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
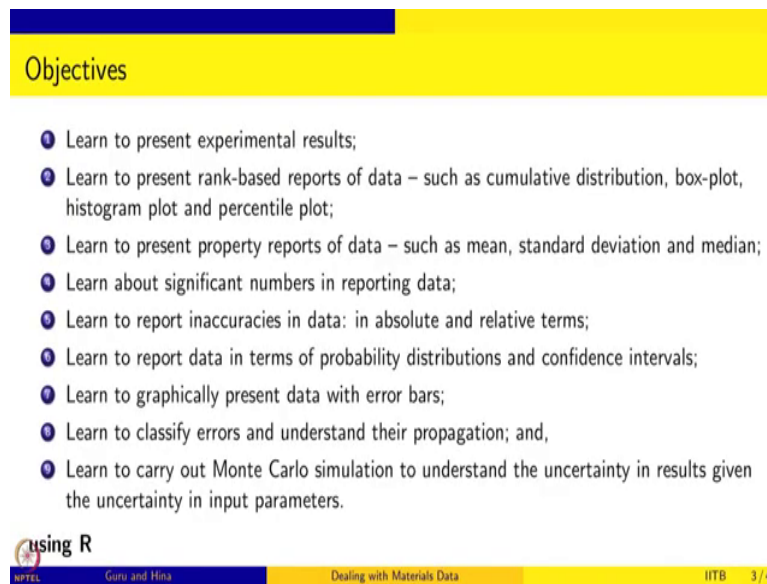**Lecture No. 19**
Descriptive Statistics

Welcome to Dealing with Materials Data. In this course, we are going to learn about Collection, Analysis and Interpretation of materials data and we have completed one module which is on introduction to R. In this module again we are going to use R and we are going to learn how to do Descriptive Statistics?

(Refer Slide Time: 0:42)

So, this is a module on Descriptive Statistics using R.

(Refer Slide Time: 0:46)

And here are the objectives of this module. First thing, we want to learn to present experimental results and most of the times it is not possible to present all the experimental data that we have, so we have to learn to present rank based reports of data. What are these? These are things like cumulative distribution, box-plots, histogram plots, percentile plots and so on and so forth. So, you have the raw data and you do some amount of manipulation or analysis on that and then report the data.

You can also present property reports of the data such as mean, standard deviation and median and so on and so forth. And we will see examples of both sometimes it is useful to combine them both and to look at the data to better understand data. And sometimes just giving these property reports are sufficient they completely describe the data. Sometimes they are not sufficient, sometimes it is essential to give other information like cumulative distribution or histogram plot.

And while reporting data we should know about the significant numbers. Again if you just measure data, then you will get some numbers and up to the accuracy to which it is measured using your equipment that information will be there. But, when we do manipulations on the data, we will have some numbers that turnout and they might not have the same accuracy or same significant numbers. So, we have to make a conscious decision as to up to what number we are going to keep. So, we will learn about significant numbers in reporting data.

Obviously, when the experimental results are obtained and analysis is done, we will learn about the inaccuracies in our data and we should be able to present them. It is a good practice to report not just the numbers but the associated inaccuracy. And we can do it in two ways, in

absolute terms and in relative terms and we are going to learn how to do that in this module. And we will also learn partly, how to report data in terms of probability distributions and confidence intervals.

But, we are going to learn about probability distributions and how to use R to look at probability distributions or manipulate them and then so on. So, we will come back to this aspect again later in the course and pay closer attention but to some amount of reporting as probability distribution we will do in this module.
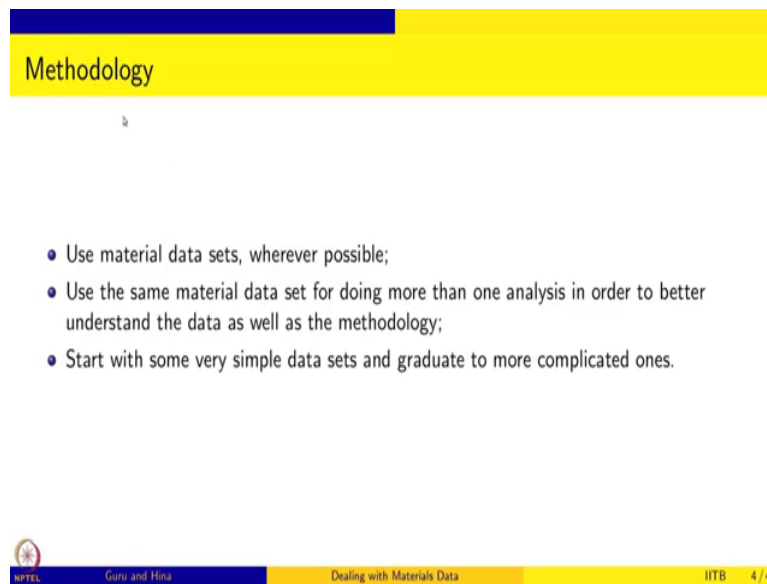
And how to graphically present data with error bars is something that also we should know. So, we will learn that and we will also learn you know, the experimental data you can take, you can analyse, you can manipulate, you can present and you can give errors, you can give confidence intervals, you can graphically present.

But, it is also essential to know about the errors in data and classify the errors and understand how they propagate? Most of the times it is not just the measurement but based on the measurements we carry out other calculations. So, if the measurement has an error, what is its effect on the further calculations that we do is very-very important.

And the finally, this is going to be an important chunk of this module is to learn to carry out Monte Carlo simulation to understand how uncertainty in results are going to be given the uncertainty in the input parameter. So, we want to estimate the uncertainty in the result given some uncertainty in the input parameters. So, we are going to look at some very-very simple cases to just understand, how this is done.

And we want to do all this using R. So, that is the purpose of this module. So, this is to do a descriptive data analysis using R. So, that is the aim of this session, this module. And we are going to do all of this at some level and some of them we are going to like probability distribution and confidence interval. We will do only partly we will come back and re do it later after we do the probability distribution part of the course.

(Refer Slide Time: 5:31)

And, so the methodology as usual is to use materials data sets wherever possible and we are also going to try and use the same material data set for doing more than one analysis and this will help us better understand the data. It will also help us understand the methodology. So, familiarity with the data and better understanding of the data will also help us understand the methodologies of these some of these analysis themselves.

And, we are going to start with some very simple data sets and we will graduate to more and more complicated ones. Even then, they are not data sets which are very large like sometimes are used in the literature. But, we hope that our sessions will prepare you for dealing with such really large data sets and more complicated these data sets.

So, this is the session on doing Descriptive Data Analysis or Descriptive Statistics using R. Welcome and we will go through each of these objectives one by one using materials data sets. Thank you.