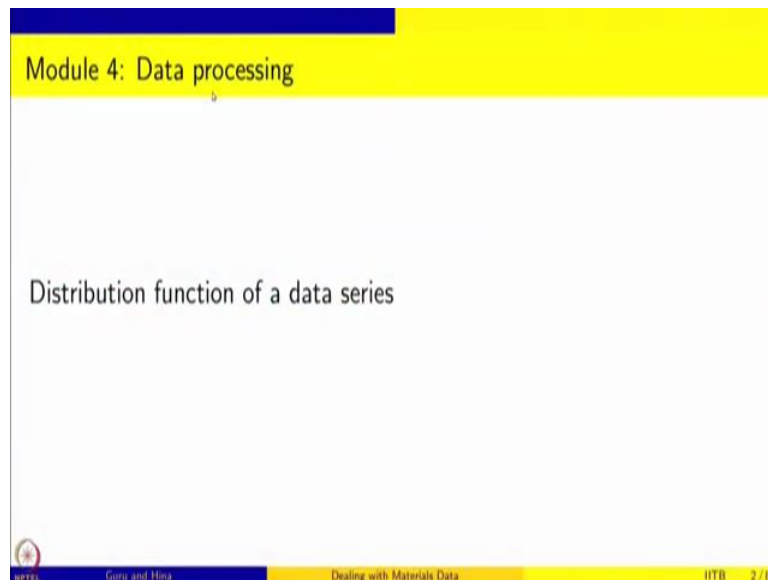


Dealing with Materials Data
Professor M P Gururajan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 61
Distribution Function of a Data Series

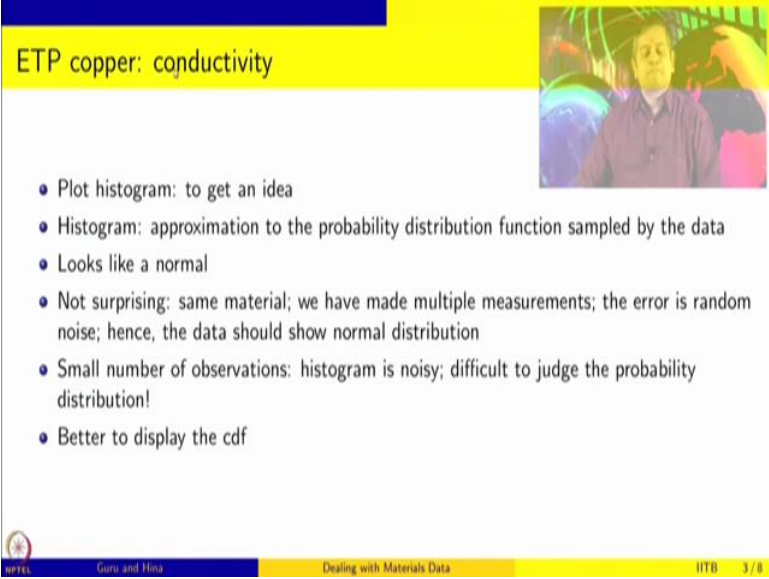
Welcome to Dealing with Materials Data, in this we are looking at the collection, analysis and interpretation of data from Material Science and Engineering. And we are in the module on data processing.

(Refer Slide Time: 00:27)



And in this session we want to look at the distribution function of a data series. So, you have a set of numbers, you have some data that is available to you and you want to say something about the distribution function of the data that you have got.

(Refer Slide Time: 00:47)



ETP copper: conductivity

- Plot histogram: to get an idea
- Histogram: approximation to the probability distribution function sampled by the data
- Looks like a normal
- Not surprising: same material; we have made multiple measurements; the error is random noise; hence, the data should show normal distribution
- Small number of observations: histogram is noisy; difficult to judge the probability distribution!
- Better to display the cdf

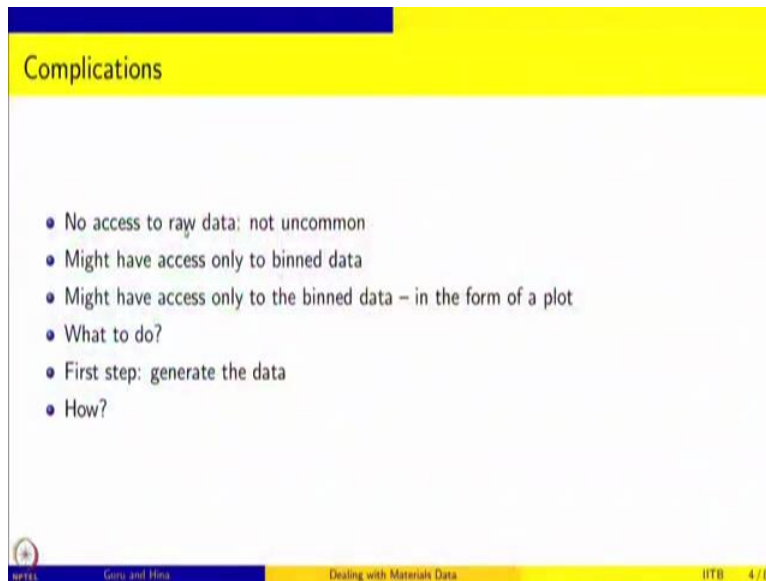
NPTEL Guru and Hina Dealing with Materials Data IITB 3/8

So, there are 2 things that we are going to do, one is the ETP copper, we will take the conductivity data. And the first thing when you take data like that is to plot the histogram. A histogram gives you an idea about the approximately what the probability distribution function that is sampled by the data.

Of course, in the case of copper conductivity, we know that it looks like normal. It is not surprising, because it is the same material and we have made more than one measurement and so, the noise is random.

And if that is so, if you have random errors or noise, then you do get normal distribution. But if you have very small number of observations, a histogram can be noisy and it might be difficult to judge the probability distribution. In those cases, it is better to display the cumulative distribution function. So, we are going to do this exercise for the ETP copper we will do all this.

(Refer Slide Time: 01:44)



The slide is titled "Complications" and features a yellow header bar. Below the header, a list of five bullet points is displayed on a white background. At the bottom of the slide, there is a footer bar with a blue and yellow gradient, containing the text "Guru and Hina", "Dealing with Materials Data", and "ITD 4/8".


- No access to raw data: not uncommon
- Might have access only to binned data
- Might have access only to the binned data – in the form of a plot
- What to do?
- First step: generate the data
- How?

But there are sometimes complications. It is not uncommon to not have access to raw data. So, in this case, for example, I have data for all the 20 measurements, but typically that is not what is published. It is very rare for people to list out all the measurements that they make. And what is worse, sometimes you might have access only to binned data. That is data has already been analyzed, and values which lie within a range, they are all just counted, without really telling what exactly the values were.

And they are just given in the form of a histogram plot. And so even the binned data is not available to you in raw form is what I am assuming. You might have access only to the binned data and only in the form of a plot. Now, can we do any analysis on that data? Is the question, the answer is yes.

But to do that, first you have to generate the data. So you have to go from the plot to the data and then from the data, we can go back and do the analysis. And how do we do it?

(Refer Slide Time: 02:56)



Some tools!

- Disclaimer: there are too many! The following is based on my preference, choice and/or knowledge
- Engauge Digitizer: Read the data from the image
- gimp: To pull out the data histogram as a jpeg file to feed to the digitizer
- LibreOffice: export from digitizer directly or enter the data into the file by hand

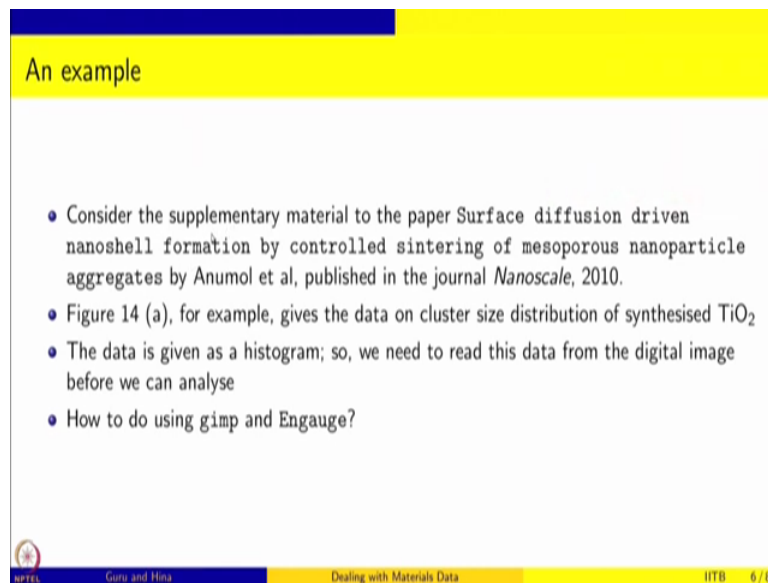
NPTEL Guru and Hina Dealing with Materials Data IITB 5/8

So at this point, a disclaimer is necessary. There are too many different ways doing or achieving this. And different people have different preferences but based on my preference, choice and experience, so I am going to give you an introduction to a couple of tools. Introduction in the sense that I will just show how it is done.

One is called Engauge Digitizer, this is used to read data from image. The other one is called the gimp, this is to pull out the data from a PDF file into a JPEG file, in the form in which you can then feed it to the digitizer. Ofcourse, you can use Libre Office to get this data from engage digitizers after reading the data into Libre Office directly or you can do it by hand.

And in the following exercise that I am going to show you the data was entered by me by hand, but ofcourse I will show that digitizer also can do it automatically, which we are not going to use so much. So in this session, we are also going to look at how to use these 2 tools to generate data from PDF file, of the type which is binned data, that is what we are going to look at, but you can use this for everything.

(Refer Slide Time: 04:27)



An example

- Consider the supplementary material to the paper Surface diffusion driven nanoshell formation by controlled sintering of mesoporous nanoparticle aggregates by Anumol et al, published in the journal *Nanoscale*, 2010.
- Figure 14 (a), for example, gives the data on cluster size distribution of synthesised TiO₂
- The data is given as a histogram; so, we need to read this data from the digital image before we can analyse
- How to do using gimp and Engauge?

npTEL Guru and Hina Dealing with Materials Data IITB 6/8

And specifically, what we are going to do is to consider the supplementary material to the paper, surface diffusion driven nanoshell formation by control sintering of mesoporous nanoparticle aggregates by Anumol et al, is published in nanoscale and there is figure 14 A, which gives the data and cluster size distribution of synthesized Titanium and it gives that in the form of a histogram and that is what the data that we are going to read out.

And we will use, we have used the digitizer and the gimp to do this. So I will show you the data and we will do the analysis on the data. But I will also show you how to use gimp and Engauge by taking figure 14 B as an example. So, that will be part of this session.

(Refer Slide Time: 05:14)

The slide is titled "Data with different statistical weights" and contains the following bulleted text:

- ETP copper conductivity data: each data point has equal weight ($= \frac{1}{n}$) where n is the number of measurements (20 in this case)
- Binned data: statistical weights are different
- Cluster size data: bins of size 20
- Original data: we have no access
- Let us assume that 200 nm cluster size, for example, means all clusters of sizes in the range 190 to 210 nm
- Each bin: gets a statistical weight w_i
- w_i : number of observations in the i -th bin to the total number of observations

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

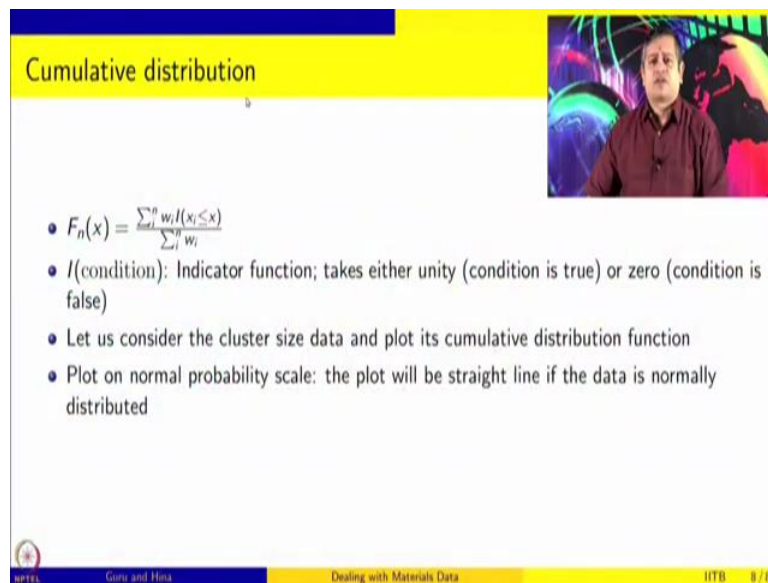
And when you have binned the statistical data like that, there is data with different statistical weights. Why is that so? In the case of ETP copper conductivity data, each point had equal weight, $1/n$, where n was a number of measurements because all data points were just measurements.

But when you have binned data, the statistical weights of different bins are different. For example, in the case where we are looking at this Titanium size, the cluster size data was given in bins of size 20 nanometers, so, we do not have access to the original data. And if you look at something like 200 nanometer cluster size, it means all clusters of sizes 190 to 210 were put in this bin, right.

So, we need to give a statistical weight w_i to each of the bins. And the w_i is nothing but the frequency in that bin divided by the total number of observations. So w_i is the number of observations in the i th bin to the total number of observations. So, this is the way to give different statistical weights and it becomes important in such binned data.

Like I said, it is not very uncommon to see such binned data being published. So, sometimes if you want to do analysis with the existing data, that is there in the literature, this kind of exercise becomes essential.

(Refer Slide Time: 06:45)



Cumulative distribution

- $F_n(x) = \frac{\sum_{i=1}^n w_i I(x_i \leq x)}{\sum_{i=1}^n w_i}$
- $I(\text{condition})$: Indicator function; takes either unity (condition is true) or zero (condition is false)
- Let us consider the cluster size data and plot its cumulative distribution function
- Plot on normal probability scale: the plot will be straight line if the data is normally distributed

MPPL Guru and Hina Dealing with Materials Data IITB 8/8

And so then we plot cumulative distribution, in the case of connectivity data, for example, it is rather straightforward to plot the cumulative distribution function, ECD, a function we will do and we have already done it once, but we will just repeat for the sake of completion. In the case of binned data, of course, you have to get the cumulative distribution by considering the weighting factors, right.

So, i is an indicator function, it takes either unity when the condition is true or 0 when the condition is false. So, what is within i , the argument is a condition and if the condition is satisfied it will take unity and if it is not satisfied it will take 0. So, we will consider the cluster size data and we will calculate the cumulative distribution function using this because cumulative distribution function x means the probability that it is less than or equal to x itself.

$$F_n(x) = \frac{\sum_i^n w_i I(x_i \leq x)}{\sum_i^n w_i}$$

So, and then we will plot it down the normal probability scale and the plot will be a straight line if the data is normally distributed. So, that is how we know what is the underlying distribution. So, you can look at the cumulative distribution function and make out easily what type of data you have. Of course, you can also get it from histogram data, if the data is good you can make out fairly easily.

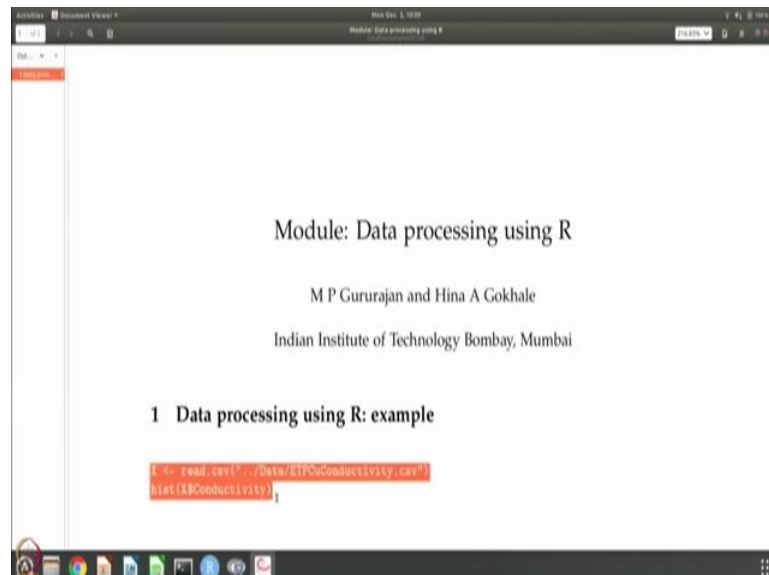
But otherwise, it is an approximation because noise can throw you off. And there are places where it is very difficult from histogram to actually understand, for example, log normal or variable might be very difficult to distinguish. So the cumulative distribution function is a

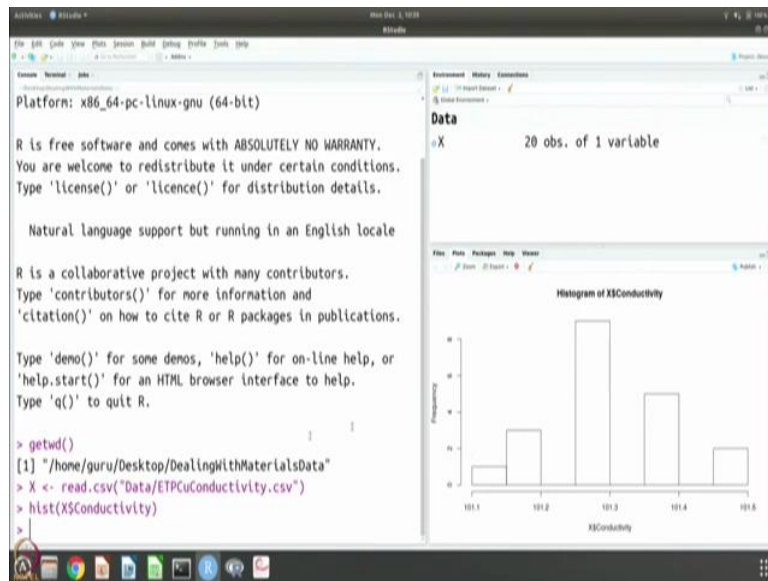
slightly better way of understanding the distribution from which the data comes. So let us go do this exercise.

So, first we want to take the copper conductivity data, do the analysis, then we want to understand how to take PDF file, specific figure you want to cut out and then generate a JPEG out of it in such a way that it can be fed to digitizer. And in digitizer, then we need to know how to read the values, which can then be entered into Libre Office and you can have a csv file which can be used for further analysis.

So I am going to do that as the second exercise to show you how to do this digitization and followed by the reading of numbers from such figures. And finally, we will take such, one such binned data and do the analysis on that data in this session.

(Refer Slide Time: 09:26)

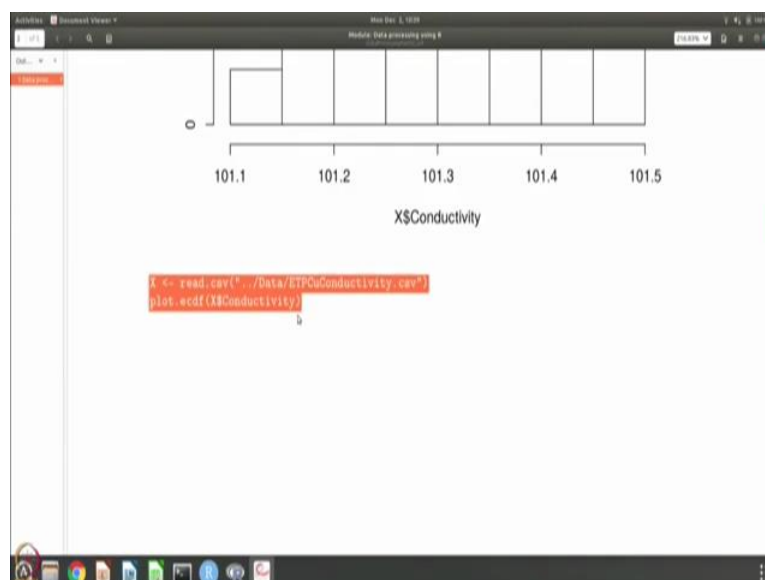


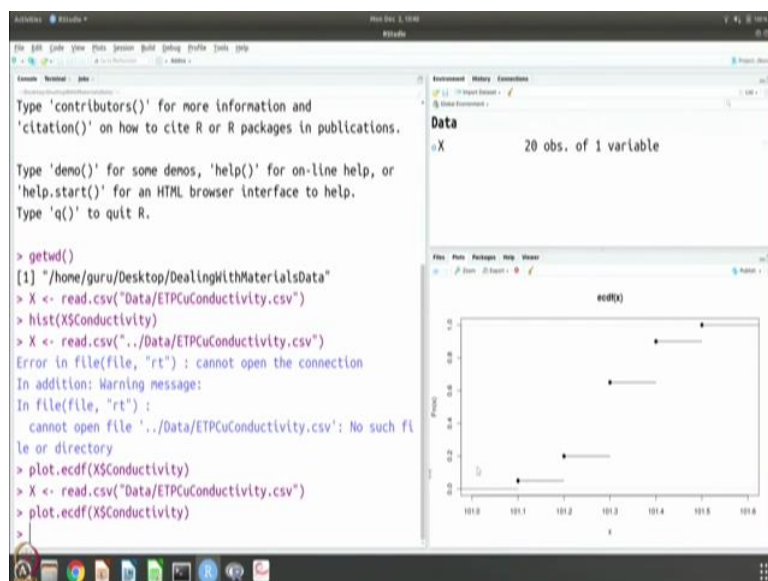


So as usual, the first exercise to do is to start with R. Okay, so it is a good idea to know what is the working directory, so we are in the right directory, and R version is 3.6.1. So the first exercise is to read data and plot the histogram. So let us do that. So we want to read the ETP conductivity data, and then we want to plot the histogram.

So of course you can look at the histogram, this is the frequency versus data plot. And it does look like normal distribution, may be slight skewing. So, this gives you an idea that this might be normal distribution, so that is a first exercise.

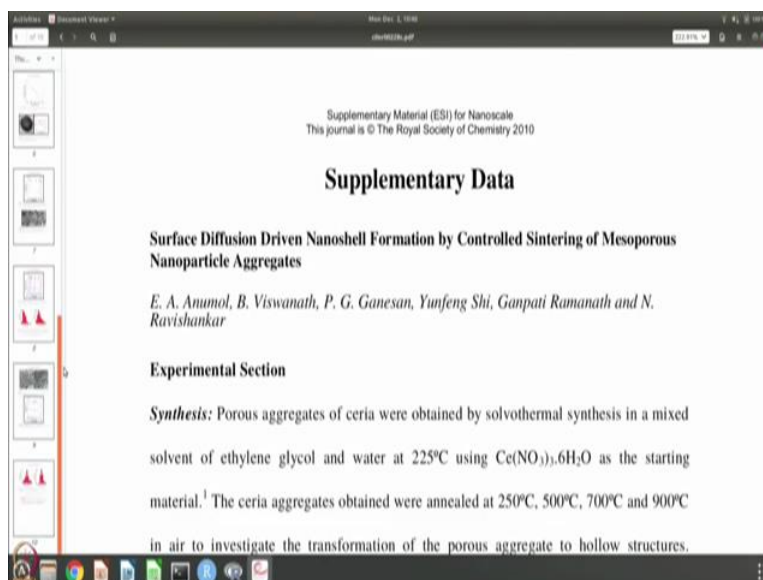
(Refer Slide Time: 10:20)





And the second exercise of course, is to plot the cumulative distribution function. So, this is the cumulative distribution function, and this also indicates that this could be okay. So, this also indicates that this could be a problem with the, this could be a normal distribution. And of course, it is difficult to make out, so we would like to make the y axis to be scaled as a probability distribution. And then you will find that it is normal. And this exercise we have done in the past, once.

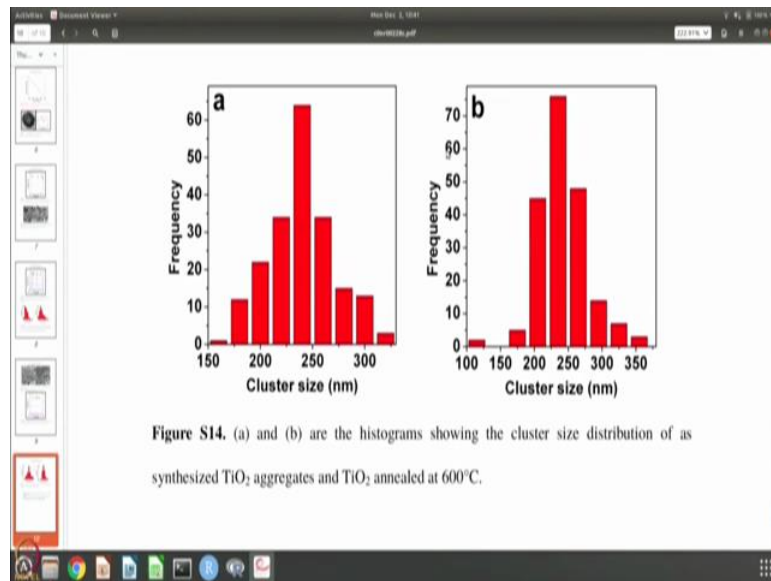
(Refer Slide Time: 11:20)



Now, let us go to the second exercise. Let us consider this supplementary data. So this is a supplementary data on Surface Driven Diffusion of Nanoshell Formation by Controlled Sintering of Mesoporous Nanoparticle Aggregates, by Anumol et al. And it is published in

nanoscale. So, this is a supplementary information to the paper. And the paper has lots of data and one of the data that we are interested is here.

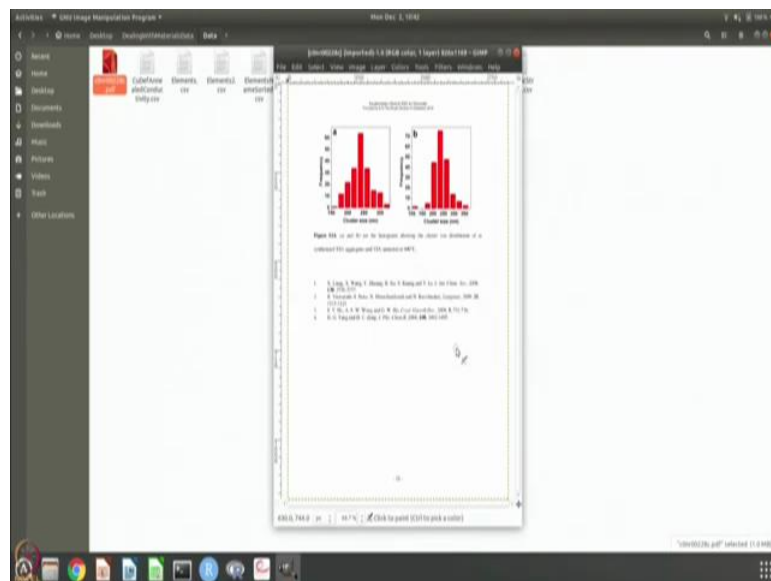
(Refer Slide Time: 11:39)

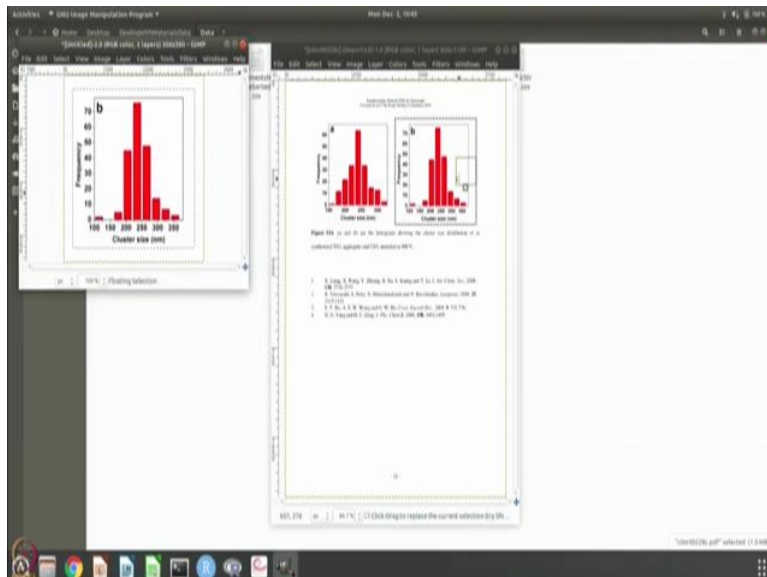
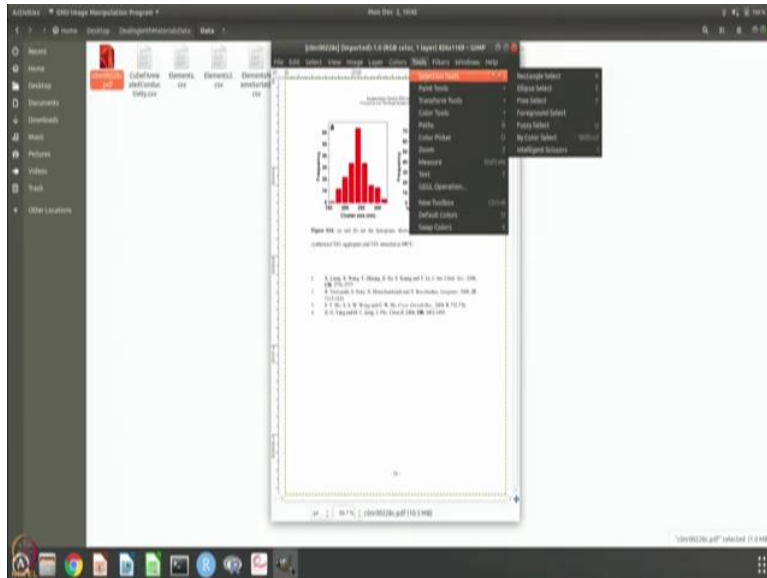


So, it is a cluster size in nanometer versus frequency. So, the figure caption says that it is a histogram showing the cluster size distribution of a synthesized Titanium aggregates. And this is the cluster size distribution for Titanium annealed at 600 degree Celsius.

So, this is the data that I have taken and generated the raw data for our further analysis. But I want to show how I did that using this as an example. So first thing we need to do is to take this figure and generate a JPEG file out of it so that this can then be fed into the digitizer.

(Refer Slide Time: 12:31)





To do that, I am going to open this file using the application GIMP. GIMP is a gnu image manipulation program. And it tells me which page I should open. Of course, I want only this page. So I say this page. Once I have this page, gimp has lots of tools, the tool that is of relevance to us is a selection tool.

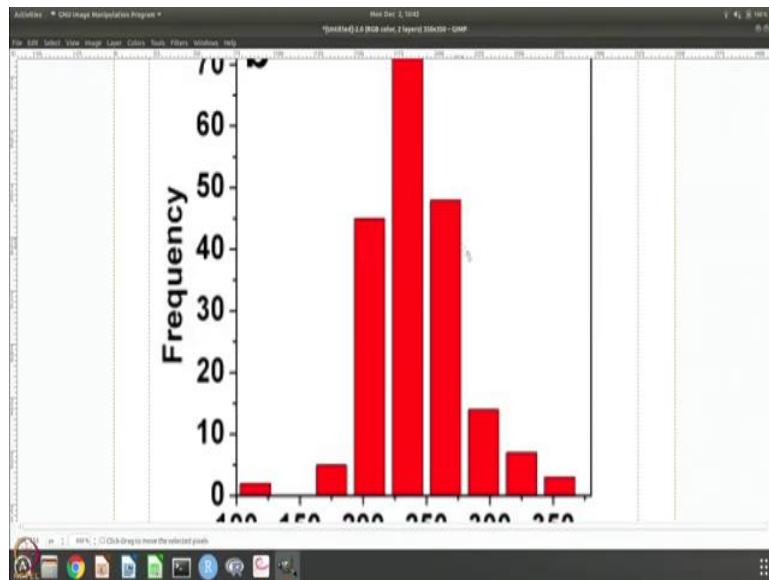
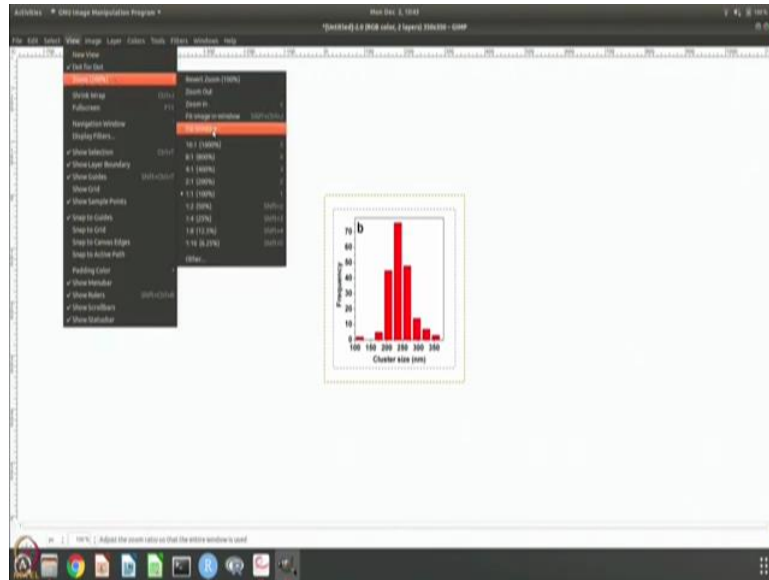
So I am going to say rectangle select. What is it that I am going to select, so you can mouse click and get this. And then you can say, copy or cut that portion and you can make a new file. And so you need to give the size of the file. So the size you can read off from here, for example right.

So this is 250, this is 500, this is 750. So we have about probably 300 or 350 in width. So I am going to say 350 in width. And you can see in the height also. So this is 250 and this is 500. So

we again probably have another 300, 350 in height, so I am going to say 350 by 350. Then I am going to say paste.

So you can see that the figures that is here, I cut and I pasted it here, and then I am going to do some analysis here. I am going to transform.

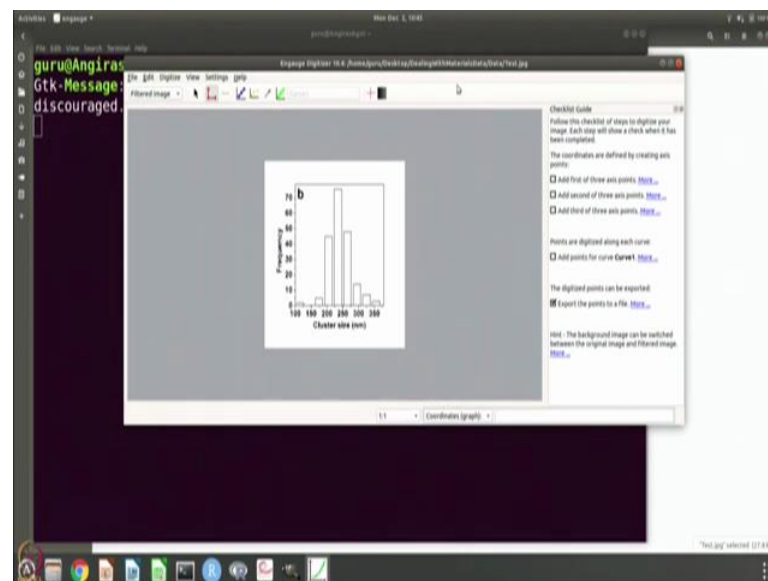
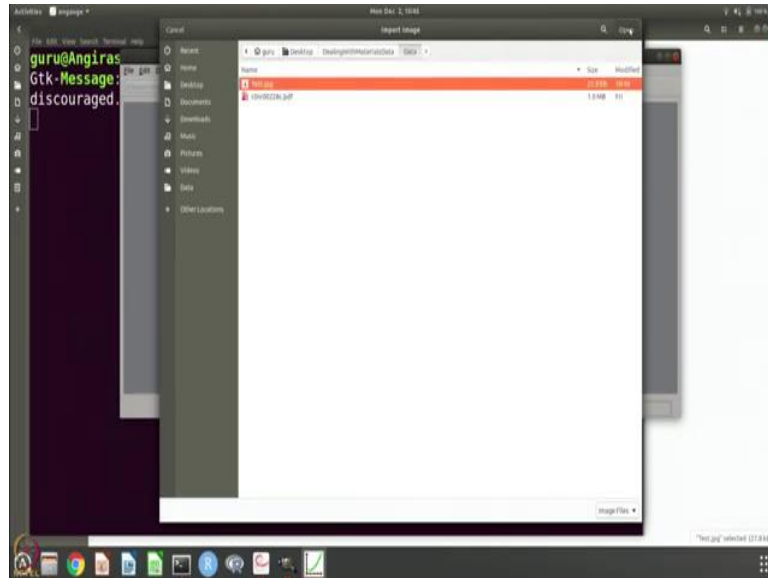
(Refer Slide Time: 14:05)



No, I do not need rotations. So I just want to make the image a little bit bigger, so that will be easier. So let us say that I want to zoom it to some 400. Okay, so here is the figure. Now I am going to save this figure as yeah, so let us go here. Let us go to data and this is, let me save it as test.cf so that you can later do further analysis. But we actually want the data, the figure to be in JPEG format, right.

You can export the image as JPEG and as good quality as possible, so we are going to export. So the figure has been exported and if you go here you can see that that is this test.jpg.

(Refer Slide Time: 15:17)

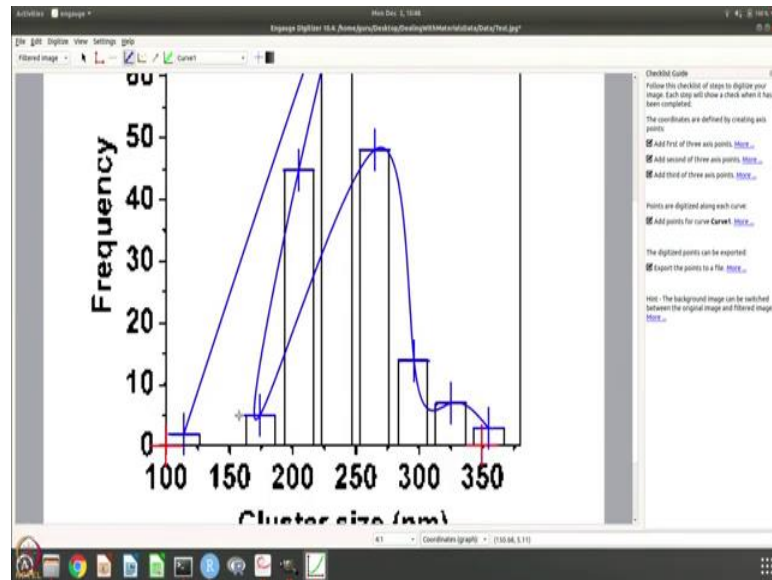


Now we are going to say, now we are going to use this yeah, so engage digitizer is this and I am going to say file import. The file that I need to import is the file that I recently generated. That is in the data file here, the test.jpg. So I am going to open and okay, so the data has been imported here.

The first thing that we need to do is to identify the x y axes so that the program can read the data. So that is the first step. So you need to identify 3 points on the axes. The origin, some distance along x, and you have to tell what that point is, some distance along y and you have to tell what that y point is.

So, that the sort of distances on this plot is mapped for the program. So it knows if you go some distance, how much does it correspond to in the x units or y units.

(Refer Slide Time: 16:40)



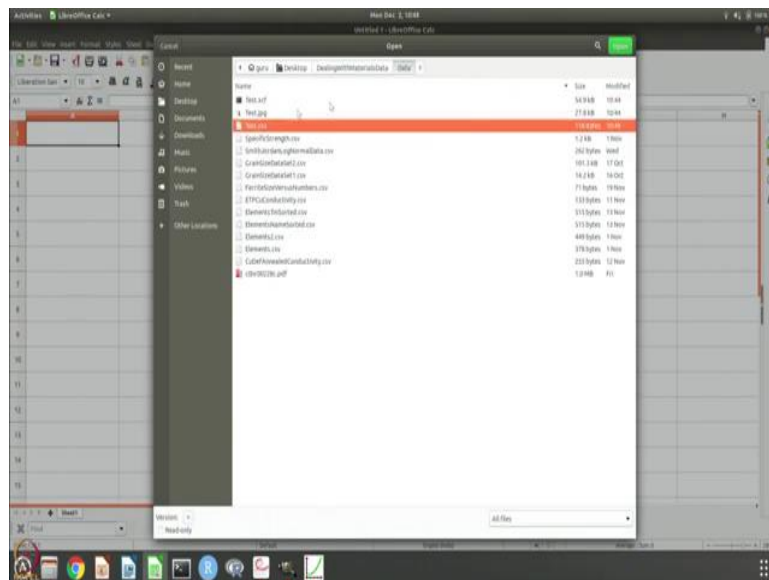
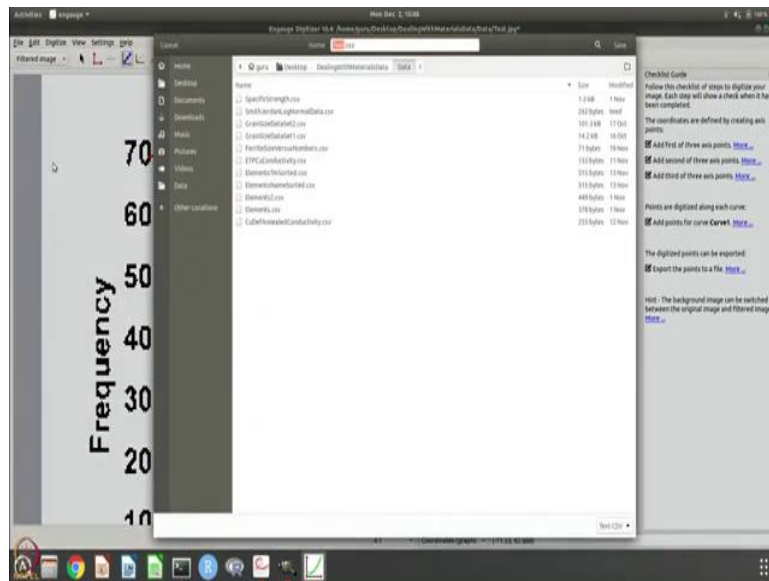
So let us do that. So first, we want to do that and it is, I am going to use some zoom. It is easier for me to deal with if I have this. So let us say that this is the first point we want to mark and this is nothing but 100 in x and 0 in y. Okay, so we know that this is the point and this point is 350 and 0, so let us mark this.

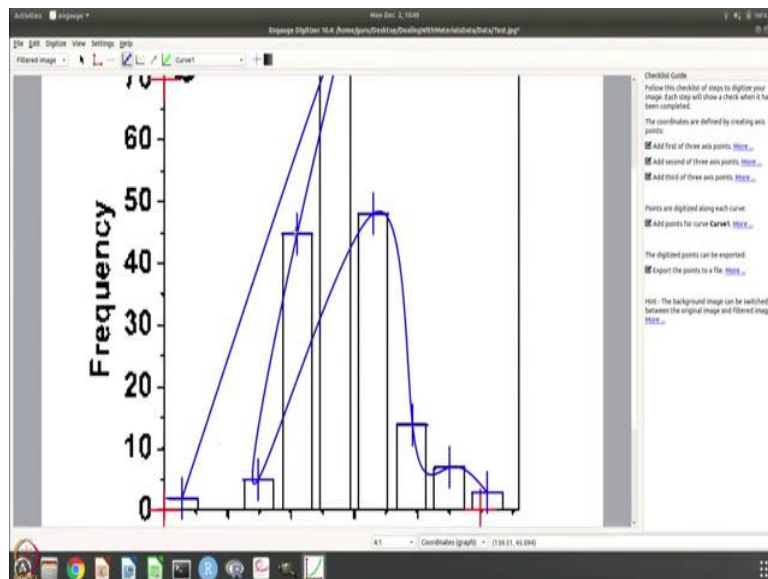
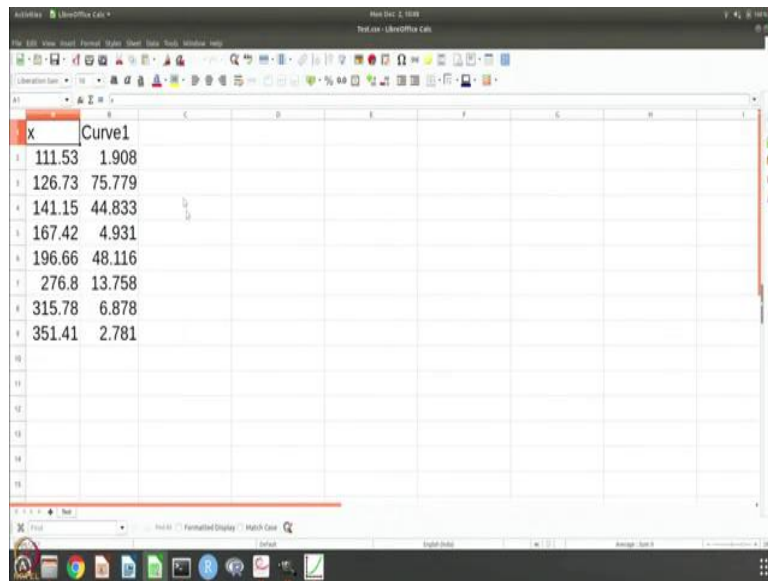
So it is 350, y is 0, okay. And then we want to mark this one. So we know that this is 0 in x axis and 17 y axis. Okay, very good. So now, the engage digitizer knows that this is for example 350. So if I go here, it will know that it is 250. So, as you can see here, if I hover the mouse over these values, then it gives those numbers, right. So, for example, if I want to know what this point is, so, it tells me that it is about 205, right.

And what this point is, so it is about 175 okay. So, once we have defined 3 points is all sufficient. Now the curve point tool is the one which actually can identify the curve points, right. So, for example, let me just start putting points. So let us say that I have this point, I have this point, I have this point, I have this point, and I have this point, I have this point, I have this point and I have this point, right.

So, the program can actually get the points and you can then import the points.

(Refer Slide Time: 19:19)



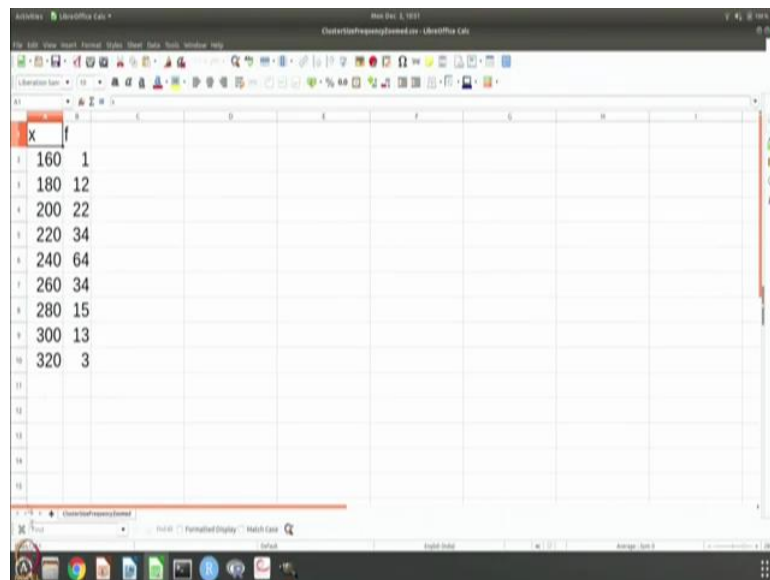
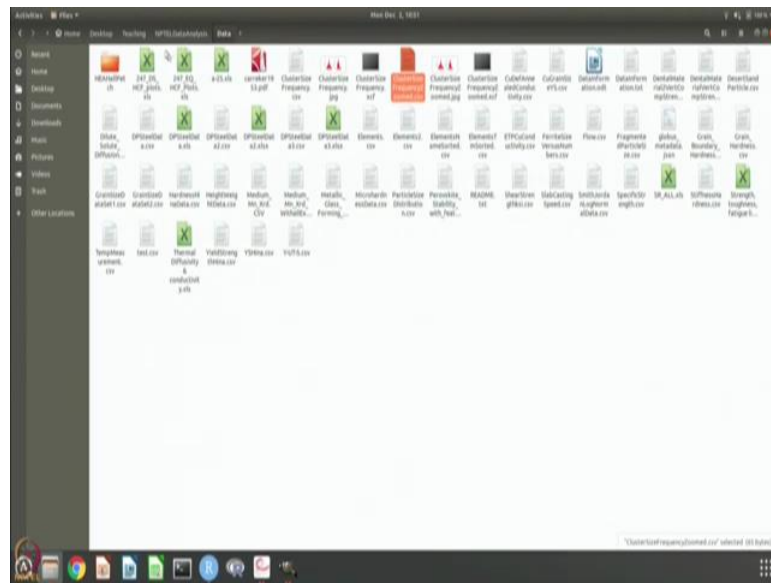


So, if you just say export and so test.csv, then it will save the data and you can open the test.csv. Okay, so it gives you the x y points, but this is not very useful for me, because, I mean of course, it puts the curve crazily, the curve should actually go like this.

So, what I did instead is to actually hover and read the data, like this is 175 and for 175 as cluster size, the frequency is about 5 right. And for this is 115 and the frequency turned out to be like 2 and this is 45 and that corresponds to 205 and so on.

So, in a similar fashion so and so this you can use for reading any digital data and it comes very handy. So, sometimes if you see data and literature and it is not in a form in which you can analyze it, you can use these programs. And but these are not the only ones. So, there are other programs that are available.

(Refer Slide Time: 21:10)



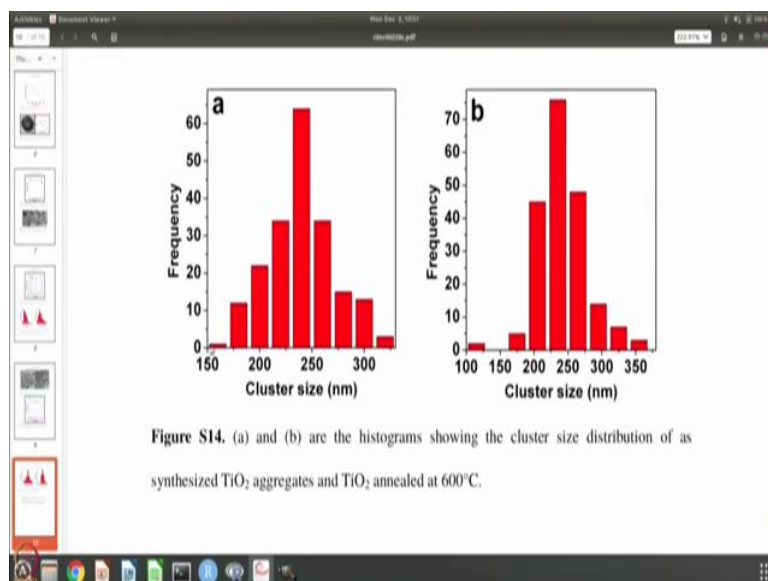
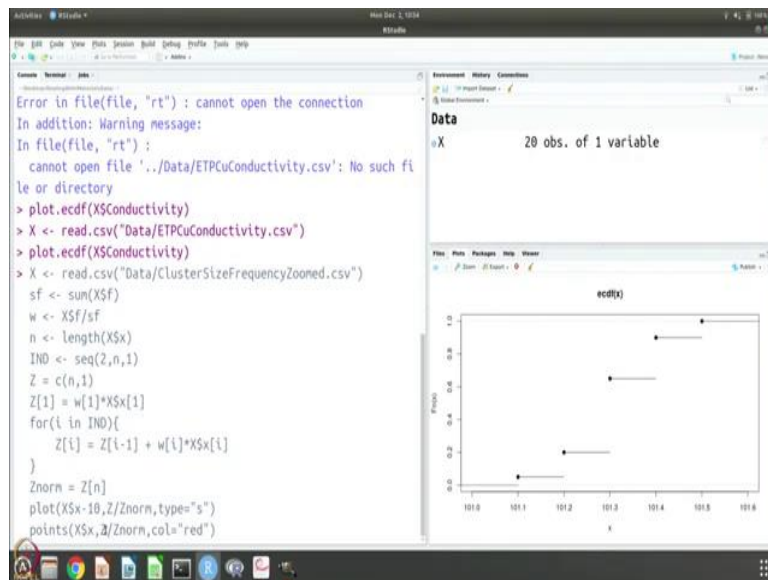
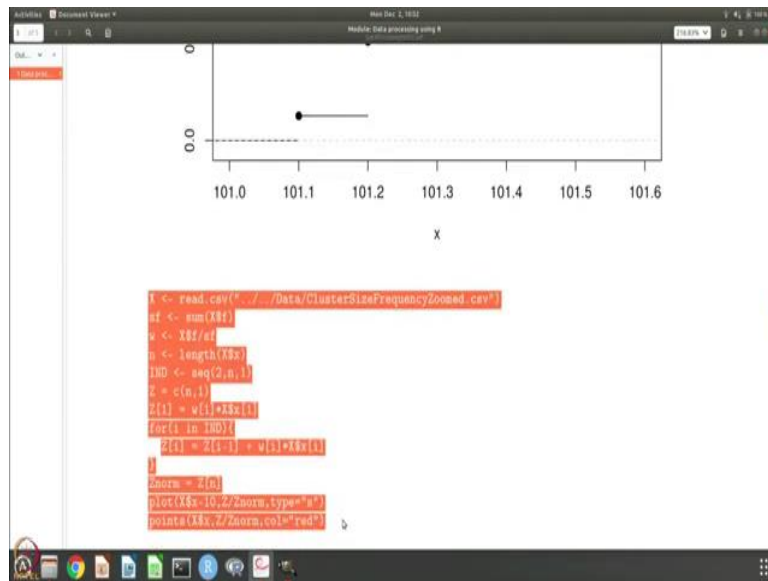


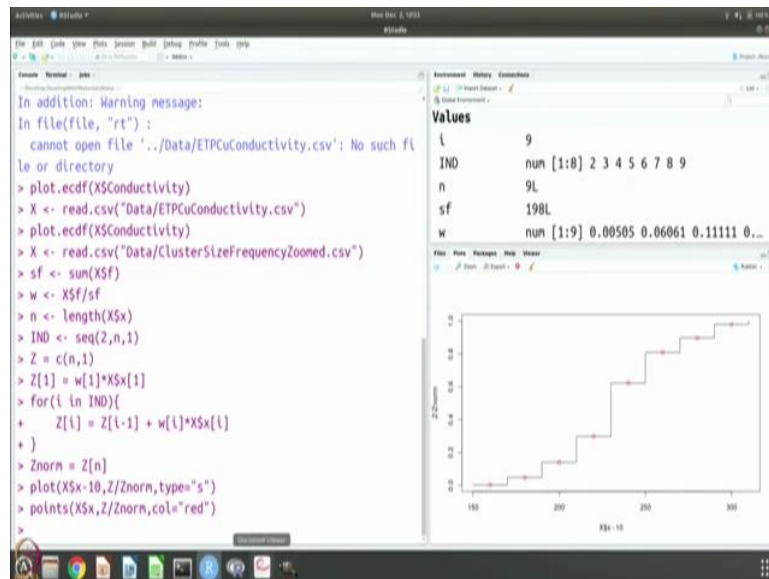
Figure S14. (a) and (b) are the histograms showing the cluster size distribution of as synthesized TiO_2 aggregates and TiO_2 annealed at 600°C .

So, what I have read, I have made such a data set and let me copy that data set. Yes. So, this is the csv file I need, cluster size, frequency, let me put it here. So, this is the data that I have got 160 to 320 in 20s, and 1, 12, 22, 34, etc, which you can see from this figure. So it is 1, 12, 22, etc. So it is like 160, 180, 200, 220, 240, 260, etc. up to 320. So, this is the data that I have digitized using GIMP and Engauge.

And that is the data that is shown here, x versus frequency. This is in nanometer, the size and the cluster size and the frequency of such clusters. So, that is what is the data that is given here. This is the data that we are going to use now and do our further analysis right.

(Refer Slide Time: 22:45)





So, let us see what is the analysis that we are trying to do. So, first thing you have to read the cluster size frequency data and sum of frequency is just sum of that column. And the weight is nothing but the particular number of observations in that bin divided by the total number of observations, so that is the weight.

And the number of bins that we have is given by the length of x, because remember it just had x and f as the 2 columns. And then we are going to make a vector called z and it has n rows. And we are also going to make another sequence which goes from 2 to n, 2 to n because z1 I am going to take as w1 times the x and then for other z, it is the previous z plus wi times that x.

So, this is for the cumulative distribution and you can normalize by the total, so that it goes to 1. That is why the last value is taken as zn and everything is divided by that. And when we are plotting we should remember that the x value is actually read, it has the spread of 20. So, when we have these values, we want to plot the cumulative distribution function in such a way that it has a step of that size.

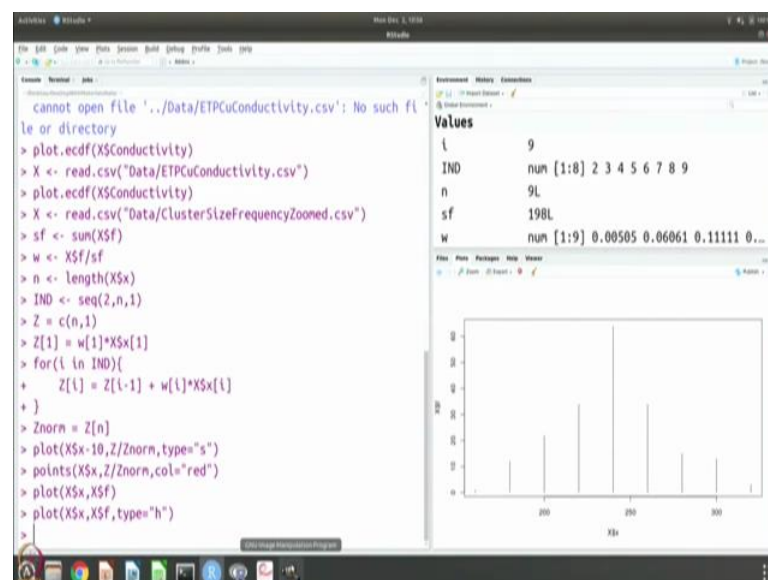
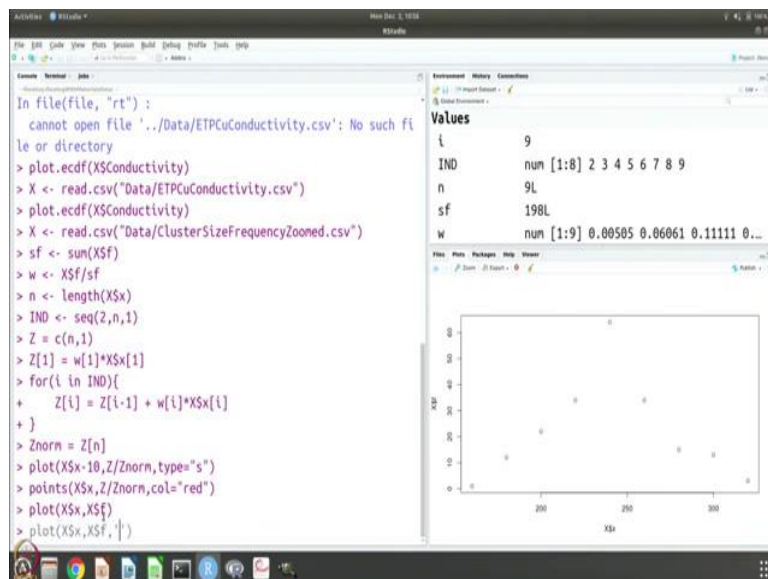
So, you do not want the value to jump at that value itself, but you want it to jump 10 after that. So this shifting is done to make sure that the step is of the right size. And I am also adding a point exactly at the center of that step to indicate that, that is where we have read the value. The step actually indicates the uncertainty in that value.

Anything that lies within that range will actually be binned to that point. So, that is what this indicates. So let us plot this and see. So, you this is what I said. So 160 is where we have taken

a data point, but we know that 160 actually means 150 to 170. So we want a step and that is the reason why this minus 10 is there.

So it will draw this line and it will show step then and wherever the actual values that we have read from the histogram because the histogram is giving these bins 160, 180, 200, etc. So I have put a red point to indicate that, that is where we have read the value, but this is the uncertainty or the bin size, all values that lie within this range are actually clubbed and put it at this value. So, this is the plot.

(Refer Slide Time: 26:26)

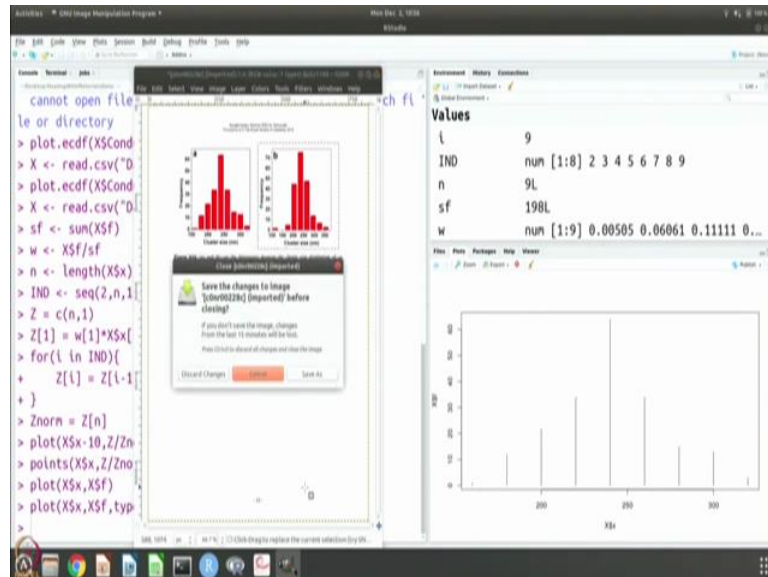


So, obviously, it will be more helpful if you can ofcourse, we have seen the, seen the data in histogram form and you can also plot this one to actually see the histogram. And in fact, if you

want to get the histogram feel, you can also make it type yeah so, so, you can see that the heights of these lines are equal to the frequency.

So, it does show you this nice histogram which was what was there in the data itself, right.

(Refer Slide Time: 27:05)



So, we have seen this data and so this is basically the same thing. So, we have read these values and we have plotted it here. Okay, so, we can do that, the more useful thing is to of course, do the y axis to be a probability scale.

(Refer Slide Time: 27:24)

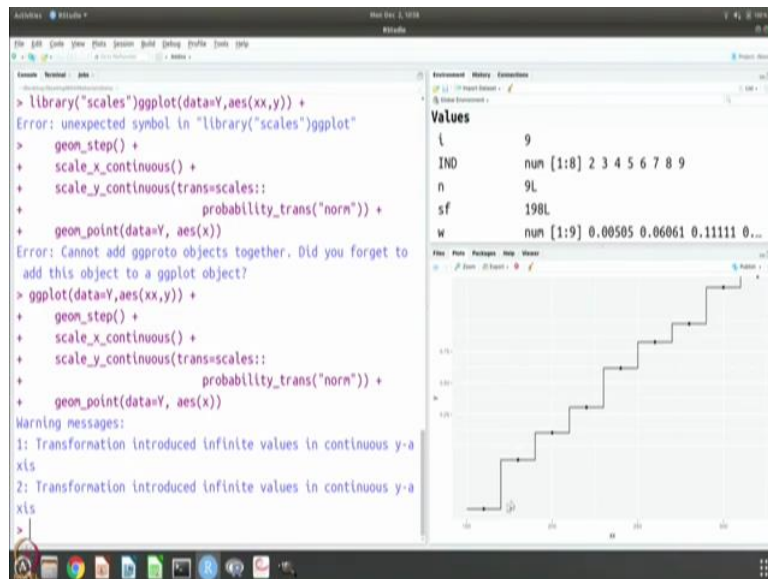
```
150      200      250      300
X$x - 10

#<- read.csv("../Data/ClusterSizeFrequencyZoomed.csv")
#f <- sum(X$f)
#v <- X$f/#f
#n <- length(X$x)
#IND <- seq(2,n,1)
#Z <- c(n,1)
#Z[1] = v[1]*X$x[1]
for(i in IND){
  #Z[i] = Z[i-1] + v[i]*X$x[i]
}
#Znorm = Z/n
#Y <- data.frame("x" = X$x, "xz" = X$x-10, "y" = Z/2norm)
library("ggplot2")
library("scales")
```

```
library("scales")

ggplot(data=Y,aes(xz,y)) +
  geom_step() +
  scale_x_continuous() +
  scale_y_continuous(trans=scales:
probability_trans("norm")) +
  geom_point(data=Y, aes(xz,y))

## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Transformation introduced infinite values in continuous y-axis
```



So we are going to do that. Okay, so let us take a look at the command that we have. So let us go back here. So, it is the same, some the frequencies make the weights, find out the number of data points and generate the cumulative and then normalize it and then plot it. So, that we have done already, but the only extra thing is to use GG plot now, and use that to scale the y axis.

So, that is the command that we have done here. So, we have made this new data called Y. And what is Y, Y is just a data frame it takes x as x and X-x as x minus 10 because remember we wanted to make the steps, y as z by z norm so that the values will go to 1. So, this data frame now we take and plot and while plotting, we have the GG plot.

So it is x-x axis y. So the geometry is of step, so it will do this step plot. And then we scale the y axis to be normal probability scale. Then we just add the points at these centers like we did in the previous case. So, you can see that this is more or less a sort of straight line, indicating that this data might also be normal.

So, to summarize, we can have data and if it is raw data, we can deal with it directly. If the data is only in published papers, if you have access to the PDF files, it is possible to generate the some form of data from those plots yourself. There are lots of tools that are available online for you to do that.

And both the Gimp and Engauge digitizer that I showed you are freeware, so you can freely download them and use them on your computers to get the data from the paper into a digital format. So, you can then use Libre Office which is another freeware to have the data entered in csv format.

Once you have, ofcourse the data in csv format you can use R to do all the analysis and we have shown one example of how to plot the histograms and cumulative distribution functions using R from a data that is given only in the form of a histogram plot.

And the histogram plot also tells us how to deal with data which has different statistical weights. So all the points in that data do not have the same weight. Some values for example, have frequency 1 or 2, at some values there are about 50 or 60. And we know that when we have one bin, all values which lie in the range of that bin are actually binned into that single bin.

So there is an uncertainty in the numbers. When we say 200 nanometers, it is 190 to 210 nanometers, any aggregate in that size will actually be counted in that bin. So, we need to give different weights and we have to understand that each bin has an uncertainty in terms of the actual value itself. But this is very common, I mean, hardly ever you will get raw data of all the cluster sizes for example.

And so, once you have this kind of data, using this it is possible to proceed with the analysis and that is an example that we have shown and you will have more exercises this week to do similar analysis from data of similar type that we will give. Thank you.