Dealing with Materials Data

Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay

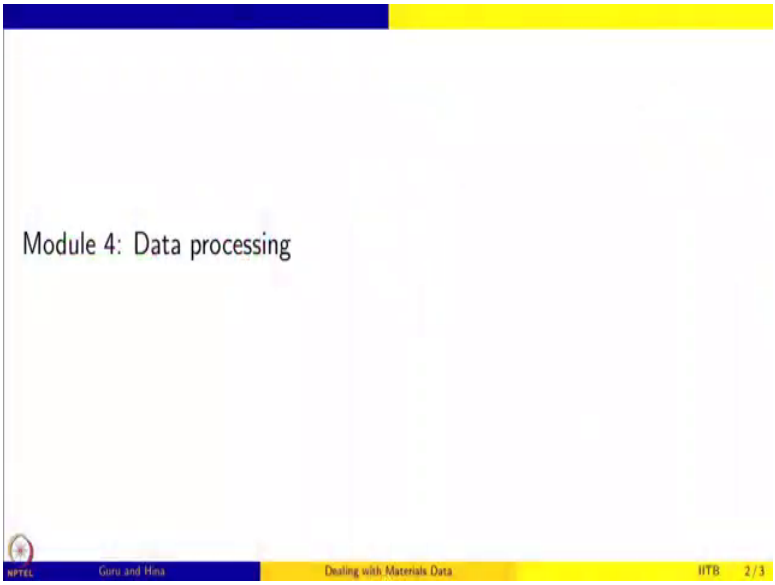
Lecture 60

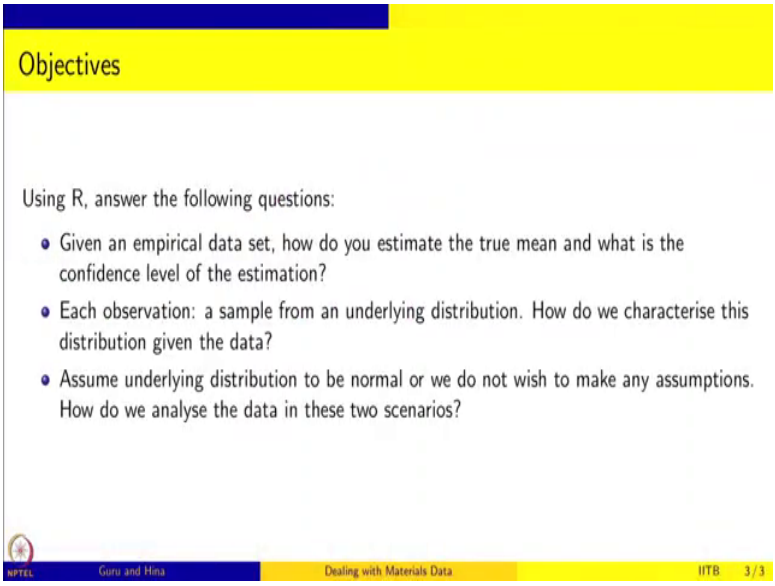
Data processing: Introduction

Welcome to Dealing with Materials Data. In this course, we talk about the collection, analysis and interpretation of data from material science and engineering. Till now, we have done 3 modules. First is the introduction to R, second one is descriptive statistics using R and the third one is probability distributions.

So, we have pretty much every tool that we need. So, we are now ready to go take up data and analysis and draw conclusions. So, in this session we are going to do one such analysis.

(Refer Slide Time: 00:52)





So, this module is for processing of data and in this module using R, we want to answer the following questions. Suppose, if you are given a empirical data set, how do we estimate the true mean and what is the confidence level of that estimation? Because we make lots of measurements and at the end of the measurement, we want to give one number as the mean value of the parameter that we are probing.

And when we give that value, we also want to say how confident we are that, that value is correct or in what interval does that value lie, for a given confidence level, say 90 percent probability this value will live only in this range. So, there is also, so it is not just the mean, it is also about the distribution of the data, the spread of the data.

So, we want to give this information. So, we know that each observation is a sample from an underlying distribution. So, given the data can be characterized the distribution is the next question. Of course, if we assume that the under data set is normal, we can do some analysis. If we do not have that assumption available to us, or if we do not have reasons to believe that the underlying distribution is normal, how do we go about analyzing the data?

So these are the three things that we want to answer. So take a dataset, can we say something about the properties like mean and spread? And can we say something about the underlying distribution by looking at the data? And if we do not have the assumption that it is a normal distribution, can we say anything about the data? So, these are the type of questions that we want to answer.

So obviously, these are simple questions. There are more complicated questions that we will ask that we will do in the next session. But in this session, we are going to simply answer these questions. And because now, we have had some experience with R, I am going to show all these examples using the same data set that we had been using, namely the conductivity of electrolyte tough pitch copper, so that you are familiar with the data.

And you can see the different analysis done on the same data. So, you can understand how this analysis work. But from this exercise on, we are also going to give lots of other data for you to do the same analysis. So, some of the analysis that I am going to do is repetition, you might have seen it, but this is the kind of set of steps that you need to follow if you get a new data set and we will have more exercises of that type also in this module.

So, we will take up these objectives one by one and use the copper conductivity data and one more data that is important for a completely different reason, namely, from the existing literature, if there are some data that becomes available to you. First, how to get that data in a form in which you can analysis and then second, how do you analysis? So, these are the things that we are going to discuss in this module. Thank you.