

## **Dealing with Materials Data: Collection, Analysis and Interpretation**

**Professor M. P. Gururanjan**

**Professor Hina A Gokhale**

**Department of Metallurgical Engineering and Materials Science**

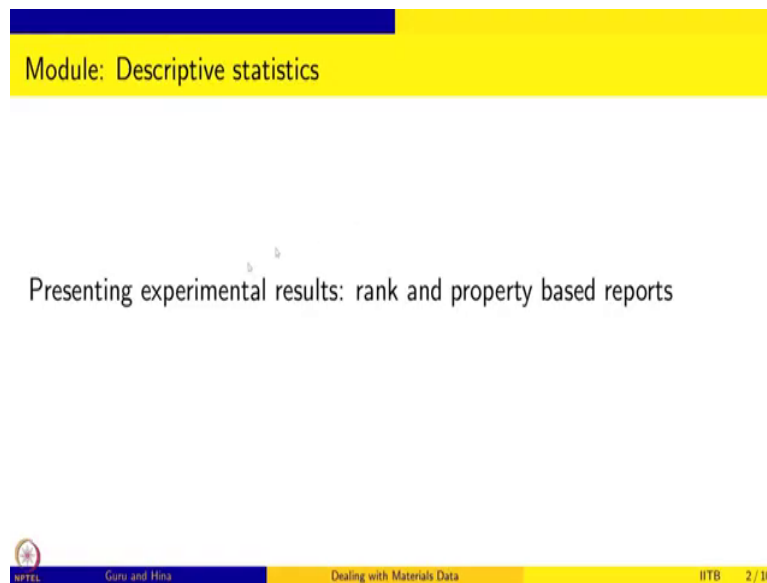
**Indian Institute of Technology, Bombay**

**Lecture No 20**

### **Presenting experimental results: Data on conductivity of ETP copper**

Welcome to Dealing with Materials Data. In this course we are going to learn about Collection, Analysis and Interpretation of data using materials data sets as examples and specifically in this sessions, we are going to learn about using R to do this.

(Refer Slide Time: 0:41)



Module: Descriptive statistics

Presenting experimental results: rank and property based reports

NPTEL Gurni and Hina Dealing with Materials Data IITB 2/10

And, this is the second module, this is the module on Descriptive Statistics and in this module we are going to learn about presenting experimental results and specifically we are going to prepare rank and property based reports of experimental results. So, we are going to take some data sets and prepare these reports and we will learn how to prepare them and as well as how to report them. So, that's what this session is going to be.

(Refer Slide Time: 1:03)

The slide features a yellow header with the title 'Electrical conductivity of ETP copper'. Below the header, there is a list of four bullet points, each preceded by a blue circular icon containing a white number. The footer of the slide is divided into three colored sections: a blue section on the left with the IITB logo and the text 'IITB', a yellow section in the middle with the text 'Guru and Hina', and a blue section on the right with the text 'Dealing with Materials Data' and 'IITB 3 / 10'.

- 1 ETP: electrolytic tough pitch copper;
- 2 Conductivity measurement: eddy current method;
- 3 Units: % IACS (International Annealed Copper Standard);
- 4 Consider the data on copper conductivity;
- 5 Measurements carried out by Dr N Harshavardhana, and reported in his PhD thesis submitted to IITB.

Okay, as the first example we are going to use the electrical conductivity of ETP copper. ETP copper is electrolytic tough pitch copper and it is very pure copper, it is commercially pure copper and it is used in many practical applications where its conductivity is very important. So, typically in the industrial setting, the conductivity is measured using the eddy current method.

And the units of the conductivity is given in percentage IACS. IACS stands for - International Annealed Copper Standard. So, with respect to this standard what is the conductivity that is measured in the sample, that is what is given in this conductivity measurements. And so we are going to consider the data on ETP copper conductivity and these measurements were carried out by Doctor N Harshavardhana and these are reported in his PhD thesis submitted to IIT, Bombay.


So, we are going to use this data set and so this is one measurement. This is some 20 times he has measured in different parts of the sample and reported as the table so, that is the data that is given here.

(Refer Slide Time: 2:30)

ETP copper: conductivity data

101.4	101.3	101.3	101.4	101.2
101.3	101.4	101.5	101.2	101.3
101.2	101.3	101.4	101.3	101.3
101.5	101.4	101.3	101.3	101.1

20 measurements of ETP copper using eddy current method; the values are in % IACS.  
Data stored in ETPCuConductivity.csv

 RPTEL Gurus and Hina Dealing with Materials Data IITB 4 / 10

So, using eddy current method and what is reported is percentage IACS and so 20 different measurements gives 101.4, 101.3, 101.4 and so on and so forth. So, this is the raw data and this is the most complete reporting of data. We have made 20 measurements and each of the measurement value is given and typically it is also given in the same order in which the measurement has been made.

So, first measurement is 101.4, second measurement is 101.3, third measurement is 101.3 and so on and so forth. And if these are made on different parts of a sample and sometimes you can even give a schematic of the sample and locate where the first, second, third etcetera the measurements are made. These are all made on same piece of copper and so the idea is to do these measurements to get the conductivity of the sample.

And as usual we want to deal with this data so we want to store it as a csv file and that is done we have a file called ETPCuconductivity dot csv. And is the file that we are going to use when we are going to do the R program.

(Refer Slide Time: 3:54)

The slide is titled "Presenting data" and contains the following content:

- The most complete: report all the numbers you obtained – as we saw in the table;
- Not practical: we are expected to do data reduction and report the reduced data with the methodology;
- Current standards: do a data reduction analysis, report the reduced data with methodology, and, also share the 'raw' data (along with the relevant codes, scripts etc) as supplementary material

The footer of the slide includes the IITB logo, the text "Guru and Hina", "Dealing with Materials Data", and "IITB 5 / 10".

So, once the data is given, of course the most complete is to just present the table and this table is actually there in Doctor Harshavardhana's thesis. So, he has even reported but it is not possible to keep reporting numbers like this for every measurement that you make. His thesis for example, contain so many conductivity measurements and this is the first and only time where the complete measurement is given.

Just to give an idea to the reader as to, how these measurements are made, what the numbers look like and when we report our means and standard deviations and so on and so forth later, people can understand, what is the type of data that we are dealing with?. So, that is the purpose so it is always better to give the complete report, all the numbers if possible.

But many a times it is not practical so you will see in Doctor Harshavardhana's thesis that apart from this, there are not many places where the repeated measurements, values are given and because it is not practical to keep giving numbers like this. We are expected to do data reduction and report the reduced data. But, it is important then also to tell, how the reduction itself has been carried out.

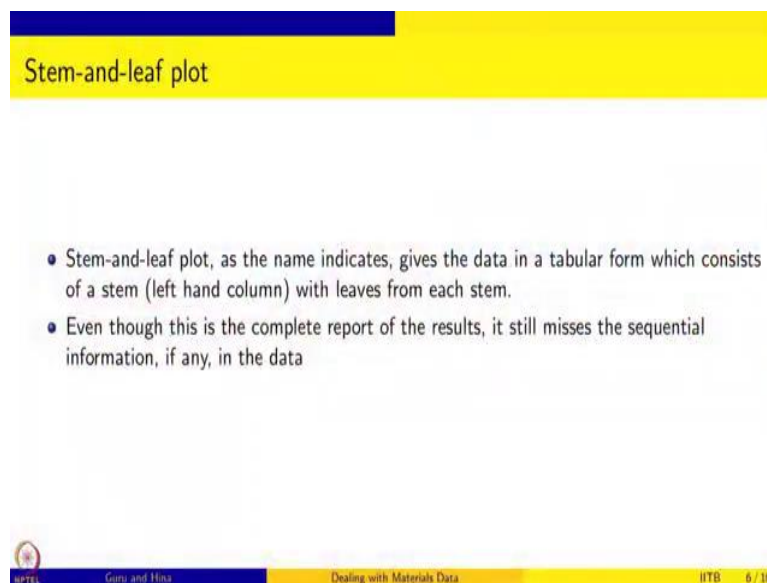
So, you have to give the reduced data along with the methodology that you used to reduce the data so that is the most common thing that is done in scientific reports thesis, papers and so on and so forth. However, having said that the current standards are also changing, nowadays when you do a data reduction analysis and report the reduced data with the methodology. You are also expected to share the raw data that is the like the data that we saw in the table the complete data along with relevant codes and scripts etc. that you have used.

This is typically given as a supplementary material and this is good practice and as you will also see later we are going to use some of these data that is there in the open literature for carrying out our own analysis to understand some of these methodologies to learn about dealing with materials data. It also allows others to actually carry out the same analysis or if they have a different methodology, then they can apply to this data.

So, it is very important to have this raw data available and so the most preferred current standard is not just do data reduction analysis and report the methodology but also store the raw data at some place along with the scripts and codes used for the reduction. And make it available to everybody so that people can independently do the same analysis or if they need or want to do a different analysis.

And we will also show you some papers in the recent literature which do this very well, which do a commendable job of reporting this raw data and those should be the standards we should aspire to.

(Refer Slide Time: 7:22)



Stem-and-leaf plot

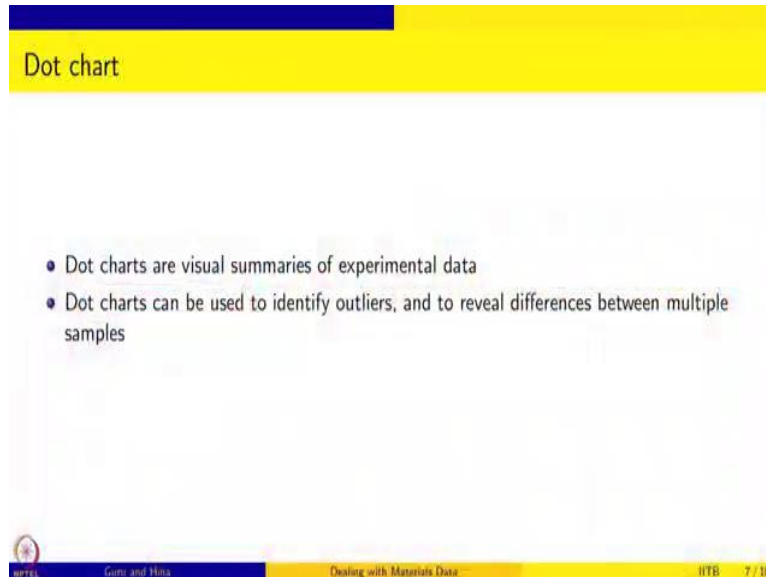
- Stem-and-leaf plot, as the name indicates, gives the data in a tabular form which consists of a stem (left hand column) with leaves from each stem.
- Even though this is the complete report of the results, it still misses the sequential information, if any, in the data

HPTEL Gurs and Hira Dealing with Materials Data IITB 6 / 10

So, the first reduced data is the stem-and-leaf plot and in fact stem-and-leaf plot has the same amount of information as your table data. Except that there is still some modification that is done. It loses the information on the order in which the data was obtained. The stem-and-leaf plot as the name indicates gives the data in a tabular form and it consist of a stem which is a left hand column. And there are leaves from each of the stems. So, we will show for the conductivity data how this looks like.

So, like I said even though it is a complete report of the results, it is still misses the sequential information if any in the data. So, we are going to order the data and then we are going to plot stem-and-leaf which means we are going to lose the information about the original order in which the data was obtained.

(Refer Slide Time: 8:15)



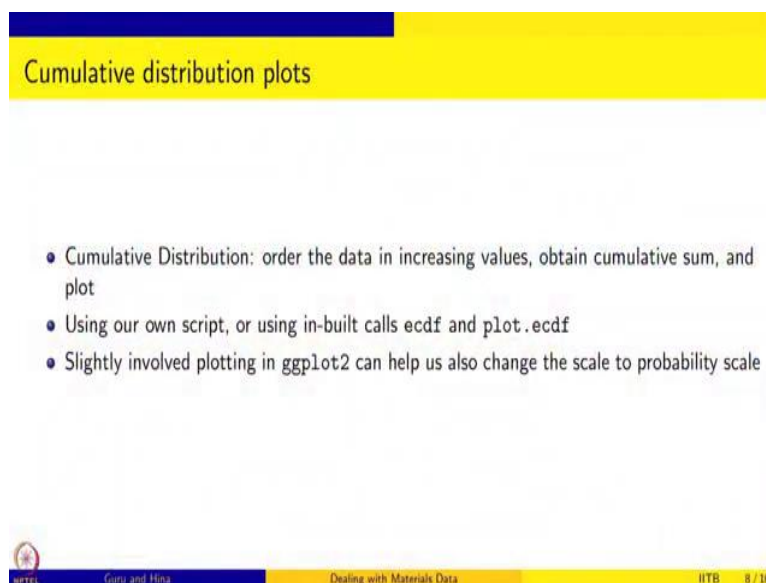
Dot chart

- Dot charts are visual summaries of experimental data
- Dot charts can be used to identify outliers, and to reveal differences between multiple samples

NPTEL Gouri and Hima Dealing with Materials Data IITB 7 / 10

Dot chart is another way of presenting results again it is going to give the complete information. But, they are nice visual summaries of experimental data and they can be used to identify outliers and if you have more than one data set, they will also reveal the differences between the different data sets. So, we are going to see examples of both as we go along in the course.

(Refer Slide Time: 8:50)



Cumulative distribution plots

- Cumulative Distribution: order the data in increasing values, obtain cumulative sum, and plot
- Using our own script, or using in-built calls `ecdf` and `plot.ecdf`
- Slightly involved plotting in `ggplot2` can help us also change the scale to probability scale

NPTEL Gouri and Hima Dealing with Materials Data IITB 8 / 10

And the third way of presenting data is, to give cumulative distribution plots. Here, again we order the data in increasing values and we obtain the cumulative sum and we plot and there are several ways of doing cumulative distribution plots. You can use your own script or you can use in-built calls like `ecdf` or `plot dot ecdf` and so on to get the cumulative distribution.

And you can also do slightly involve plotting in the `ggplot` which helps us change this scale of the y-axis in this cumulative distribution plots to probability scale. To know whether the data follows the normal distribution or not. So, this might be important in some cases. We will see examples of that and we will also so there is more involved analysis that one can do. This is just for step changing the y-axis to probability scale but, but we will do. We will see an example of, how to do this.

(Refer Slide Time: 09:50)

Histograms and box-and-whisker plots

- We can bin the data and plot histograms;
- Or, we can produce box-and-whisker plots which indicate the distribution of the data.
- Or, use commands like `quantile` to get information on the spread of data!

NPTEL Guru and Hina Dealing with Materials Data IITB 9 / 10

And the next set of plots that we present are histograms and box-and-whisker plots. So, you can bin the data and present histograms, so in this range, how many measurements have shown up or in the next range, how many measurements have shown up and so on and so forth. These histogram plots are very-very important specifically if the distribution is not normal or is not what people expect or if you want to give explicit information about the distribution then histogram plots are important.

Box-and-whisker plots also have similar information. They indicate the distribution of the data and you can also use commands like `quantile` to get these spread of the data.

(Refer Slide Time: 10:38)

Property based reports

- Mean, median and standard deviation
- Can combine with graphical methods to better understand the data

NPTEL Guru and Hina Dealing with Materials Data IITB 10 / 10

And finally, there are also property based reports that one has to make or one can make. These are mean, median and standard deviation variance so on and so forth and in these tutorials we will also see in addition to getting the property based reports how to combine the property based reports along with the rank based reports in a graphical methodology.

So, you plot them you also put this information of the property based reports on the same plots and that gives us better information about the data or help us understand the data better. So, we are going to do all this, so this is session on reporting rank and property based data. So, we are going to use the electrical conductivity of ETP copper as the example case for doing all this analysis. So, let us do that.

(Refer Slide Time: 11:46)

Let us consider the data on the electrical conductivity of copper.

101.4	101.3	101.3	101.4	101.2
101.3	101.4	101.5	101.2	101.3
101.2	101.3	101.4	101.3	101.3
101.5	101.4	101.3	101.3	101.1

This data is stored as a csv file: ETPCuConductivity.csv. Let us load the data in R and plot a stem and leaf plot to report the complete data:

```
R <- read.csv("Data/ETPCuConductivity.csv", header=TRUE)
stem(I$Conductivity)
```

```
##
## The decimal point is 1 digit(s) to the left of the |
##
## 1011 | 0
## 1012 | 000
## 1013 | 0000000000
## 1014 | 000000
## 1015 | 00
```

Note that even though the stem-and-leaf plot gives all the numbers, the actual order of data is lost in such a presentation. So, this data is already reduced even though all the 20 numbers are reported. From the stem-and-leaf plot, it is also clear that the mode (number appearing the most often) is 101.3.



```

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> X <- read.csv("Data/ETPCuConductivity.csv",
+             header=TRUE)
> X

```

```

12 101.3
13 101.4
14 101.3
15 101.3
16 101.5
17 101.4
18 101.3
19 101.3
20 101.1

> x<-X$Conductivity
> stem(x)

The decimal point is 1 digit(s) to the left of the |

1011 | 0
1012 | 000
1013 | 000000000
1014 | 000000
1015 | 00

```

So, first thing to do, so let us open R and let us read the data and make a stem-and-leaf plot. Well it is very easy, so we are going to read into X the data on ETP conductivity and so let us first do that and then we are going to say X so this is the data. So, I am going to say X we are going to save in small x, the conductivity data.

So, now if you say stem x. You get this stem-and-leaf plot as you can see the decimal point is 1 digit to the left of the pipe. So, pipe is here so decimal point is here that we know the data is 101.3, 101.4 etcetera so 101.1, 101.2, 101.3 etcetera. And you can see that there is one data point 101.1. There are three 101.2, there are nine 101.3 and five 101.4 and two 101.5.

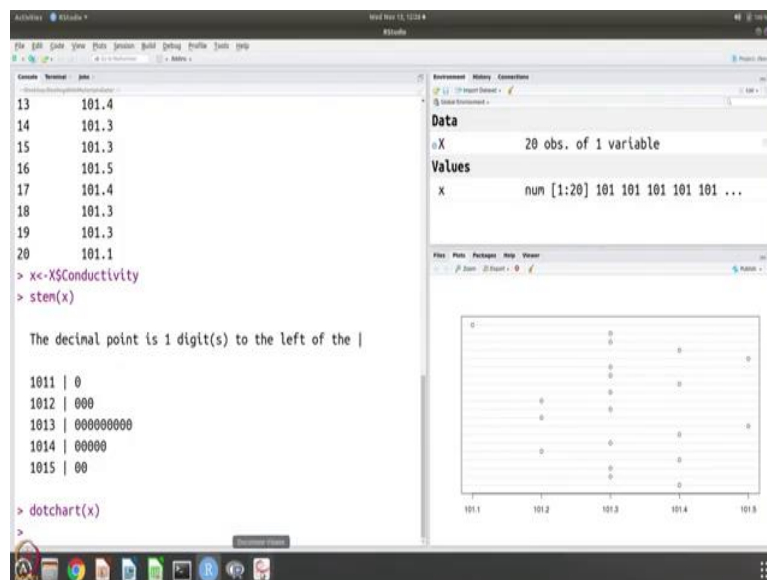
So, if you add them all up there is total of 20 data points. 4, 9 plus 4 13 and plus 5 18 plus 2 20. So, this is the stem-and-leaf plot and this is called the stem and this is called the leaf and you can also see sort of the distribution of the data. So, you can see, how the data is distributed

and it looks like normal distribution at least looking at these numbers. So, this is the first one, so get us stem-and-leaf plot.

So, like I said stem-and-leaf plot is complete in the sense that whatever 20 data points that we saw are all here except now that they are ordered. This information that you know 17 measurement gave 101.4 and then 18 and 19 gave 101.3 and 20 gave 101.1 that information is missing here. The sequential information, if it is important for example, 0.5, 0.4, 0.3, 0.1 is there some reason why if you make these measurements it will reduce like this. If there is any such information that is missing from the stem-and-leaf plot otherwise it has all the data.

So, it is complete in one sense that it has all the data and it also by looking at it you cannot only see how the data is distributed, you can also see which is the, you know most repeated number, the mode of the data so that you can see clearly.

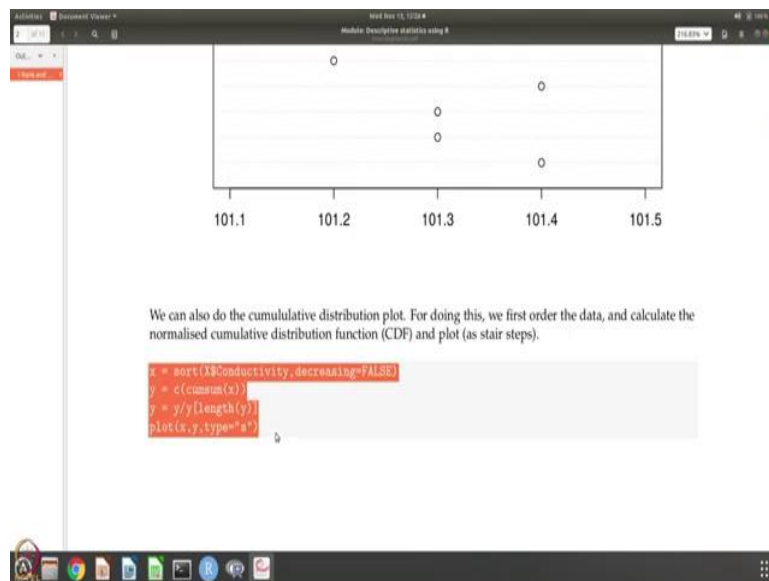
(Refer Slide Time: 14:56)



Now, let us do the dot chart so dot chart of x. Now, here is the dot chart and again 101, 101.2, 101.3 etc. so you can see two data points here, five data points here, nine data points here and three data points here and one data point here. The dot chart is also plotted in such a way that it tells you about the values and 101.1 is shown like this because it is sort of an outlier.

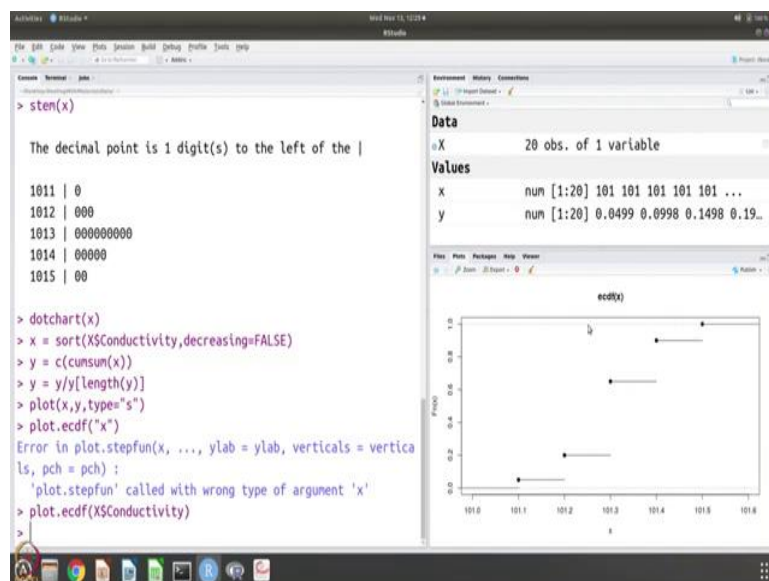
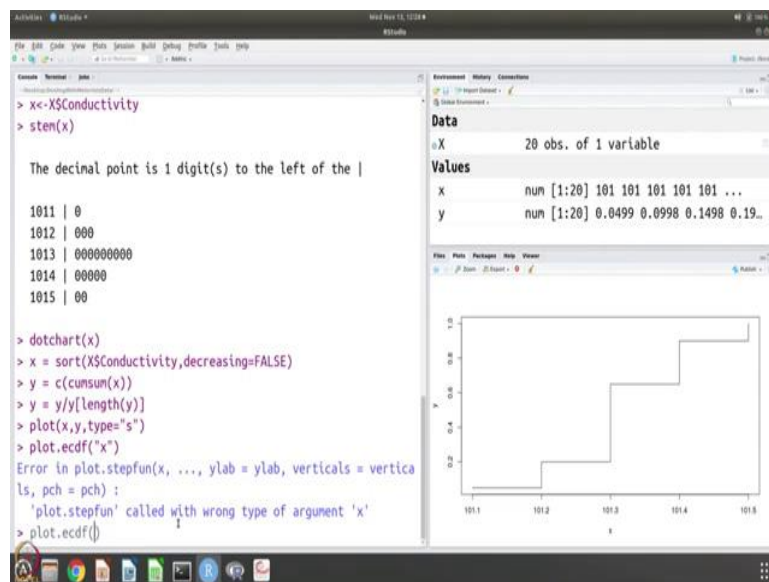
And dot charts are useful to identify such outliers and so we will see why this is an outlier, we will see later. But, at least by looking at the dot chart again dot chart also gives all the information. It only loses the information about the sequence of the measurements but in addition it also sort of shows you where the outliers are and in this case this happens to be within the outlier.

(Refer Slide Time: 16:10)



We can also do the cumulative distribution plot. For doing this, we first order the data, and calculate the normalised cumulative distribution function (CDF) and plot (as stair steps).

```
x = sort(X$Conductivity,decreasing=FALSE)
y = c(cumsum(x))
y = y/length(y)
plot(x,y,type="s")
```



```

The decimal point is 1 digit(s) to the left of the |
1011 | 0
1012 | 000
1013 | 00000000
1014 | 000000
1015 | 00

> dotchart(x)
> x = sort(X$Conductivity,decreasing=FALSE)
> y = c(cumsum(x))
> y = y/[length(y)]
> plot(x,y,type="s")
> plot.ecdf("x")
Error in plot.stepfun(x, ..., ylab = ylab, verticals = verticals,
+ ls, pch = pch) :
+ 'plot.stepfun' called with wrong type of argument 'x'
> plot.ecdf(X$Conductivity)
> help(ecdf)
> plot(ecdf(X$Conductivity))

```

**Data**

nX	20 obs. of 1 variable
<b>Values</b>	
x	num [1:20] 101 101 101 101 101 ...
y	num [1:20] 0.0499 0.0998 0.1498 0.19...

**Empirical Cumulative Distribution Function**

**Description**

Compute an empirical cumulative distribution function, with several methods for plotting, printing and computing with such an "ecdf" object.

**Usage**

```
ecdf(x)
```

**ecdf(x)**

```
# S3 method for class "ecdf"
plot(x, ..., ylab="ecdf", verticals = FALSE,
     col.fill = "gray80", pch = 10)

# S3 method for class "ecdf"
print(x, digits=getOption("digits") - 2, ...)
```

**Arguments**

```
x, x[1:n] numeric vector of the observations for ecdf. For the methods, an object inheriting from class "ecdf".
```

The next step is that so we have done the dot chart. Let us do the cumulative plot and this is how the cumulative plot is done. So, let us put it here. So, we are going to sort the conductivity and decreasing is false, so it is going to be in the increasing order, lowest to highest and we are going to get the cumulative sum as y of x and we are going to normalise.

So, length of y will give you what is the number of elements in that vector and so the last point of that because it is cumulative sum you know final sum will be the total. And that we are going to divide by so that the values go from 0 to 1 and then we are going to plot the x value and the sorted conductivity value with the cumulative sum, normalised cumulative sum. And the type is basically step like plot and that is what type equal to s means.

So, let us do this and look at so as you can see, 101 these steps is what because we say type equal to s and this is the conductivity values and the cumulative sum of the conductivity values are here. So, this is the CDF, there are other ways of plotting this for example, you can say plot dot ecdf that stands for Empirical Cumulative Distribution Function, I think. So, you can say CDF, ok so sorry, **plot ecdf X dollar conductivity**.

So, you can see it is the same plot as we got except that the plotting style is slightly different and it shows you the Empirical Cumulative Distribution. So, so you can look **at help ecdf**, so let us Empirical Cumulative Distribution Function so that is what we got. That is another way of course you can say plot and what should be plotted we can say, Empirical Cumulative Distribution Function of conductivity.

So, again it is the same plot so either you can say plot dot ecdf or plot ecdf of this. So, there are two, three different ways of doing it but they all give you the same result namely that you have the cumulative distribution plot from the data.

(Refer Slide Time: 19:17)

The top screenshot shows a plot with a single data point at x=101.1 and y=0.0. The x-axis ranges from 101.0 to 101.6. Below the plot, the following R code is shown:

```
library("ggplot2")
library("scales")
ggplot(data=X,aes(Conductivity)) +
  stat_ecdf() +
  scale_y_continuous(trans=scales::
    probability_trans("norm"))
```

A warning message is displayed: **Warning: Transformation introduced infinite values in continuous y-axis**.

The bottom screenshot shows a terminal window with the following R code:

```
> dotchart(x)
> x = sort(X$Conductivity,decreasing=FALSE)
> y = c(cumsum(x))
> y = y/length(y)
> plot(x,y,type="s")
> plot.ecdf("x")
Error in plot.stepfun(x, ..., ylab = ylab, verticals = verticals,
pch = pch) :
'plot.stepfun' called with wrong type of argument 'x'
> plot.ecdf(X$Conductivity)
> help(ecdf)
> plot(ecdf(X$Conductivity))
+ )
> library("ggplot2")
library("scales")
ggplot(data=X,aes(Conductivity)) +
  stat_ecdf() +
  scale_y_continuous(trans=scales::
    probability_trans("norm"))
```

The plot window on the right shows the resulting ECDF plot titled "ecdf(X\$Conductivity)". The x-axis is labeled "x" and ranges from 101.0 to 101.6. The y-axis is labeled "P-Value" and ranges from 0.0 to 1.0. The plot shows a step function with points at (101.1, 0.0499), (101.2, 0.0998), (101.3, 0.1498), and (101.4, 0.1997).

x	y
101.1	0.0499
101.2	0.0998
101.3	0.1498
101.4	0.1997

Activities RStudio

```

> plot(x,y,type="s")
> plot.ecdf("x")
Error in plot.stepfun(x, ..., ylab = ylab, verticals = verticals,
pch = pch) :
  'plot.stepfun' called with wrong type of argument 'x'
> plot.ecdf(X$Conductivity)
> help(ecdf)
> plot(ecdf(X$Conductivity))
+)
> library("ggplot2")
RStudio Community is a great place to get help:
https://community.rstudio.com/c/tidyverse.
> library("scales")
> ggplot(data=X,aes(Conductivity)) +
+   stat_ecdf() +
+   scale_y_continuous(trans=scales::
+                       probability_trans("norm"))
Warning message:
Transformation introduced infinite values in continuous y-axis

```

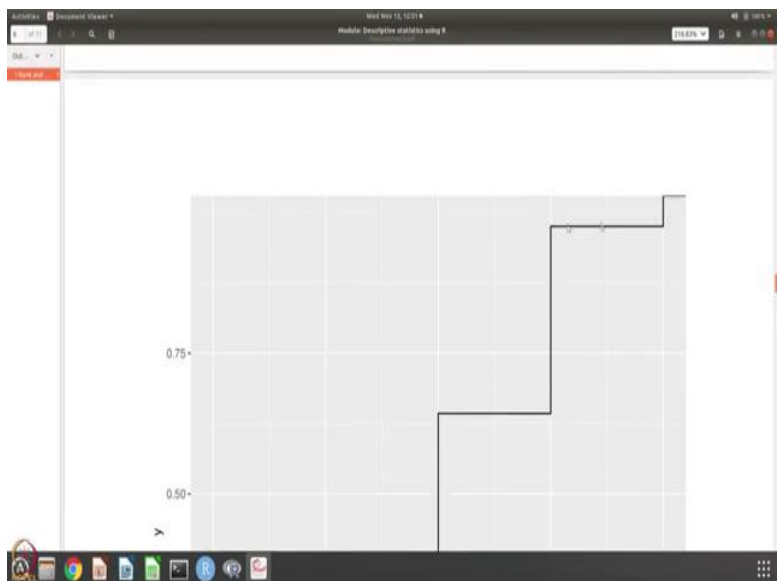
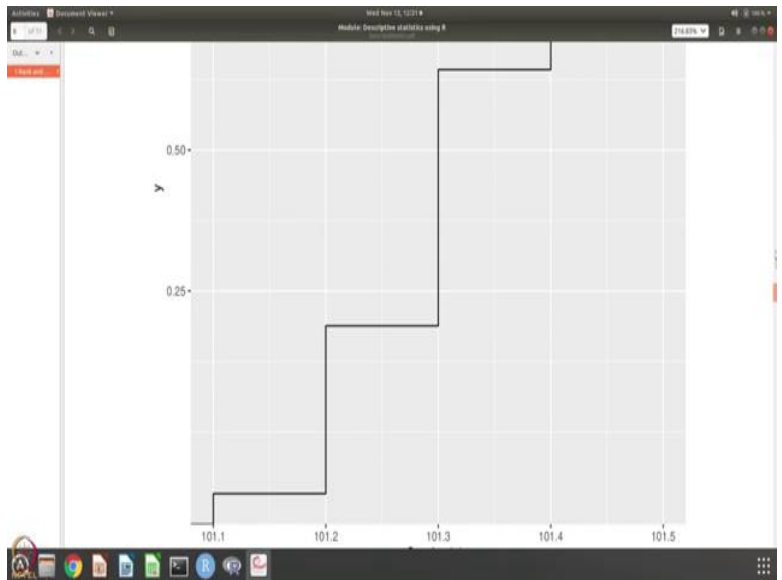
Environment History Connections

Data

```

>X      20 obs. of 1 variable
Values
x      num [1:20] 101 101 101 101 101 ...
y      num [1:20] 0.0499 0.0998 0.1498 0.19...

```



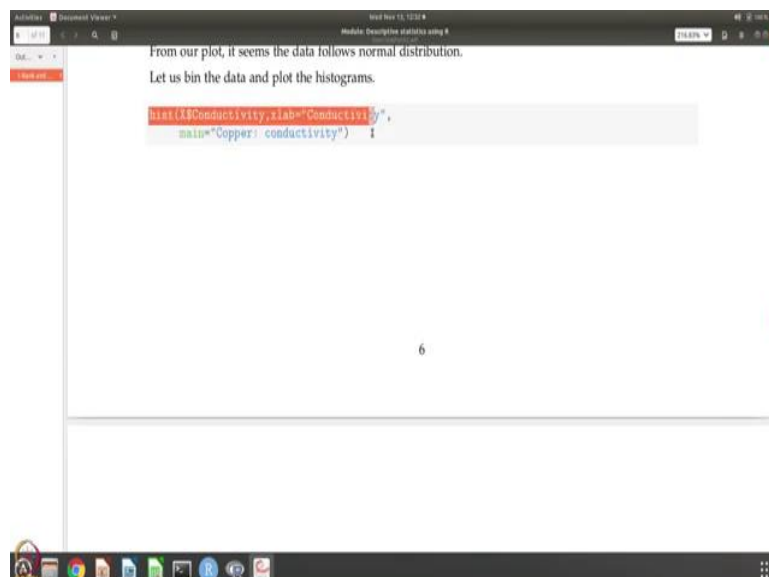
Now, we want to use ggplot, we want to change this range. This range to probability scale so obviously we have to use library ggplot and scales and so let us do that. So, we invoke the library ggplot2 we invoke the library scales and as usual ggplot you have to tell the data, we have to tell the aesthetics so conductivity is all that is there in the data so that is what we want to plot and we want to do the statistical analysis namely, Cumulative Distribution Function.

And that is what we want to plot and the scale of y we want to transform to probability scale and the probability scale if it is a normal distribution then, how the data would look like and what does the actual data look like. So, that is what we want to compare and that is why we want to do this scale transformation.

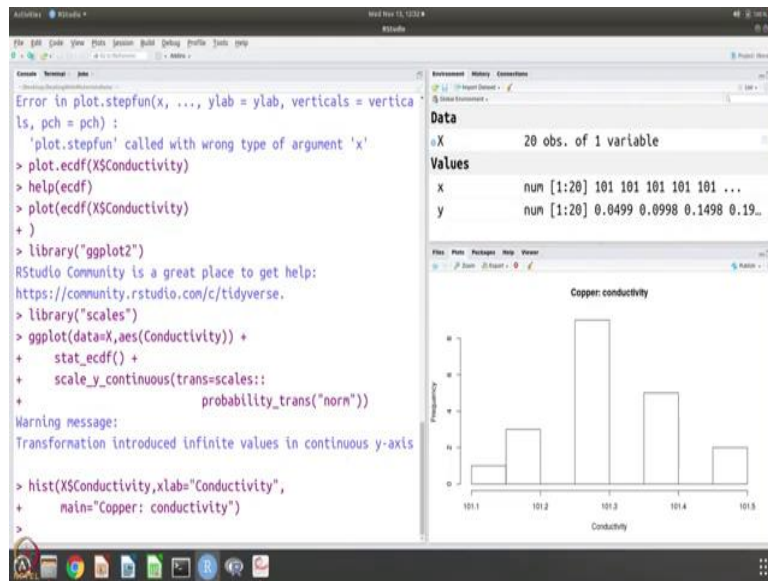
So, let us do this and you have this. And if this is sort of like a straight line that indicates that the data is having a normal distribution. But, anyway there is a warning message so transformation introduce infinite values in continuous y-axis but I do not think that is very important. But, it is important to read and pay attention to them, in this case this is not important.

So, as you can see the scale is now slightly different and so it basically tell us, 25 percent of the data is here and 50 percent of the data is here and 75 percent of the data is here and 100 percent data is here. So, it sort of gives you not just the data but the distribution of the data.

(Refer Slide Time: 21:19)

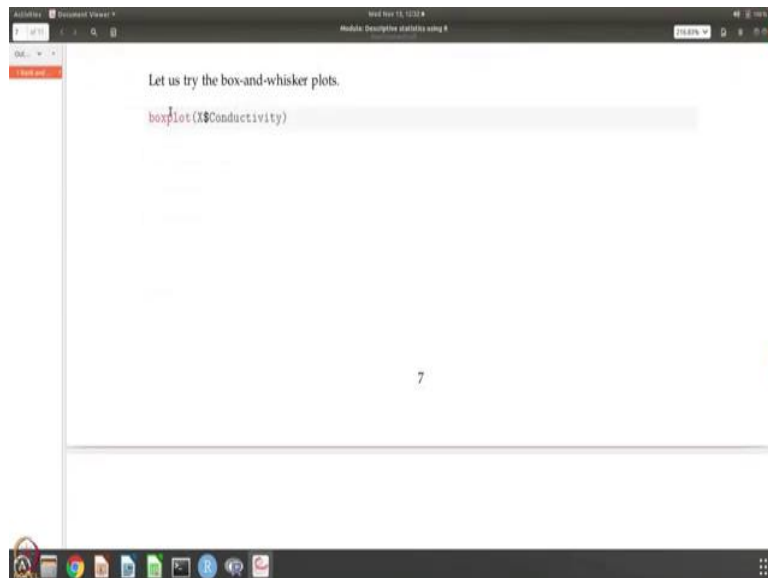




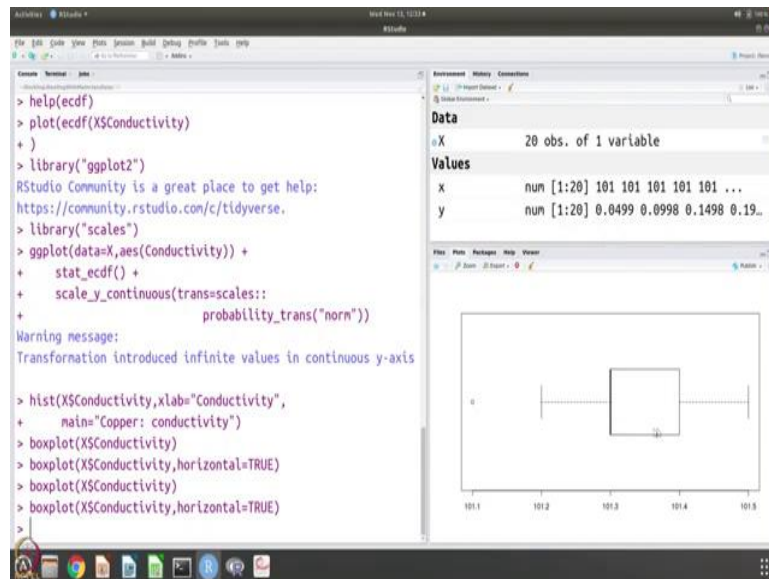


Now, the next one we want to do is to do a histogram plot, so let us do the histogram plot. So, we take X and conductivity of course x label is conductivity and the title is copper conductivity. You can see the histogram plots, histogram plots again are very nice so you can see the distribution of the data. So, there is one data point 3, 9, 5 and 2 that we already know and so that how the data is. So, this is a histogram plot.

(Refer Slide Time: 21:58)







Now, the next one that we wanted to do is to do a box-and-whisker plot. Again the command is very easy and as we have been doing it is easy to actually put help box plot for example and learn more about these commands. So, here is a box plot and the box plot again indicates where the data is, where the sort of median data lies and so on and so forth and we can also do, yes. So, what this has done? Is that it just flipped the, see the original box plot was like this and this 101.1 is an outlier as you can see here again and we can flip it to be horizontal so the conductivity values are here? And this means that this box actually has the second and third quantile data.

And so this is the median value and you can see that there is one data point which is a really lying outside. So, this is an outlier 101.1 as an outlier. We got a hint of this earlier too from the dot plot and we will see why that is so later.

(Refer Slide Time: 23:39)

We can also get the data spread information using the quantile command:

```
quantile(X$Conductivity)
```

```
## 0% 25% 50% 75% 100%  
## 101.1 101.3 101.3 101.4 101.5
```

```
quantile(X$Conductivity, probs=seq(0,1,0.10))
```

```
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
## 101.10 101.20 101.28 101.30 101.30 101.30 101.30 101.40 101.40 101.41 101.50
```

Let us prepare some property based reports and present them along with graphical representation of the data:

```
> library("ggplot2")  
RStudio Community is a great place to get help:  
https://community.rstudio.com/c/tidyverse.  
> library("scales")  
> ggplot(data=X, aes(Conductivity)) +  
+ stat_ecdf() +  
+ scale_y_continuous(trans=scales::  
+ probability_trans("norn"))  
Warning message:  
Transformation introduced infinite values in continuous y-axis  
  
> hist(X$Conductivity, xlab="Conductivity",  
+ main="Copper: conductivity")  
> boxplot(X$Conductivity)  
> boxplot(X$Conductivity, horizontal=TRUE)  
> boxplot(X$Conductivity)  
> boxplot(X$Conductivity, horizontal=TRUE)  
> quantile(X$Conductivity)  
0% 25% 50% 75% 100%  
101.1 101.3 101.3 101.4 101.5  
>
```

Data

x	20 obs. of 1 variable
x	num [1:20] 101 101 101 101 101 ...
y	num [1:20] 0.0499 0.0998 0.1498 0.19...

We can also get the data spread information using the quantile command:

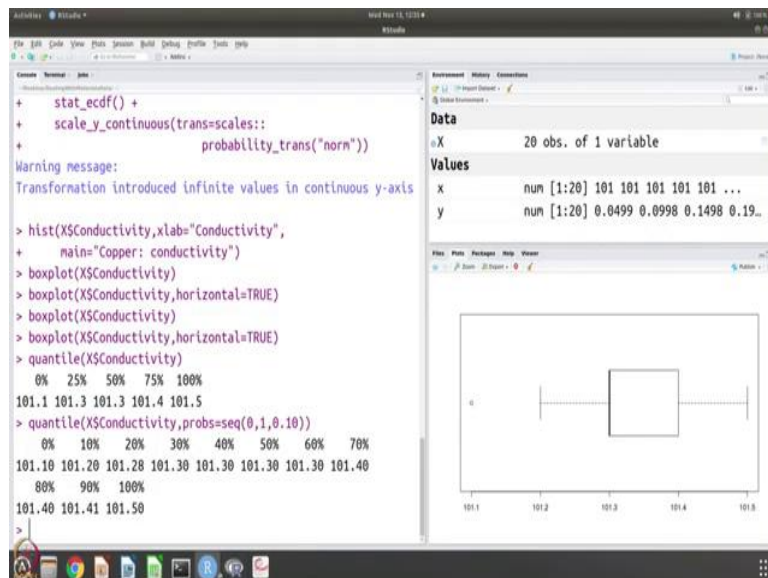
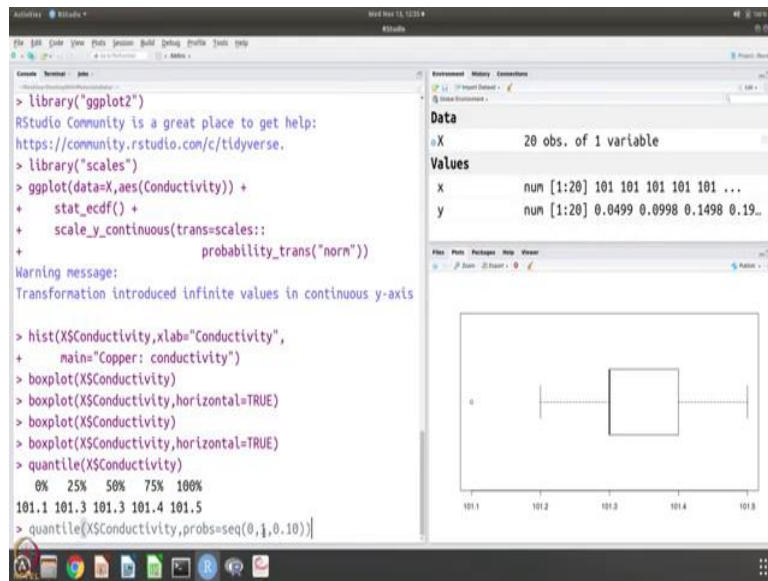
```
quantile(X$Conductivity)
```

```
## 0% 25% 50% 75% 100%  
## 101.1 101.3 101.3 101.4 101.5
```

```
quantile(X$Conductivity, probs=seq(0,1,0.10))
```

```
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
## 101.10 101.20 101.28 101.30 101.30 101.30 101.30 101.40 101.40 101.41 101.50
```

Let us prepare some property based reports and present them along with graphical representation of the data:



So, now let us also do this, you can also say quantile. This is not a graphical representation but it will give you the numbers. So, it says that 25 percent of the data is achieved at 101.3, 50 percent also at 101.3 and 75 percent of the data is actually 101.4 and 101.5 so, this is the quantile. How much of data is in which range.

You can increase so it gave you only 0, 25, 50, 75 etc. so you can decide that you want to have more information than that by explicitly giving that. So, here we again one quantile but instead of the default 0, 25 percent etc. we want to go in 10 percent so 0, 10, 20 etc. so you can get the quantile. So, this again gives you some information about the spread of the data. And these are all the rank based ways of representing the data.

(Refer Slide Time: 25:06)

The top screenshot shows a console window with the following content:

```
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
## 101.10 101.20 101.28 101.30 101.30 101.30 101.30 101.40 101.40 101.41 101.50
```

Let us prepare some property based reports and present them along with graphical representation of the data:

9

```
mu = mean(X$Conductivity)  
median(X$Conductivity)
```

```
## [1] 101.3
```

```
sigma = sd(X$Conductivity)  
CI1 = mu-2*sigma  
CI2 = mu+2*sigma  
var(X$Conductivity)
```

The bottom screenshot shows the RStudio interface with the following content:

```
> hist(X$Conductivity,xlab="Conductivity",  
+ main="Copper: conductivity")  
> boxplot(X$Conductivity)  
> boxplot(X$Conductivity,horizontal=TRUE)  
> boxplot(X$Conductivity,horizontal=TRUE)  
> boxplot(X$Conductivity,horizontal=TRUE)  
> quantile(X$Conductivity)  
 0% 25% 50% 75% 100%  
101.1 101.3 101.3 101.4 101.5  
> quantile(X$Conductivity,probs=seq(0,1,0.10))  
 0% 10% 20% 30% 40% 50% 60% 70%  
101.10 101.20 101.28 101.30 101.30 101.30 101.30 101.40  
 80% 90% 100%  
101.40 101.41 101.50  
> mu = mean(X$Conductivity)  
> median(X$Conductivity)  
[1] 101.3  
> mu  
[1] 101.32  
>
```

The Environment pane shows:

Data	
x	20 obs. of 1 variable
Values	
mu	101.32
x	num [1:20] 101 101 101 101 101 ...
y	num [1:20] 0.0499 0.0998 0.1498 0.19...

The console also displays a boxplot for the variable 'Conductivity'.

So, let us now calculate the summary reports from property based reports. So, the mean is let us call it mu and the median, so you can get mean to be 101.32 and median is 101.3.

(Refer Slide Time: 25:44)

The top screenshot shows the following R code in a script editor:

```
mu = mean(X$Conductivity)
median(X$Conductivity)

## [1] 101.3

sigma = sd(X$Conductivity)
CI1 = mu-2*sigma
CI2 = mu+2*sigma
var(X$Conductivity)

## [1] 0.01010526

plot(X$Conductivity,xlab="Observation",
      ylab="Conductivity",main="Cu:conductivity")
```

The bottom screenshot shows the R console output for the same code:

```
> quantile(X$Conductivity)
 0%  25% 50% 75% 100%
101.1 101.3 101.3 101.4 101.5
> quantile(X$Conductivity,probs=seq(0,1,0.10))
 0%  10%  20%  30%  40%  50%  60%  70%
101.10 101.20 101.28 101.30 101.30 101.30 101.30 101.40
 80%  90% 100%
101.40 101.41 101.50
> mu = mean(X$Conductivity)
> median(X$Conductivity)
[1] 101.3
> mu
[1] 101.32
> sigma = sd(X$Conductivity)
> CI1 = mu-2*sigma
> CI2 = mu+2*sigma
> var(X$Conductivity)
[1] 0.01010526
> sigma
[1] 0.1005249
```

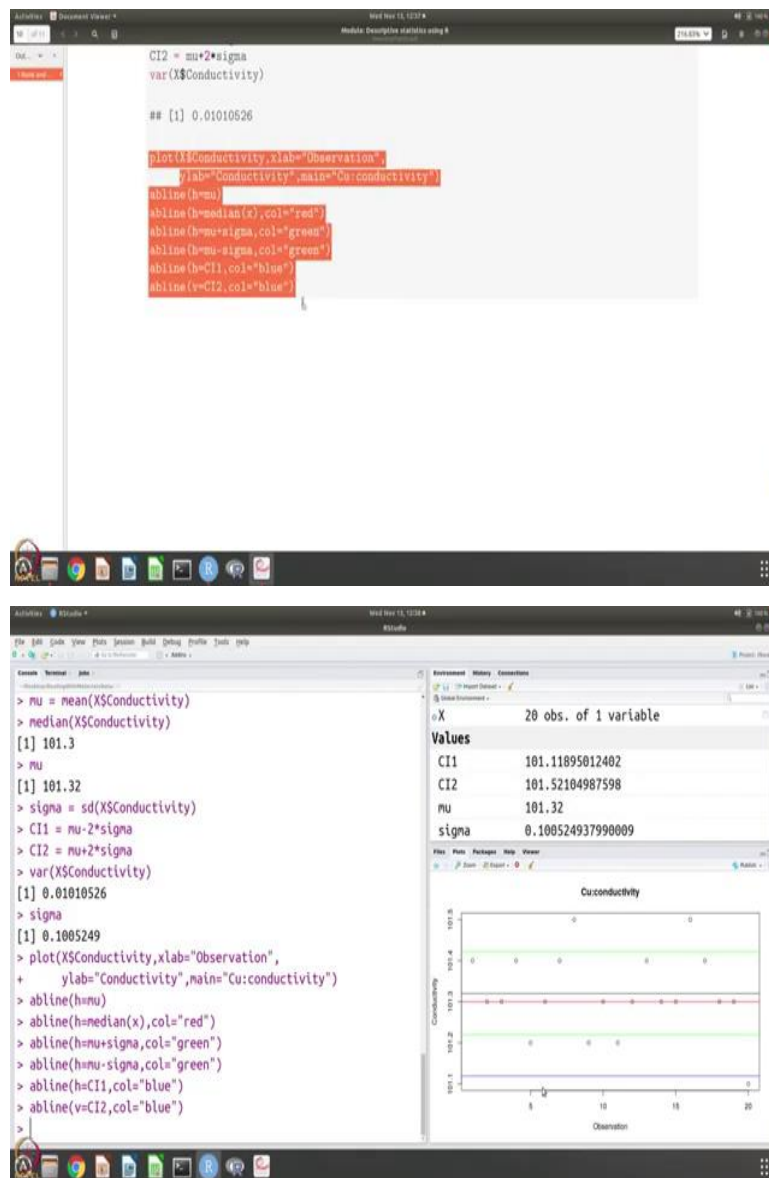
The summary window on the right shows the following values:

Variable	Value
X	20 obs. of 1 variable
CI1	101.11895012402
CI2	101.52104987598
mu	101.32
sigma	0.100524937990009

A boxplot is also displayed, showing the distribution of the 'Conductivity' variable with whiskers extending from approximately 101.1 to 101.5.

Now, you can get the sigma which is a standard deviation and variance which is var. Let us do that. So, sd is for standard deviation and we are going to store it as sigma, Mu is a mean sigma as a standard deviation and this is mu minus 2 sigma and mu plus 2 sigma is the 2 sigma range about Mu. And the var is basically variance which is. So, the variance is 0.1 and you can get sigma which is 0.1.

(Refer Slide Time: 26:28)



So, now let us do the final thing let us plot and let us also put this property based reports together with this. So, what we are going to do? We are going to plot the conductivity and we are going to start drawing lines at the mean, at the median and mu plus sigma, mu minus sigma, mu plus 2 sigma, mu minus 2 sigma etc.

You do that, you can see that okay so here is the data and the black line is basically the mean. The median is 101.3 so that is the red line and this is 1 sigma from the mean so this data points are lying within 1 sigma and if it is a normal distribution and we have been thinking that this is a normal distribution, you would expect that large percentage of data.

99 percent of data should lie between 2 sigma about the mean so and you can see that it falls but there is one data point that lies just outside of the 2 sigma. So, this is the 2 sigma line, mu

minus 2 sigma. That is why this is an outlier and this has been indicated in the dot chart and in the other plots also. Not so much in histogram plot but in the box plot we did see that this is an outlier.

So, and here also we see why it is an outlier. So, this is an another way of looking at the data, in this we have both put the data as well as the property based reports and things like histograms and cumulative distribution etcetera will be called as rank based representations.

So, we have taken a simple data set on conductivity and we have prepared both rank based and property based reports and we have learned how to present them using R. So, we will continue with some more data which can be little bit more complicated than this in the sessions to come. Thank you.