, what should we do.

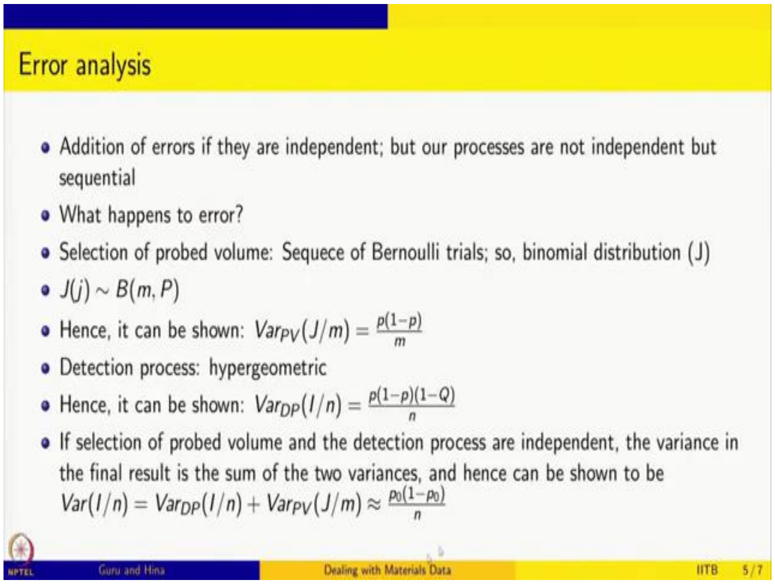
(Refer Slide Time: 8:39)



Dependent processes: error analysis

- Selection and detection are sequential processes; not independent
- Have to establish the distribution law D for i A atoms to be detected; then calculate the variance
- If $M = m$, $i = H(j, m - j, n)$; let us sum this over m so that we get $D(j, n)$
- Recall Marginal distribution discussed earlier
- Can be shown: $D(j, n) \sim B(n, P)$
- Thus, the variance is $\frac{p_0(1-p_0)}{n}$

Guru and Hina Dealing with Materials Data IITB 6 / 7



Error analysis

- Addition of errors if they are independent; but our processes are not independent but sequential
- What happens to error?
- Selection of probed volume: Sequece of Bernoulli trials; so, binomial distribution (J)
- $J(j) \sim B(m, P)$
- Hence, it can be shown: $Var_{PV}(J/m) = \frac{p(1-p)}{m}$
- Detection process: hypergeometric
- Hence, it can be shown: $Var_{DP}(I/n) = \frac{p(1-p)(1-Q)}{n}$
- If selection of probed volume and the detection process are independent, the variance in the final result is the sum of the two variances, and hence can be shown to be $Var(I/n) = Var_{DP}(I/n) + Var_{PV}(J/m) \approx \frac{p_0(1-p_0)}{n}$

Guru and Hina Dealing with Materials Data IITB 5 / 7

So, because they are sequential and not independent we have to establish a distribution law D for i A atoms to be detected. So, we have n atoms which are detected out of which i of them or of type A, we have to find out what is the distribution law D for i A atoms to be detected and calculate the variance based on that. If this random variable M results in the realization m, then the i is basically hypergeometric function, but we know that M is a variable, so it can change.

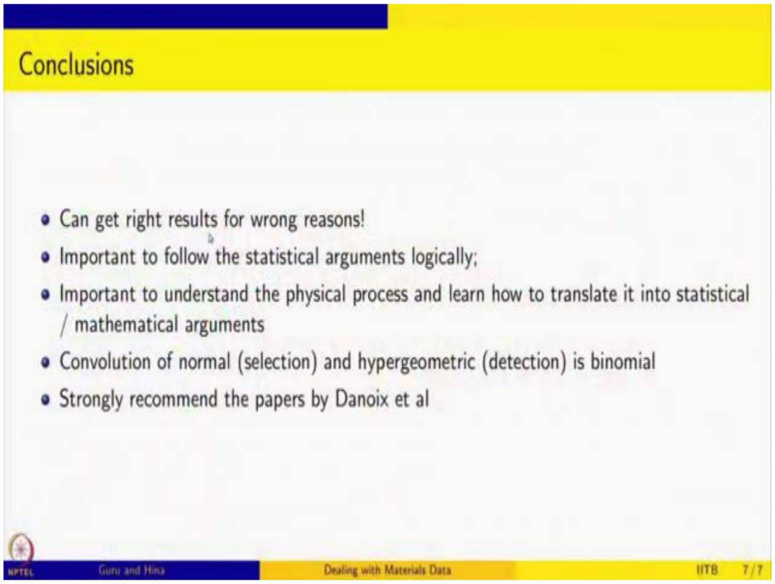
So we want to sum this probability distribution over m so we will get something which is only a function of j and n and if you do the summing. This is the definition of marginal distribution, we discussed when we started these sessions on probability distribution. So, if you have a joint probability distribution, you can sum over one of the variables and what you get is a marginal distribution.

So that is the idea that is being used here. So you sum and you get a D, and D you get in terms of j and n, and you can show that this goes as binomial. Because it goes as binomial and it goes as binomial with n and P as the parameters, because it goes as binomial with n and P as parameters, you can show that the variance is p_0 into 1 minus p_0 by n. This is what I said about the surprising result at the end after doing all the analysis.

So, here is a process where we assume that they are independent and we just added and approximately we got this result and here is a process where we have done the actual analysis, the correct analysis assuming that they are sequential and they are not independent, so do the marginal distribution and get the variance for the marginal distribution, we get the same result.

So, even though the classical error analysis was wrong, the result was not wrong, which means that Danoix paper shows the right analysis, but it does not rule out the conclusions that that were drawn from the previous studies. So, that is the surprising thing that you do all this detail analysis and you find that accidentally you happen to have the same variance, which also tells you that sometimes you might measure some variants and that might be a right variance, but you might be having a wrong understanding of things. So which is a little bit problematic.

(Refer Slide Time: 11:29)



Conclusions

- Can get right results for wrong reasons!
- Important to follow the statistical arguments logically;
- Important to understand the physical process and learn how to translate it into statistical / mathematical arguments
- Convolution of normal (selection) and hypergeometric (detection) is binomial
- Strongly recommend the papers by Danoix et al

So, that is the first conclusion you can read get right results for wrong reasons. So you have to be careful, you have to keep analyzing your logic, your way of thinking and your analysis. But, and it is important to follow statistical arguments a logically if you do not do, if you just do a recipe kind of thing without understanding exactly what is what you might do wrong analysis and the this is also a nice case study to understand how we have to understand the physical processes and translate them into statistical or mathematical arguments.

So, which is infinite sample, which is a finite sample, which is sampling without replacement and so on and so forth. So, there are at every stage you have to understand the actual process that happens in the experiment, then we have to find out what is the right statistics to describe that process and it so happens that the convolution of normal which is a selection process and hypergeometric which is a detection process is a binomial.

Why is the selection process normal? Because we are going to just pull out some m atoms from the sample, we are going to pull them out randomly. So, any deviation you would find is only because of noise or thermal fluctuations and we are going to learn about this distribution called a normal distribution, which is a continuous distribution. It is very-very important. One of the reasons why it is very important is also the fact that random errors typically lead to normal distribution.

So, so, that is why the selection process was normal and the detection was the hypergeometric and so the convolution of these two normal plus hypergeometric happens to be binomial and that is why we got the binomial as the solution. So, I once again strongly recommend reading the papers by it Danoix et al and it is very nicely written and very logically explained, it will give you a good idea about this process of atom probe.

It will also give you a flavor for how these statistical distributions are important in material science and engineering and finally, it will also tell you how to mix the different components that we have. We did error analysis and we are doing probability distribution, but real life problems do not come labeled as probability distribution problem or error analysis problem, it comes as a mixture and in fact, it is in this case coming as a mixture of some normal distribution and hypergeometric distribution.

So, it is two different distributions one continuous and one discrete and actual process that you do happens to be a combination of these two out of which then you have to do the error analysis and come up with the right value for error and so on. So, that is why it is a nice case study, where you how to bring to bear all the information you have and all the concepts you have learned and all the methods you have learned and use it to solve problem of practical interest.

So, these are some of the important distributions, discrete distributions we have looked at, and also important type of problems that you will encounter in terms of finding out errors and so on. So, we will continue to work with the more distributions and so far we have done Bernoulli trials, binomial, negative binomial, and the hypergeometric. So then we will move on to one more discrete probability distribution that we have, which is a Poisson distribution.

So, that is what we will do in the next session. So, I will end this session here and I again strongly recommend that you read the paper of the Danoix et al to have a better appreciation in the second paper there is also the 3D analysis that is described and again, there is a interesting conclusion there that the detector efficiency also is going to contribute to the error so, I recommend that you take a look at the paper. Thank you.