

Course on Dealing with Materials Data: Collection, Analysis and Interpretation
Prof. M P Gururajan,
Professor. Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 30
Summary: Descriptive statistics

Welcome to dealing with materials data. In this course we are trying to understand the collection analysis and interpretation of materials data. We have done two modules, so far, one is introduction to R. The second one is using descriptive statistics using R. This is the summary of the second module on descriptive statistics using R.

(Refer Slide Time: 0:42)

Summary

- Visualising data: scatter plot, dot-chart and stem-and-leaf plot
- Rank-based reports: cumulative distributio, histogram, box-and-whisker plot
- Summary reports: mean, median, variance, standard deviation and quantiles
- Significant digits in a given data
- Reporting absolute and relative errors in data
- Presenting data with error bars to understand trends
- Classify errors
- Understand uncertainty propagation and carry out Monte Carlo simulations for error propagation

IITB 3/4

So, we have learned how to visualize data. So, we have learned how to use scatter plot, dot-chart and stem-and-leaf plot for visualizing data. Then we have learned how to prepare rank-based reports of data which includes cumulative distribution, histogram and box-and-whisker plots. We also have learned how to prepare summary reports for data. Mean, median, variance, standard deviation and quantiles are the quantities that one calculates in the summary-based reports.

One has to learn about the significant digits in a given data. Sometimes when you do this analysis, computer returns a large number of decimal points but beyond a point, some of these numbers are not meaningful and so we should not report them.

This is like, for example, if you say that price of some three things is 100 rupees, and each one costs how much. So, we are not going to report the resulting number beyond

second decimal place because below paisa there is nothing. There is no meaningful number that you can quote.

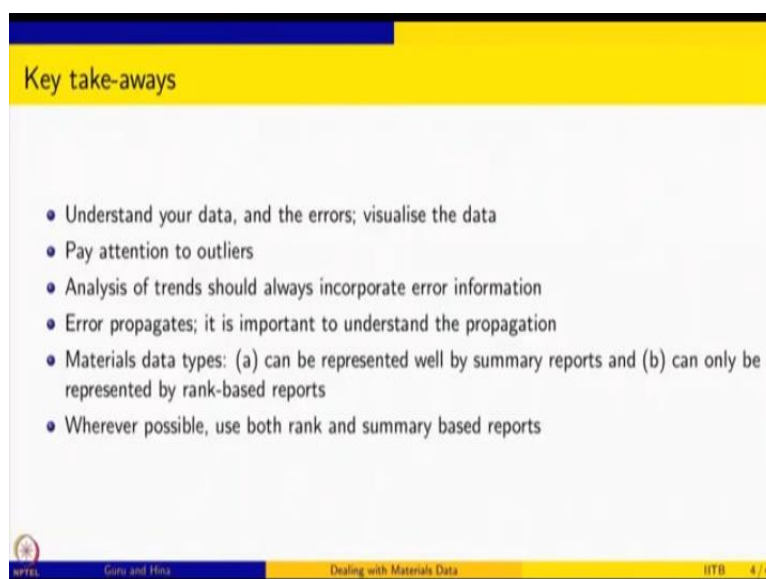
So, this is similar in all cases. For example, we have seen that conductivity measurement itself is up to first decimal place. Then it makes sense to report the means and standard deviations etc up to first decimal place.

This does not mean that when you are doing the calculation, you will always do up to first decimal place. It is a good idea to keep the extra digits and round off only at the last step, when you report the numbers. And so but it is important to know the significant digits in any given scenario and report data only up to that.

We have also learned how to report errors. You can report it in absolute terms and in relative terms and we have learned how to present data with error bars because in order to understand trends in any given data, it is not sufficient just to look at the mean values but you should also look at the error bars in the data.

And we have learned how to classify errors and we have learned that error propagates and we have learned how to quantify this uncertainty propagation, either by using some analytical calculations or by using Monte Carlo simulations. So, either way, you can find out how the error propagates.

(Refer Slide Time: 3:14)



Key take-aways

- Understand your data, and the errors; visualise the data
- Pay attention to outliers
- Analysis of trends should always incorporate error information
- Error propagates; it is important to understand the propagation
- Materials data types: (a) can be represented well by summary reports and (b) can only be represented by rank-based reports
- Wherever possible, use both rank and summary based reports

NPTEL Guru and Pina Dealing with Materials Data IITB 4 / 4

So, these are things we have covered. We have also learned few important things in doing this course, in this module. First thing is, you have to understand your data and the errors. So, visualizing that data is a very good way of understanding the data.

And visualizing with error bars is a nice way to understand the errors in the data. And it is always important to pay attention to the outliers. Some amount of effort is needed to understand why they are there, and it will help you improve the experiments or understand what is happening better.

We saw one example where there was an outlier in the electrical conductivity measurement and little bit of analysis showed that the measurement methodology was not applicable for that scenario and that is why we got some meaningful and numbers which were not in tune with the rest of the numbers. So, they were no consistent with the rest of the data.

So, we have also learned that while analysing trends, it is important to incorporate the error information and we have understood the importance of propagation of errors and so if you measure some quantity and it has some uncertainty, any subsequent analysis that you do using that data also picks up the uncertainty from these quantities.

And we have also learned that materials data is broadly of two types. One, which can be represented by summary reports like conductivity for example. So, it follows nice normal distribution so it is sufficient to give the mean and standard deviation that completely describes the data. This is because every measurement gives one number and the fluctuation that you see is random. It's a noise, so it can be very well described by the normal distribution.

And on the other hand, some materials data can only be represented in rank based reports. You need to give things like histograms to describe this data and if you just use rank based reports like we found, for example, for phase two, it looked as if there were so many outliers.

It was as if you assume that the data is normal then obviously it is not normal because of which you find that compared to the mean up to 6 sigma you found the data points, and they were all only on one side. So, it is obviously not a normal distribution.

So, it is better to describe such data using the appropriate distribution. It is not a good idea to assume that it is normal and assuming that it is normal or not can also have a say on understanding the further results.

So, we have a found an example that if you are trying to calculate the uncertainty propagation using grain sized data, whether you are going to assume that it is normal

and the errors are normal or it is not, it is lognormal, for example, it is going to make a difference to your analysis.

So, it is important to know. It is also important to sometimes report this quantity. Sometimes in the literature these quantities are not reported then it becomes very difficult to understand or analyse the data.

And finally, wherever possible, of course we should use both rank and summary based reports because that is the most complete information that one can give. The best is to actually give raw data and it is recommended it give it as supplementary data but in addition when you do any analysis, we should give the analysis methodology and we should give as much of information as possible in terms of summary based and rank based reports so that a complete picture of data is presented. Thank you.