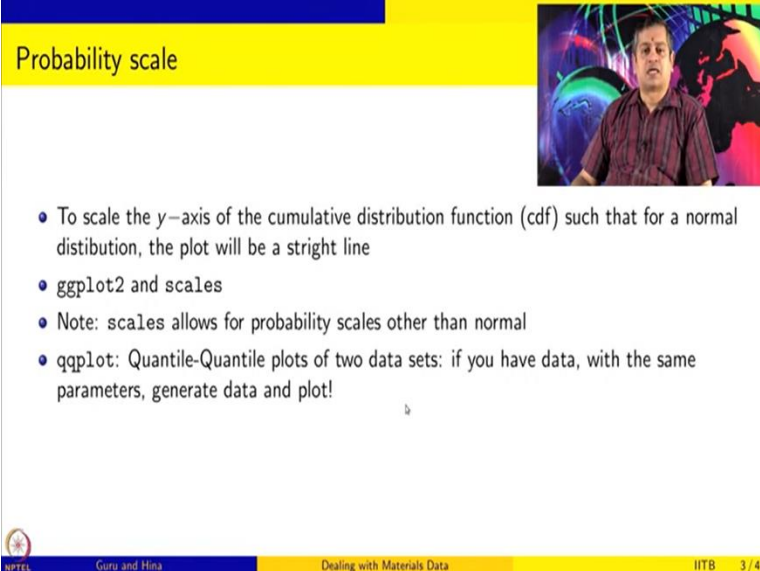


Dealing with Materials: Data Collection, Analysis and Interpretation
Professor M.P. Gururajan,
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 46
Probability scale

Welcome to Dealing with Materials Data, we are looking at the Collection Analysis and Interpretation of data from material science and engineering. We are in the module on Probability Distributions and specifically we are looking at normal probability distribution and in this session we are going to talk about probability scale and other related things to check whether a given data follows a given probability distribution. It starts with normal, so we will discuss normal first which is very, very common and I will also show you how to use R to do similar analysis for other distributions.

(Refer Slide Time: 00:57)



The slide is titled "Probability scale" and features a yellow header bar. On the right side, there is a small video inset showing a man speaking. The main content area contains a list of bullet points:

- To scale the y-axis of the cumulative distribution function (cdf) such that for a normal distribution, the plot will be a straight line
- ggplot2 and scales
- Note: scales allows for probability scales other than normal
- qqplot: Quantile-Quantile plots of two data sets: if you have data, with the same parameters, generate data and plot!

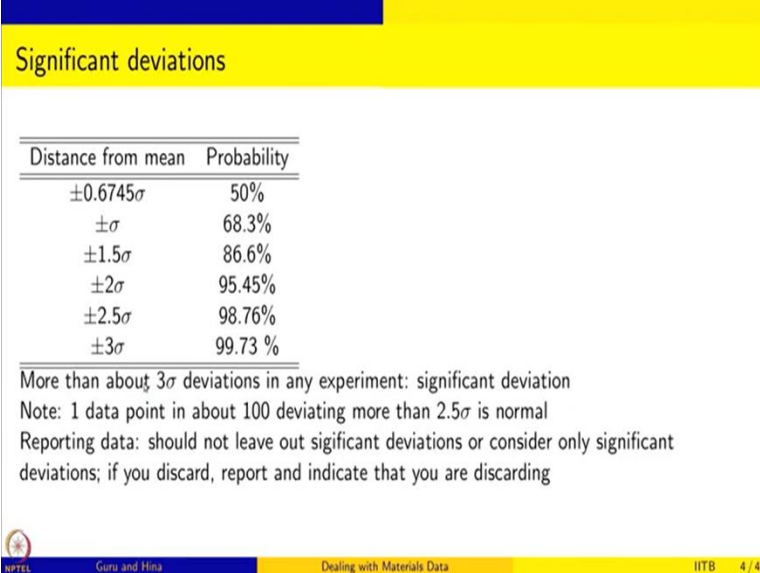
At the bottom of the slide, there is a footer bar with the following text: "NPTEL", "Guru and Hina", "Dealing with Materials Data", and "IITB 3/4".

So, the probability scale is a scale in which y-axis of the cumulative distribution function is scale such that if the data actually follows normal distribution the plot will be a straight line. We can use ggplot2 and scales to do this transformation of the scale to probability scale. We have done it once in the past but just for completion sake I am going to repeat that part of the exercise and most important point note, at that point we actually did not do this exercise but scales actually allows for probability scales other than normal.

So, I will show you an example and there are also plots which are known as qqplots, Quantile-Quantile plots from two data sets. This is also something that we have seen when we were using the library fit DISTR plus to fit the given data to known distributions. We did plot this qqplots and in that case there were two data sets one from the fitting and the other one was from the data itself imperical data.

And so the qqplots basically, you put them together the two data sets with the same parameters if they are same then you will again get a straight line and if they are different then you will see that they do not follow the straight line relationship. So that is what we are going to look in this session.

(Refer Slide Time: 02:35)



Significant deviations

| Distance from mean | Probability |
|--------------------|-------------|
| $\pm 0.6745\sigma$ | 50% |
| $\pm\sigma$ | 68.3% |
| $\pm 1.5\sigma$ | 86.6% |
| $\pm 2\sigma$ | 95.45% |
| $\pm 2.5\sigma$ | 98.76% |
| $\pm 3\sigma$ | 99.73 % |

More than about 3σ deviations in any experiment: significant deviation
Note: 1 data point in about 100 deviating more than 2.5σ is normal
Reporting data: should not leave out significant deviations or consider only significant deviations; if you discard, report and indicate that you are discarding

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

And before we proceed to use R to look at these numbers there is one more important thing which is known as significant deviation that we should know. If you are following normal distribution if your data follows normal distribution, so remember we said that if there are random noise in an experiment the data will follow normal distribution.

If so you can calculate what is the probability that a data point will lie about the mean within let us say 0.6745 sigma. So, 50 percent of the data point then you expect to fall within this distance. And, if you go for 1 sigma about 68 percent of the data will fall or any data point has 68.3 percent probability to fall between plus or minus sigma about the mean and, if you keep going 1.5 sigma is 86.6 and 2 sigma is 95.45, 2.5 sigma is 98.76, 3 sigma is 99.73.

In other words there is less than 1 percent chance for any data to fall outside of 3 sigma if you do an experiment. Remember if you do an experiment even if one data falls outside of 2.5 sigma that is understandable because if you have hundred data points for example then it is about 99 percent of the times it should fall. So, there is 1 percent probability that some data will fall outside this.

So that is expected so that is not uncommon. But if anything goes beyond 3 sigma in all probability that is a significant deviations, and while reporting data if you find data with very significant deviations you should still report the data. You might discard the data for some of the analysis that you will do, but it is still important to report and indicate that you are discarding and the reason for discarding namely that it is a significant deviation.

And in most of the cases that like we discussed in the past once significant deviations might also indicate something about the experiment or about the assumptions that you are making that might not be true or valid for that particular case or there might be other issues or phenomenon that is happening. So, it is important to pay attention to such data which has significant deviation, of course but it is also very important to report them.

It is wrong to consider data which have or data which is significantly deviating or to discard them without informing, both are wrong you cannot emphasize only on data which has significant deviation. You can also not just leave out data which has significant deviation, so it is important that you report and then indicate that you are discarding if at all you are discarding. So let us now go back and use R to look at these.

(Refer Slide Time: 05:49)

The image consists of two screenshots of an RStudio interface. The top screenshot shows a presentation slide with the following content:

Module: Descriptive statistics using R

M P Gururajan and Hina A Gokhale

Indian Institute of Technology Bombay, Mumbai

1 Probability scale

```
X <- read.csv("Data/ETPCuConductivity.csv")
library("ggplot2")
library("scales")
ggplot(data=X,aes(Conductivity)) +
  stat_ecdf() +
  scale_y_continuous(trans=scales::
    probability_trans("norm"))
```

Warning: Transformation introduced infinite values in continuous y-axis

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

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.

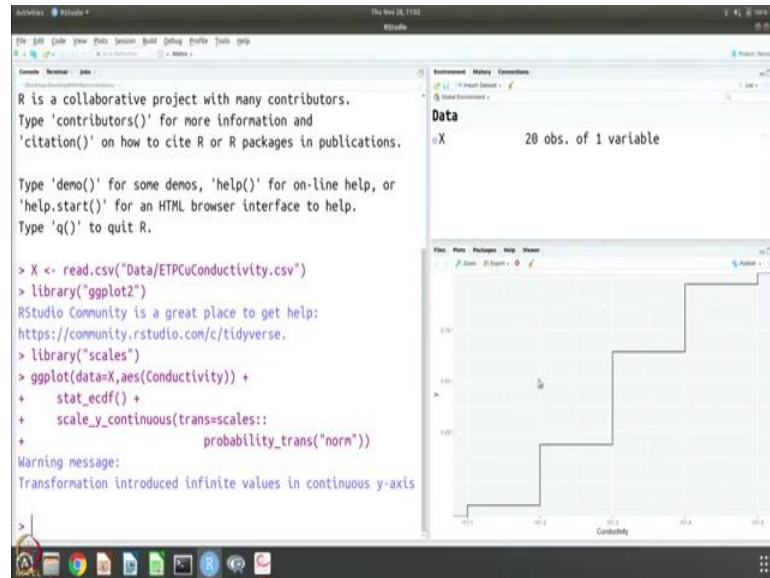
Natural language support but running in an English locale

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

Environment is empty

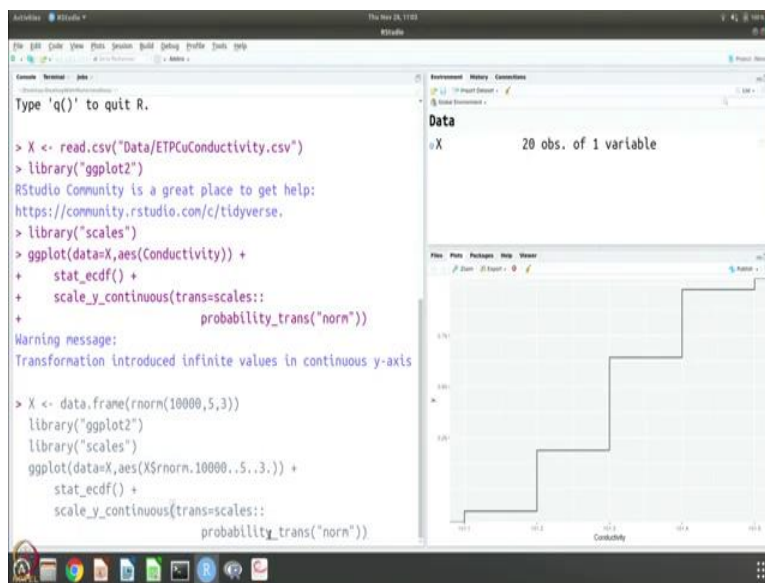
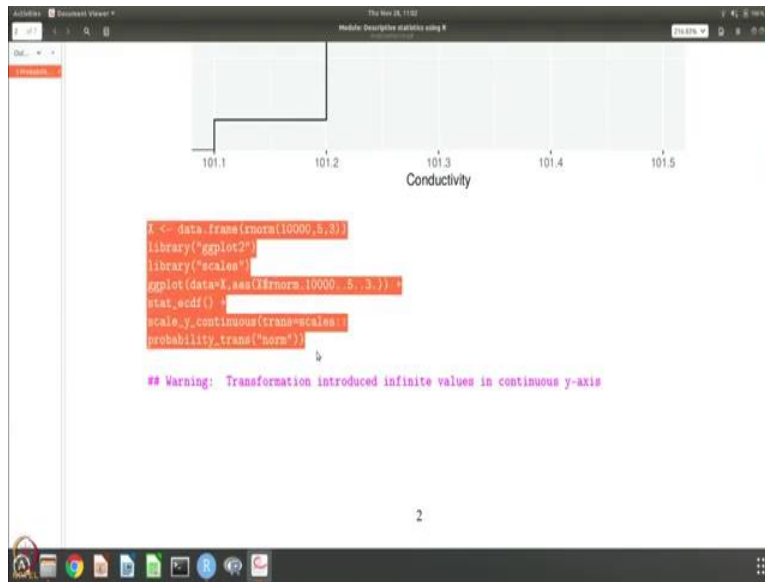


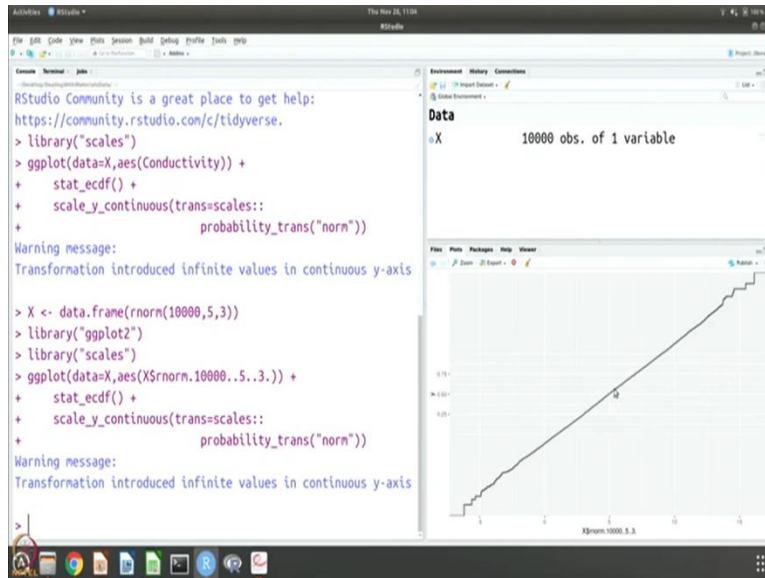
So, let us start with the probability scale problem. So, let us first start with the probability scale problem. This is something that like I said I am just repeating so this we have done in the past. So we read the ETP copper conductivity data. We invoke the library ggplot end scales and we first say that okay conductivity data has to be plotted and it is the empirical cumulative distribution function that we plot. Remember the probability scale is scaling the y axis in such a way that this ecdf plot will be a straight line and so scales actually allows for this y axis scaling.

And the probability transformation is for normal distributions. So, this is what allows us to put other probability transformations as we will shortly you can do other transformations and look at data. So, this we have done, so you can see that this is sort of a straight line. You can also find out where the mean is and what is the value of the standard deviation etcetera from such a plot, you will be able to directly read it out.

But this is very small number of data points and the straight line also is not looking quite like a straight line, here it is better. And so, so you can see that if you go this way so you will, you will find the, the data does follow more straight line.

(Refer Slide Time: 07:41)



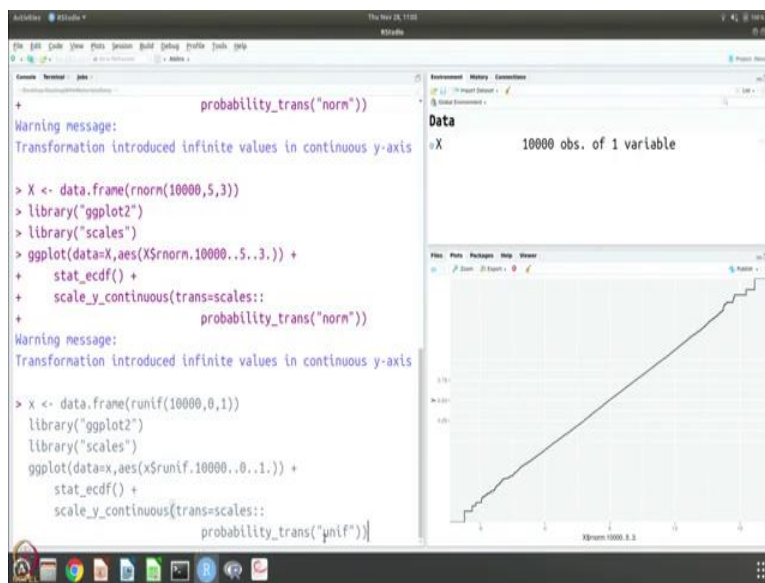
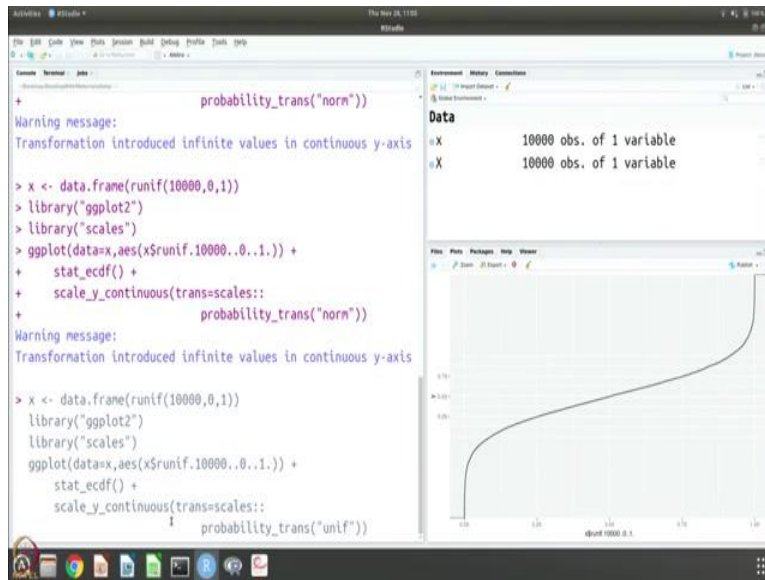


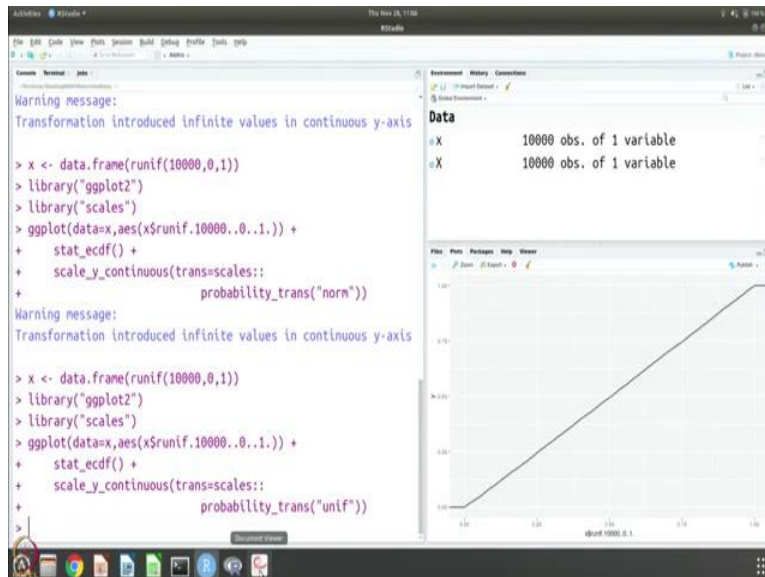
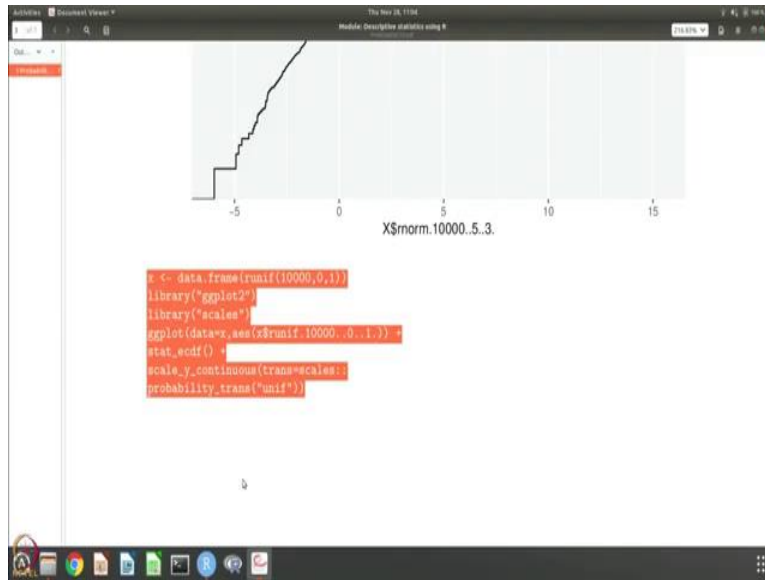
So let us try to do similar exercise but now from the data that we have generated from R itself. So, I want to generate some 10000 random normal deviates and the mean is 5 and standard deviation is 3 and I am going to do the same exercise plot the data and the cumulative distribution function of this data but I am going to change the probability scale to normal.

So when we do that, so you can see that this follows a straight line and you can see where the mean is, so that is at 5 and you can also see if where the standard deviation is, so that is 3, 5 plus or minus 3. So, you can see that it is here, 5 minus 3 which is like 2 because it is 0, 2.5, 5. So, 2 is somewhere here and similarly 8 is somewhere here its 7.58.

So, from the plot where the y axis is probability scale from the straight line you can also read of the mean and standard deviation fairly easily. So, that is one of the advantages of using this plots. It was difficult for us to do in the case of conductivity because the number of data points was very less, so it was overall a straight line behavior but you know the data was having lots of steps so we could not clearly see. But it is possible to do it. So here it is very clear that this is at 5 and this is at 2 and slightly here which is at 8, so you can find out the mean and standard deviation by looking at these plots.

(Refer Slide Time: 09:34)



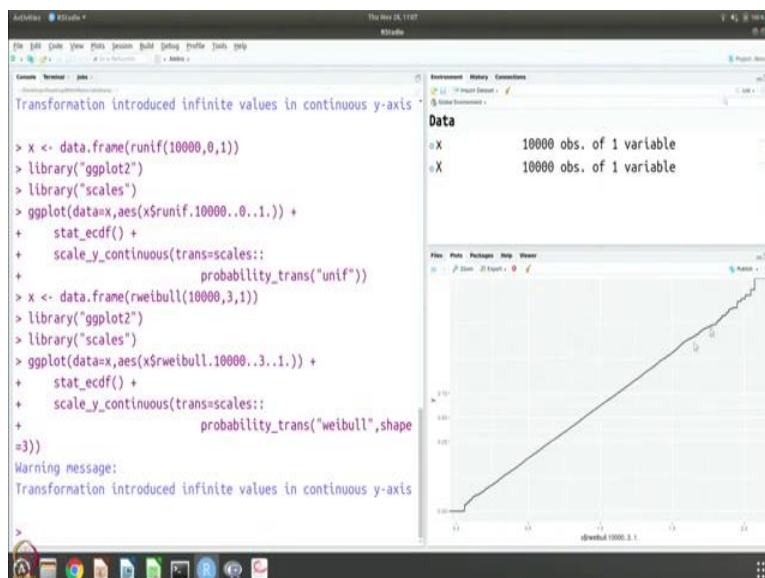
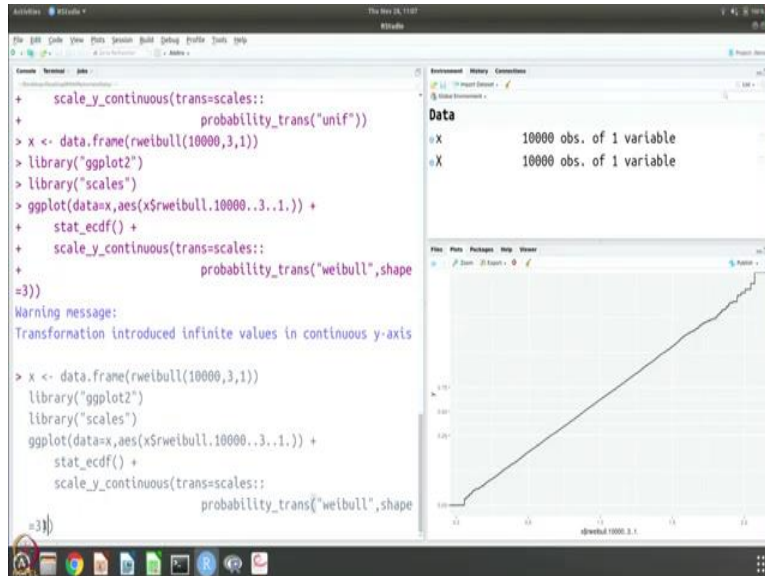


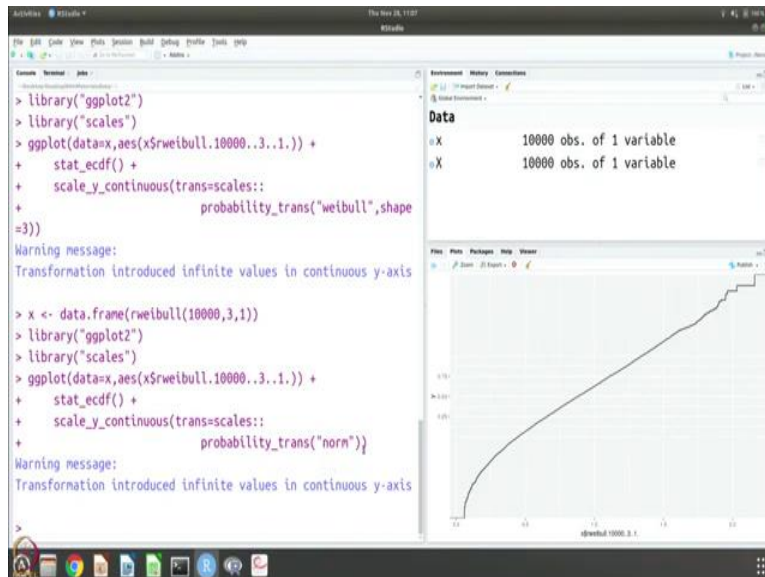
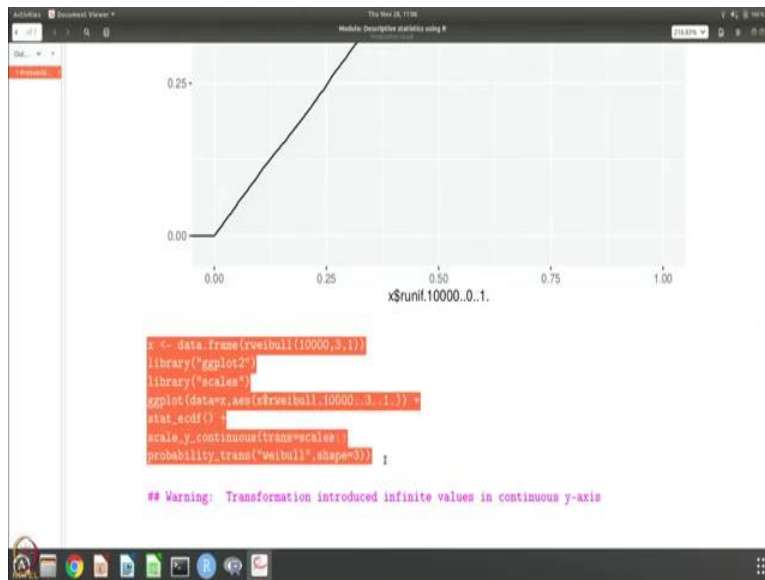
So, now let us do the next exercise so I want to show you that it is possible to use scales other than normal and get similar plots. So, in this case for example, I am using the uniform distribution and I am generating some 10000 random deviated from the uniform distribution and I am going to do the same exercise.

So, plot the cumulative distribution function scale the y axis, but I am going to scale it to be uniform. See, if I suppose scale it to be norm I will find data like this. So obviously it is not a straight line which is understandable because it is a normal scale that we have used here but whereas the data is actually generated from the uniform distribution. But if I make it uniform you will see that they data follows a straight line.

So, it is very clear that we can get information about other type of distribution, so you can look at the data and you can prepare such plots and see what happens. So, it was between 0 and 1, its uniform distribution and so you can see that the data follows a straight line when plotted by changing the y axis of the cumulative distribution function to follow the scale of uniform probability distribution.

(Refer Slide Time: 11:17)





So we can do similar thing so I am going to show you a slightly more complicated one which is for what is known as a weibull distribution. We will look at weibull distribution shortly. So, I am going to generate random deviates on weibull distribution 10000 of them and again I am going to plot the cumulative distribution function and the transformation scale I am going to make it as weibull and you will see that again it follows a straight line.

Of course we can do the other exercise instead of so remember in the case of transformation, probability transformation for weibull the shape of the, shape parameter of the weibull distribution is also important which we have used here and the same parameter should be used to get the proper

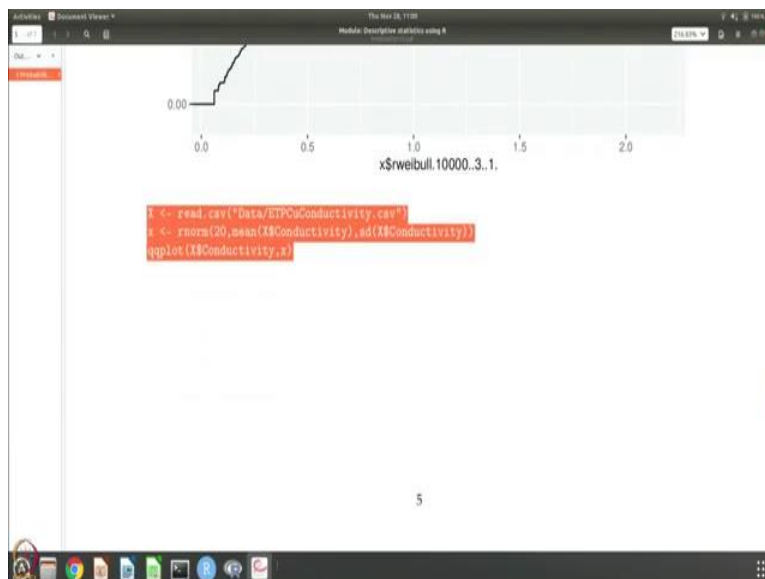
probability scale. So, that is why this extra parameter is needed to be given. Let us say that if I try to this also as norm so what happens.

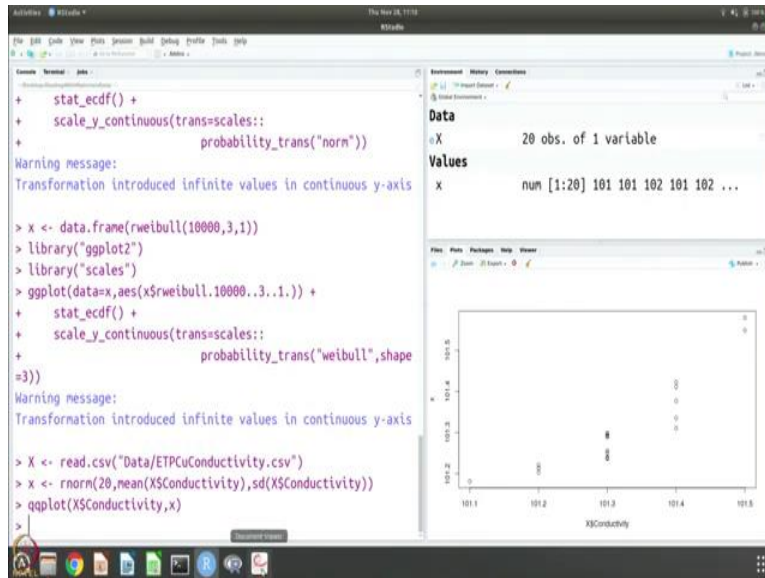
So, again you see that there is a deviation from the straight line behavior. So, any deviation you see from the straight line behavior basically indicates that it is not following. So if you had this data and if you tried to fit it for a normal distribution if you made this cumulative distribution function plot and made the y axis to be a probability scale for the normal distribution function, you will immediately see that the data is not following the normal distribution.

So, on the other hand if you make it weibull you will find that it is becoming a straight line. So, basically cumulative distribution plots with the appropriate scaling of the y axis should give you straight line or wherever it gives you straight line that scale basically tells you what is a distribution from which your data is derived? So, it is a very nice way and very useful way because many a times we have lots of data and we want to know what is the distribution that the data follows specifically most of the times we are interested in knowing whether the data follows normal distribution.

So, that is why generally probability scale commonly is the normal distribution, but of course with R now there is no such constraint. Previously there used to be graph sheets or sheets with such scaled variables for y axis on which you can plot, but with the computers now you can change these scales to anything you want and work with them. So, that is the advantage.

(Refer Slide Time: 14:02)

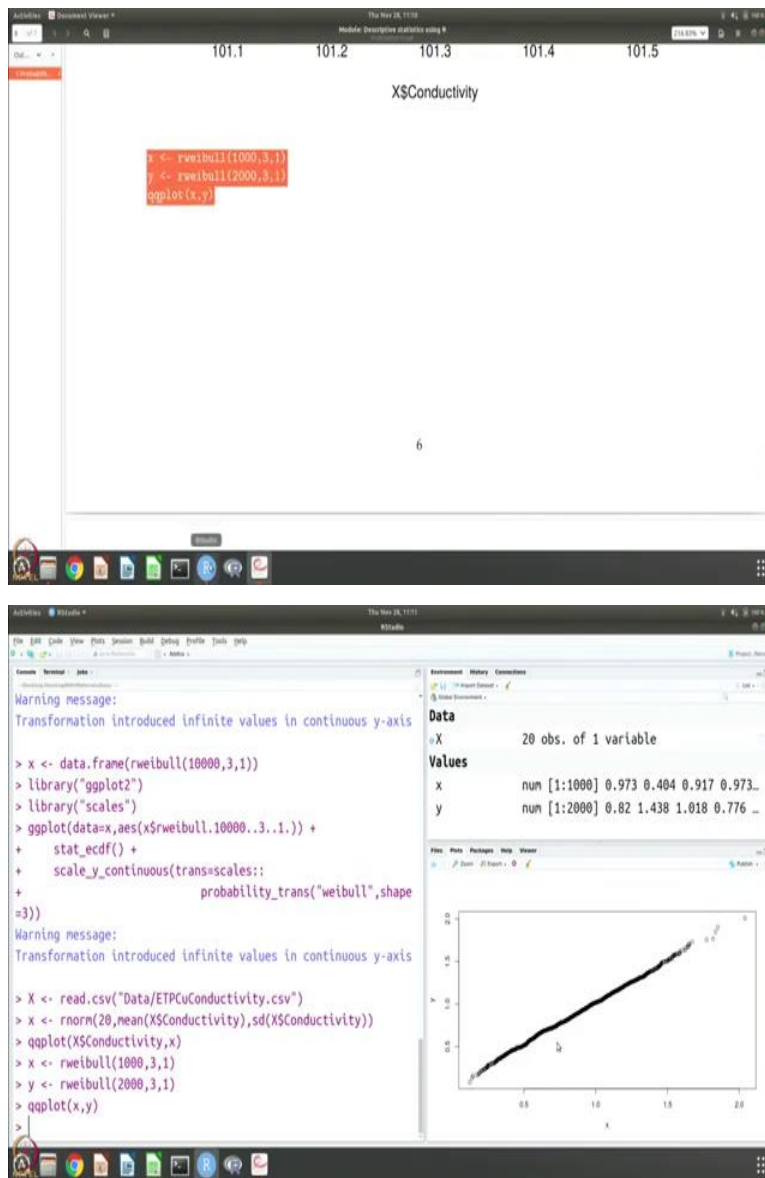




So, we will also look at an example of the qqplot which is a Quantile-Quantile plot and for doing that again I am going to go back to the electrical conductivity copper electrical conductivity example which we have seen. So, we are going to read the ETP conductivity data and I am also going to generate 20 random normal deviates with the same mean and standard deviation as my data and then I am going to make a qqplot of the data and the random normal deviates that I have made and you can see that they sort of follow this line.

So that basically tells you the from this qqplot that this is the data probably follows a normal distribution. Of course, it does not have to be just normal you can also make qqplots for other ones.

(Refer Slide Time: 15:04)



Let us do for weibull distribution and because we do not have data at the moment. Let us do the weibull distribution, let us generate 1000 and 2000 random deviates from the weibull distribution and let us generate a qqplot. So, you can see that the data actually follows a qqplot also follows this kind of behavior which tells you that the data.

So, if one of them was an empirical data the other one was a random deviate that you got from R if they followed such a straight line behavior then you know that this data is probably following weibull distribution for example. So, it is very useful to have these probability scale transformation

and qqplots and typically they were for normal distributions, because normal was considered as the most ubiquitous distribution.

So, everything else was a sort of bench mark against normal but with R, we have more freedom and we can use any distribution that we want and we can generate this plots which will give us better idea about the data and what kind of distribution probably our empirical data follows. So, we will stop here for normal and we will try to look at some more distributions and how to use R to deal with those distributions and what is a relevance for material science and engineering, thank you.