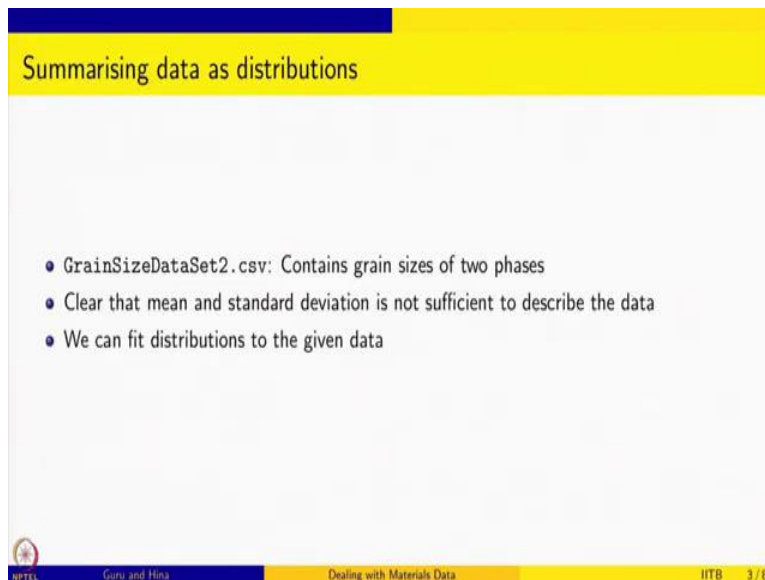


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor. M P Gururajan, Professor. Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Bombay
Lecture 28
Fitting experimental data to distributions

Welcome to dealing with materials data. In this course we are going to learn about Collection, Analysis and Interpretation of data. We are looking at the second module, which is on descriptive statistics using R and we have been looking at how to deal with distributions while presenting experimental results.

(Refer Slide Time: 00:40)



Summarising data as distributions


- GrainSizeDataSet2.csv: Contains grain sizes of two phases
- Clear that mean and standard deviation is not sufficient to describe the data
- We can fit distributions to the given data

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

Specifically, we have been looking at some grain size distributions. This is a case where the steel consists of two phases and the grain size data of both the phases is available in a csv file, and it is very clear that if you just look at the mean and standard deviation, it is not sufficient to describe the data. So, we need to find distributions that fit to the given data. So, that is what we want to try in this session. Of course, we have not done probability distributions, and we are going to do that as the next module. So, some of the ideas that we are going to use here, we are going to revisit

after we do the session on probability distributions. But at the moment, we will just use some existing libraries and use this data and look at the fit and identify what fits the given data better.

(Refer Slide Time: 01:40)



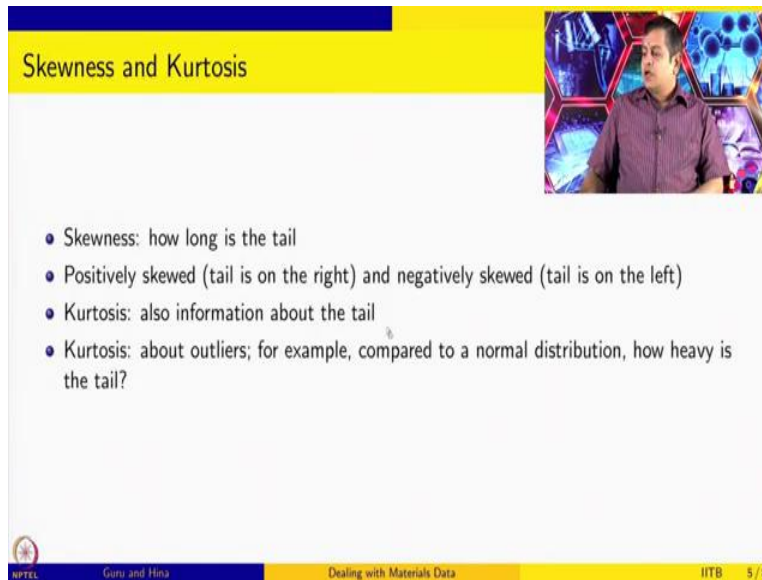
Using fitdistrplus

- We use the fitdistrplus library
- Identify the best distribution for fitting data
- Estimate the parameters of the distribution that fits data
- Very tutorial introduction: we will revisit and discuss some of the details after we go through the probability distribution module

MPTEL Guru and Hina Dealing with Materials Data IITB 4/8

So, we use fit distribution plus library, fitdistrplus. This is used to identify the best distribution for fitting the data and it also estimates the parameters of the distribution that fits the data. What we are going to give in this session is a very tutorial introduction. So, I am not going to explain many things, it is just like a command you give the command and you see the results and you will know what distribution it is, and you will give another command just to fit it for that distribution, but we will revisit and discuss some of the details after we go through the probability distribution module.

(Refer Slide Time: 02:21)



Skewness and Kurtosis

- Skewness: how long is the tail
- Positively skewed (tail is on the right) and negatively skewed (tail is on the left)
- Kurtosis: also information about the tail
- Kurtosis: about outliers; for example, compared to a normal distribution, how heavy is the tail?

IPTEL Gaur and Hina Dealing with Materials Data IITB 5/8

In order to understand how this distribution works, of course it is important to know about skewness and kurtosis and you might have already been taught about skewness and kurtosis. Skewness tells how long is the tail and it is said to be positively skewed, if the data is having a long tail on the right, and it is said to be negatively skewed, if the tail is on the left side of the data. On the other hand, for a normal distribution the data has tails on both right and left side and in fact if it is very nicely, normally distributed you will find that it is also symmetric about the mean.

On either side of the mean it will be having the same type of tail, but if the data is skewed, positively or negatively, then you will see that it has a longer tail either on the right or on the left. Kurtosis is also information about the tail and it specifically talks about the outliers and it gives you information as compared to a normal distribution how heavy is the tail in the given data, so that is what this gives. So, by looking at these two quantities, it is possible to find out what the best fit from in terms of probability distribution for the data could be and that is what we are going to do.

(Refer Slide Time: 03:48)

Moments about the mean



$$\mu_k = \sum_i (x_i - \mu)^k f(x_i) \quad (1)$$

μ_k : k -th moment about the mean

$f(x_i)$: probability distribution

μ : mean, the first moment about the origin

σ^2 : variance, second moment about the mean

skewness: third moment about the mean normalised by σ^3

kurtosis: fourth moment about the mean normalised by σ^4



$$\mu_k = \sum_i (x_i - \mu)^k f(x_i)$$

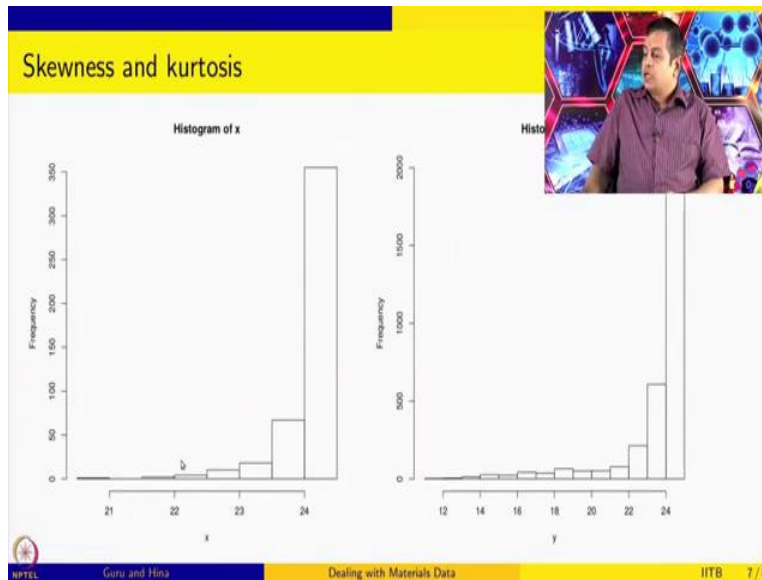
And to little bit better understand what these quantities are. We know about moments about the mean. So, μ_k is the k -th moment about the mean. It is defined as follows, so you take the data value and you take the mean which is the first moment about the origin and you take the difference to the power k and you multiply by the probability distribution. This is what we are going to discuss in detail in the next session but for now, it is enough if you understand that $f(x_i)$, basically gives you the probability that the random variable will take the value x_i and we are assuming that from that distribution is where we are getting these values. And so, this is basically a probability.

And μ is the first mean, first moment about the origin that is the mean and σ^2 is variance. So, it is the second moment about the mean itself. So, $(x_i - \mu)^2 f(x_i)$ is basically σ^2 , so that is a variance. And skewness and kurtosis are basically third and fourth moments about the mean. So, if you put cube and power 4 here, you get skewness and kurtosis, but it is not just putting 3 and 4 here, you also divide the resultant quantity by either σ^3 or σ^4 . To get skewness you divide by σ^3 and you, to get kurtosis you divide by σ^4 . So, these two numbers that you generate.

So, third moment about the mean normalized by σ^3 , where σ is the standard deviation, We know it is the square root of variance. And, so this quantities, so skewness and kurtosis is what we are going to use, to understand what is the probability distribution that

describes our data properly. Specifically, we are going to look at the grain size data and understand how it is distributed.

(Refer Slide Time: 05:56)



Skewness and Kurtosis

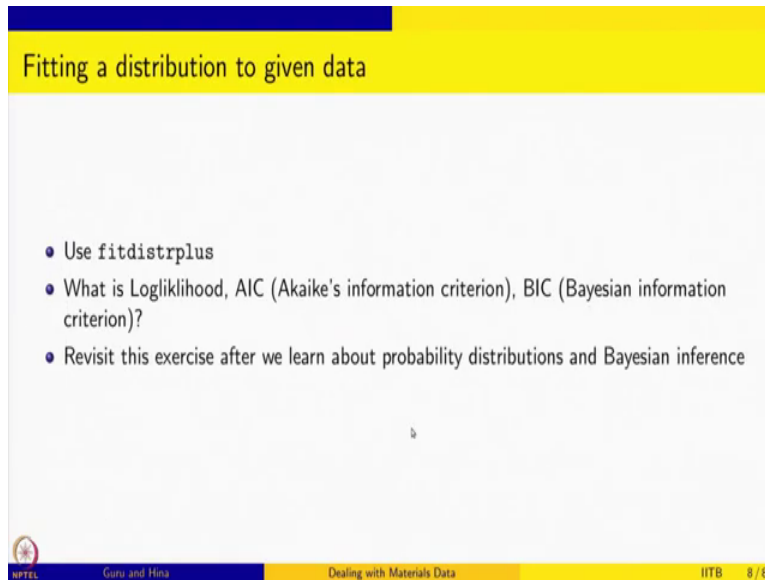
- Skewness: how long is the tail
- Positively skewed (tail is on the right) and negatively skewed (tail is on the left)
- Kurtosis: also information about the tail
- Kurtosis: about outliers; for example, compared to a normal distribution, how heavy is the tail?

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

So, you can see that our grain size data for phase 1 and phase 2, has huge skewness and also kurtosis. So, you can see that the distribution is of course one sided. So, it has a long tail to the left and this also has a long tail to the left. So, in our definition we will say that this is negatively skewed and also you can see the fatness or thickness of the tail. So, as compared to a normal distribution, of course, these tails are much more fatter. So, this is the information but we are going

to get numbers for these two quantities and they are defined in terms of moments about the mean and appropriately normalized.

(Refer Slide Time: 06:50)



The slide is titled "Fitting a distribution to given data" and contains the following bullet points:

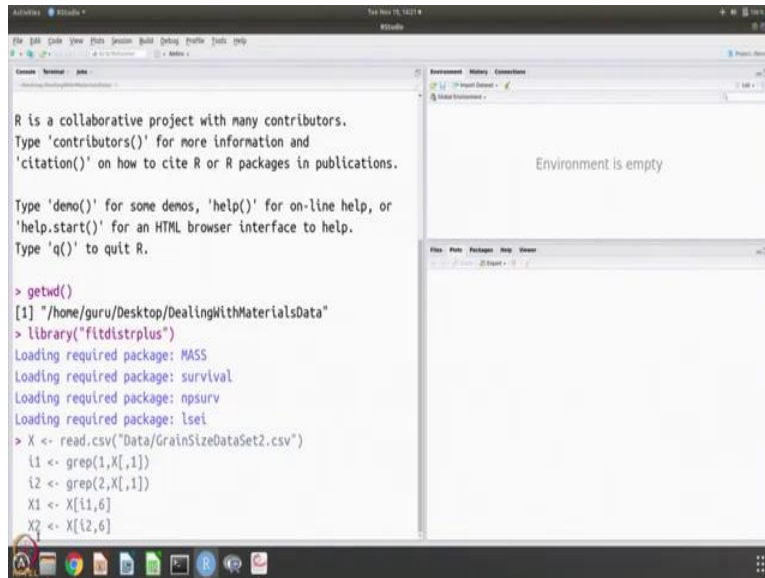
- Use `fitdistrplus`
- What is Loglikelihood, AIC (Akaike's information criterion), BIC (Bayesian information criterion)?
- Revisit this exercise after we learn about probability distributions and Bayesian inference

The slide footer includes the IITB logo, the text "Guru and Hina", "Dealing with Materials Data", and "IITB 8 / 8".

So that is what we are going to do, and for doing that we are going to use fit distribution plus library. And while we analyse the data and come up with a fit for the given data. We also have to evaluate how good is your fitting and for that there are measures. Specifically, you will see that R gives you information about loglikelihood, AIC which is Akaike's Information Criterion and BIC which is a Bayesian Information Criterion.

Of course, we will come back to understand these quantities better after we learn about probability distributions and influence and things like that, but for now, you just have to pay attention and see what are these quantities that R returns when you try to do the fitting. So, let us go and do the fitting, as usual.

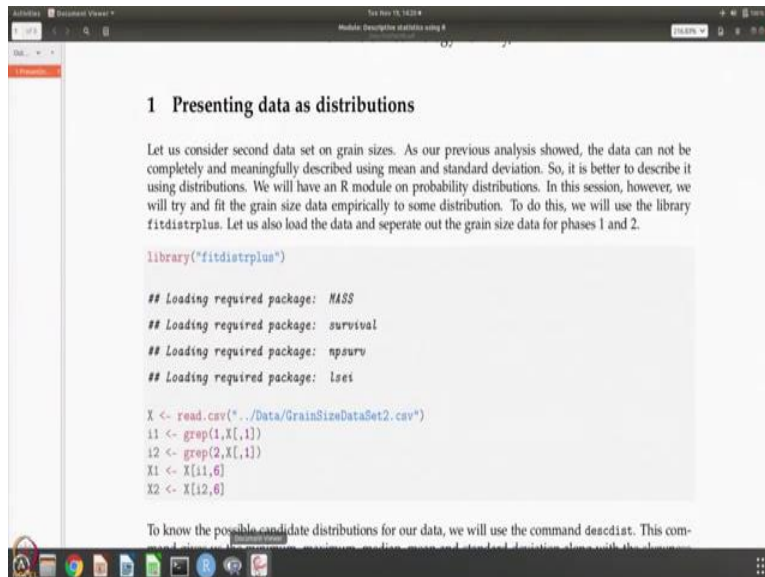
(Refer Slide Time: 07:55)



```
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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> library("fitdistrplus")
Loading required package: MASS
Loading required package: survival
Loading required package: npsurv
Loading required package: lsei
> X <- read.csv("Data/GrainSizeDataSet2.csv")
i1 <- grep(1,X[,1])
i2 <- grep(2,X[,1])
X1 <- X[i1,6]
X2 <- X[i2,6]
```



So, we will start with getting the data. So, let us start R and this is version 3.6.1 we need to know the working directory, and so we are in the right directory. So, we need to, first invoke the library and let us do that. So, we want to use the `fitdistrplus`. And then we want to read the data. So, the csv file from data, grain size data set 2.csv is read and then we are going to find out the phase identity 1 and 2 so we are going to save those row numbers in `i1` and `i2`. So, if you pull out from `X` all the `i1` ones that is for phase 1 let us call it `X1` and for `i2`, it is all for phase 2. Let us call it as `X2`.

(Refer Slide Time: 09:09)

```

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.

> getwd()
[1] "/home/guru/Desktop/DealingWithMaterialsData"
> library("fitdistrplus")
Loading required package: MASS
Loading required package: survival
Loading required package: npsurv
Loading required package: lsei
> X <- read.csv("Data/GrainSizeDataSet2.csv")
> i1 <- grep(1,X[,1])
> i2 <- grep(2,X[,1])
> X1 <- X[i1,6]
> X2 <- X[i2,6]
> descdist(X1)
1

```

```

> library("fitdistrplus")
Loading required package: MASS
Loading required package: survival
Loading required package: npsurv
Loading required package: lsei
> X <- read.csv("Data/GrainSizeDataSet2.csv")
> i1 <- grep(1,X[,1])
> i2 <- grep(2,X[,1])
> X1 <- X[i1,6]
> X2 <- X[i2,6]
> descdist(X1)
summary statistics
-----
min: 20.8 max: 24.3
median: 24.3
mean: 24.10547
estimated sd: 0.4345706
estimated skewness: -3.103136
estimated kurtosis: 15.99033
> help(descdist)

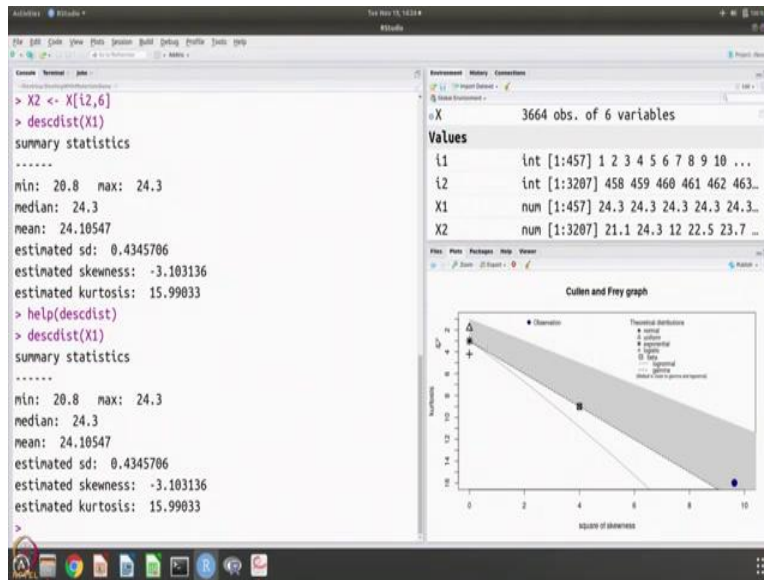
```

So, we have done and we have already seen this data. So, there are 3664 observations and there are 6 variables there and of which about 457 is for phase 1 and remaining 3200 or for phase 2. So, we have done this. So, now let us use this, command. So, it gives you the, so what is the descdist, so let us look up.

So, it is description of empirical distribution of non-censored data. There is this difference between censored data and non-censored data. Suppose for some reason to save time or because you are not able to continue the experiment for longer times. If you arbitrarily stop the experiment at some time or beyond some particular point, then that is called as censored data, what we have is not censored data. So, this is descdist is for Description of Distribution, that is what description of

distribution is what it is, and it is for the empirical data and the data should be non-censored and, in that case, you can use this.

(Refer Slide Time: 10:44)

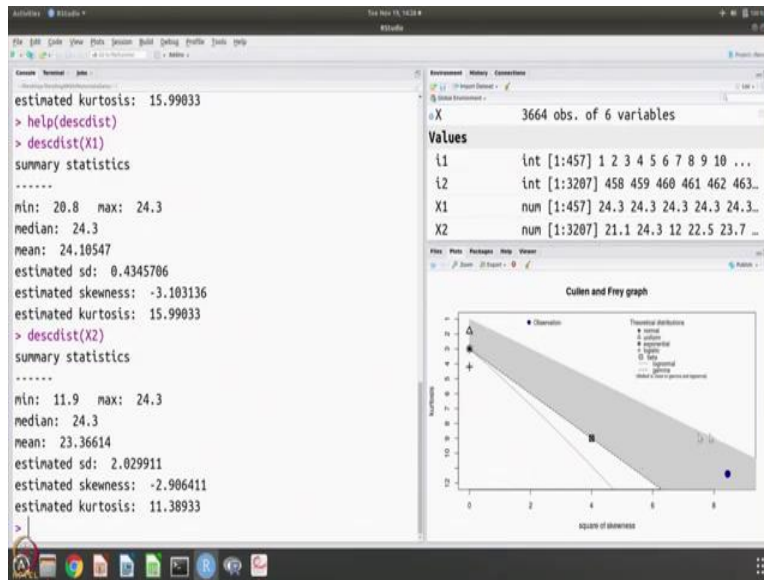


And if you look at the data that we have for phase 1 the grain size distribution data, then it lists several theoretical distributions. All of which we are going to learn in the next module normal distribution, uniform distribution, exponential, logistic, beta, lognormal, gamma and it also tells you that variable is close to gamma and lognormal.

So, if you have gamma and lognormal is the dotted point here, so the variable is close to these two distributions. And so, where is our observation? Our observation lies in this band which is for beta distribution. So, this graph it is called Cullen Frey graph and it is a graph of square of skewness versus kurtosis. And so, for example, for normal the kurtosis is here at 3 and the square of skewness is here, which is close to 0.

So, by taking these two values it knows that if some data falls somewhere here, then it must be normally distributed and so on and so forth. So, because our data falls somewhere here in the beta regime, we know that the data is, probably best described by the beta distribution and it gives you the, this we have already seen that minimum value is 20.8 and maximum 24.3 in this case. The median value is 24.3 because we saw lots of data points which were a 24.3 and the mean was 24.1 and the standard deviation was 0.4. So, it was 24.1 plus or minus 0.4 so it gives you the same data and it now in addition has estimated the skewness and the kurtosis.

(Refer Slide Time: 12:55)



So, let us do the same thing for the data for phase data for phase 2. Again, we see that this has a much larger range, minimum is 11.9, maximum is 24.3, median is again 24.3. So, you can see that the median is the same and the mean is also quite close. So, it is 23.4 and the standard deviation is about 2. So, it is 23.4 plus or minus 2, and this is 24.1 plus or minus 0.4.

So, these two data points in terms of mean and standard deviation, if you look at they are almost the same, but obviously the skewness is different or quite close it is not very different. So, it is minus 3.1 and this is minus 2.9. So, this is not very different. And kurtosis is again, this is about 16 and this is about 11.4. So, there is some difference. So, obviously it is not the same as phase 1, but it is also a beta distribution and it is slightly different from the previous one. So, we noticed that our data again falls in the beta distribution regime and but it is different from the earlier one.

(Refer Slide Time: 14:12)

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

```
X <- read.csv("../Data/GrainSizeDataSet2.csv")
i1 <- grep(1,X[,1])
i2 <- grep(2,X[,1])
X1 <- X[i1,6]
X2 <- X[i2,6]
```

Below the code, a text box explains the `descdist` command: "To know the possible candidate distributions for our data, we will use the command `descdist`. This command gives us the minimum, maximum, median, mean and standard deviation along with the skewness and kurtosis; the skewness-kurtosis plot helps us identify the distribution that describes our empirical data."

The code then runs `par(mfrow=c(1,2))` and `descdist(X1)`, producing the following summary statistics:

```
## summary statistics
## -----
## min: 20.8  max: 24.3
## median: 24.3
## mean: 24.10547
## estimated sd: 0.4345706
## estimated skewness: -3.103136
## estimated kurtosis: 15.99033
descdist(X2)
```

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

```
estimated kurtosis: 11.38933
> par(mfrow=c(1,2))
> descdist(X1)
summary statistics
-----
min: 20.8  max: 24.3
median: 24.3
mean: 24.10547
estimated sd: 0.4345706
estimated skewness: -3.103136
estimated kurtosis: 15.99033
> descdist(X2)
summary statistics
-----
min: 11.9  max: 24.3
median: 24.3
mean: 23.36614
estimated sd: 2.029911
estimated skewness: -2.906411
estimated kurtosis: 11.38933
```

On the right side of the RStudio interface, a data viewer shows the values for variables `i1`, `i2`, `X1`, and `X2`. Below the viewer, two Cullen and Frey graphs are displayed, comparing the observed data points to theoretical distributions (normal, uniform, exponential, beta, gamma, Weibull, and stable).

So, we can also try to get them both in the same figure by doing this. That will make life easy for us to compare. So, you can see that, so this goes to this point and, so these values are different. So, this is 12 and this is 16. So, this is somewhere about 16 and this is somewhere about 11 point something, and in terms of the square of skewness. So, this is somewhere about near 10 and this is less than 9 but, in both cases it falls in this band, which is called beta distribution. So, let us go back and try to fit the distribution. Now that we know that it is beta etc.

(Refer Slide Time: 15:19)

To fit the data to beta distribution, let us use the command `fitdist`. We also normalise the data to be between 0 and 1 to fit to beta distribution:

2

```
x <- X1/max(X1)
y <- X2/max(X2)
fit.b1 <- fitdist(x,"beta",keepdata=TRUE)

## [simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data, gr = gradient, dd

## Error in fitdist(x, "beta", keepdata = TRUE): the function mle failed to estimate the parameters,
## with the error code 100

fit.b2 <- fitdist(y,"beta",keepdata=TRUE)
```

```
summary statistics
-----
min: 20.8 max: 24.3
median: 24.3
mean: 24.10547
estimated sd: 0.4345706
estimated skewness: -3.103136
estimated kurtosis: 15.99033
> descdist(X2)
summary statistics
-----
min: 11.9 max: 24.3
median: 24.3
mean: 23.36614
estimated sd: 2.029911
estimated skewness: -2.906411
estimated kurtosis: 11.38933
> par(mfrow=c(1,1))
> x <- X1/max(X1)
> y <- X2/max(X2)
>
```

Variable	Class	Values
i1	int	[1:457] 1 2 3 4 5 6 7 8 9 10 ...
i2	int	[1:3207] 458 459 460 461 462 463...
X	num	[1:457] 1 1 1 1 1 ...
X1	num	[1:457] 24.3 24.3 24.3 24.3 24.3...
X2	num	[1:3207] 21.1 24.3 12 22.5 23.7 ...
y	num	[1:3207] 0.868 1 0.494 0.926 0.9...

To fit the data to beta distribution, let us use the command `fitdist`. We also normalise the data to be between 0 and 1 to fit to beta distribution:

2

```

x <- X1/max(X1)
y <- X2/asy(X2)
fit.b1 <- fitdist(x,"beta",keepdata=TRUE)

## <simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.arg, obs = data, gr = gradient, dd
## Error in fitdist(x, "beta", keepdata = TRUE): the function mle failed to estimate the parameters,
## with the error code 100

fit.b2 <- fitdist(y,"beta",keepdata=TRUE)

```

Environment History Connections

i1	int	[1:457]	1 2 3 4 5 6 7 8 9 10 ...
i2	int	[1:3207]	458 459 460 461 462 463...
x	num	[1:457]	1 1 1 1 1 ...
X1	num	[1:457]	24.3 24.3 24.3 24.3 24.3...
X2	num	[1:3207]	21.1 24.3 12 22.5 23.7 ...
y	num	[1:3207]	0.868 1 0.494 0.926 0.9...

```

> descdist(X2)
summary statistics
-----
min: 11.9 max: 24.3
median: 24.3
mean: 23.36614
estimated sd: 2.029911
estimated skewness: -2.906411
estimated kurtosis: 11.38933
> par(mfrow=c(1,1))
> x <- X1/max(X1)
> y <- X2/max(X2)
> fit.b1 <- fitdist(x,"beta",keepdata=TRUE)
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.
arg, obs = data, gr = gradient, ddistname = ddistname, hess
ian = TRUE, method = meth, lower = lower, upper = upper,
...): function cannot be evaluated at initial parameters>
Error in fitdist(x, "beta", keepdata = TRUE) :
the function mle failed to estimate the parameters,
with the error code 100

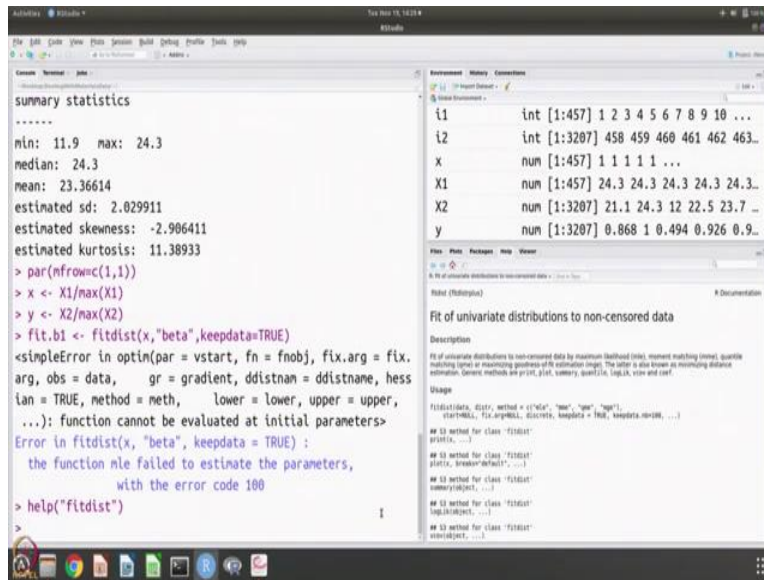
```

Cullen and Frey graph

Cullen and Frey graph

So, can be fit, for fitting to fit 2 beta you will learn that the value has to be between 0 and 1, so that is what we are going to do, you are going to normalize the values to be between 0 and 1. So, x is nothing but the X1 values divided by the maximum and why is nothing but x2 divided by its maximum. So, we have to normalize values and let us use this, let us try to fit it to beta. So, we say that fit to the distribution take the data x and fit to the distribution and fit it to beta and we are going to use the data and the fit will be saved as fit.b1. If we try, then we get the information that function MLE failed. What is MLE, MLE is Maximum Likelihood Estimation and if it failed then we can try to use other methods to fit. Let us use this MME, Moment Matching Estimation.

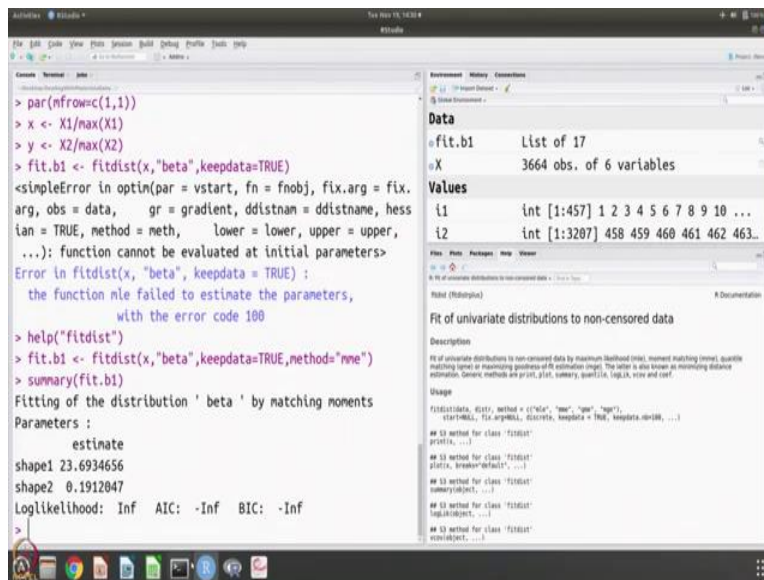
(Refer Slide Time: 16:26)



```
summary statistics
-----
min: 11.9  max: 24.3
median: 24.3
mean: 23.36614
estimated sd: 2.029911
estimated skewness: -2.906411
estimated kurtosis: 11.38933
> par(mfrow=c(1,1))
> x <- X1/max(X1)
> y <- X2/max(X2)
> fit.b1 <- fitdist(x,"beta",keepdata=TRUE)
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.
arg, obs = data, gr = gradient, ddistname = ddistname, hess
ian = TRUE, method = meth, lower = lower, upper = upper,
...): function cannot be evaluated at initial parameters>
Error in fitdist(x, "beta", keepdata = TRUE) :
the function nle failed to estimate the parameters,
with the error code 100
> help("fitdist")
```

How do we know these methods, of course you can use help, fitdist for example, you will get this information. So, it says fit of univariate distributions to non-censored data, by maximum likelihood or moment matching or quantile matching or Maximizing Goodness of fit Estimation, MGE. So, let us try the MME. So, for that you have to say method equal to this. So that fit works and you can get information about that fit.

(Refer Slide Time: 17:08)

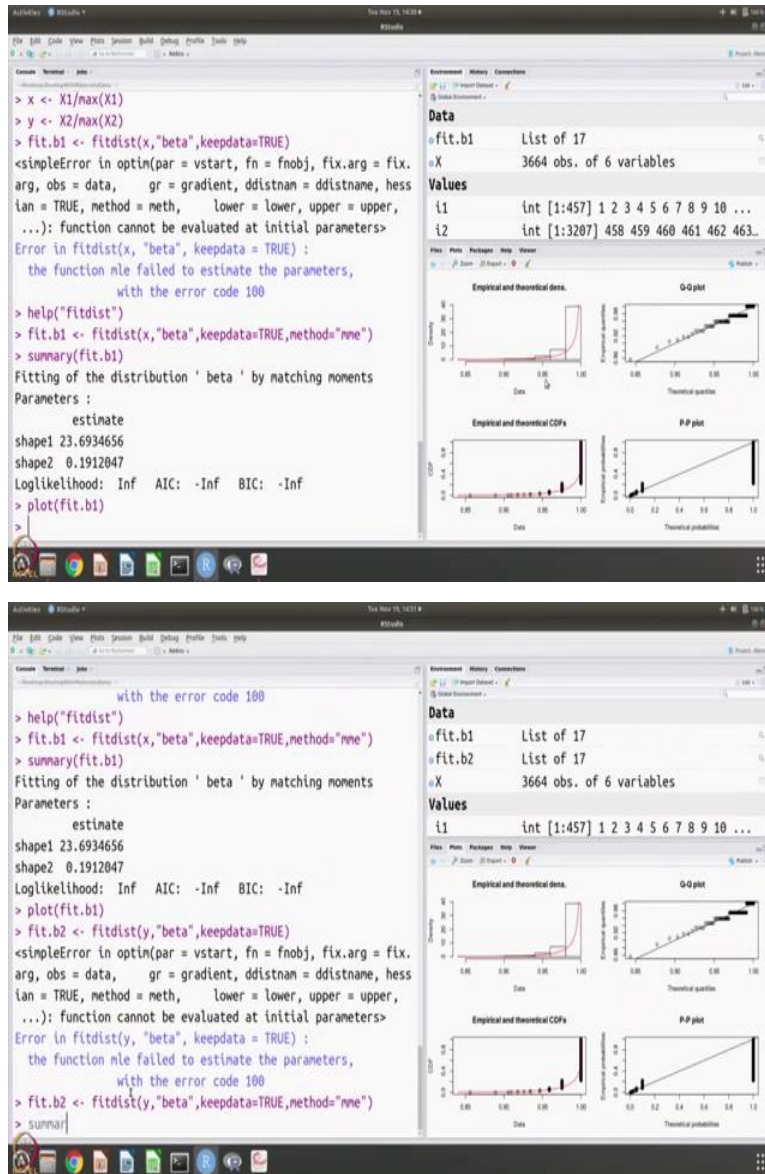


```
> par(mfrow=c(1,1))
> x <- X1/max(X1)
> y <- X2/max(X2)
> fit.b1 <- fitdist(x,"beta",keepdata=TRUE)
<simpleError in optim(par = vstart, fn = fnobj, fix.arg = fix.
arg, obs = data, gr = gradient, ddistname = ddistname, hess
ian = TRUE, method = meth, lower = lower, upper = upper,
...): function cannot be evaluated at initial parameters>
Error in fitdist(x, "beta", keepdata = TRUE) :
the function nle failed to estimate the parameters,
with the error code 100
> help("fitdist")
> fit.b1 <- fitdist(x,"beta",keepdata=TRUE,method="mme")
> summary(fit.b1)
Fitting of the distribution ' beta ' by matching moments
Parameters :
      estimate
shape1 23.6934656
shape2  0.1912047
Loglikelihood: Inf  AIC:  -Inf  BIC:  -Inf
>
```

So, you can see that fitting of the distribution beta by matching moments, and these are the parameters and it gives you this loglikelihood AIC, BIC all to be infinity. So, this is what I said, we

want to understand what these quantities are, but we will come back to it after we do some more modules and when we learn about inferences and things like that, we will come back and take a look at it. So, you can do the fitting and then of course you can plot

(Refer Slide Time: 17:48)

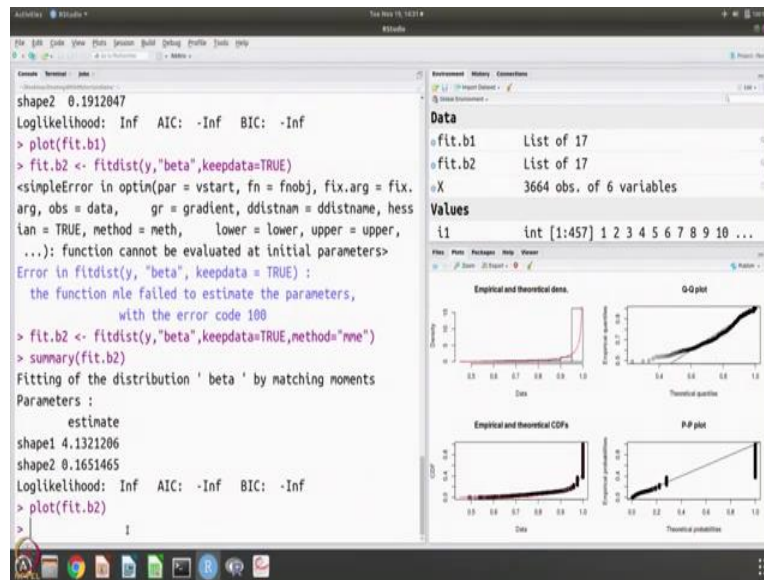


So, you can see that the data and the density plot is here, and the CDF plot is here. So, these are the data and the red line that runs through is basically our fit and you can see the Q-Q plot. And so that also seems to fit well and this is the P-P plot.

So, what are these Q-Q plots and P-P plots, we will learn when we look at the distributions and learn about these quantities, but for now, this seems to fit well and so we can do the same exercise

for the second data Y beta, so we can again see that MLE has failed. So, we will again use method to be MME and see if that works. Obviously, that works. So, you can look at what this fit is. So again, you get to these loglikelihood parameters and AIC, BIC parameters to be infinity. We will come back and understand what it is, but for now we can try to plot and see.

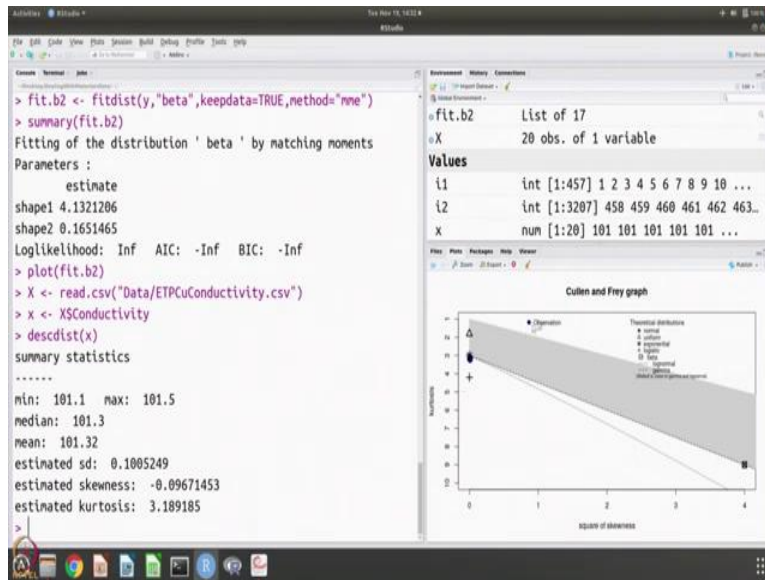
(Refer Slide Time: 19:02)



So again, you see that the empirical and theoretical densities, the empirical and theoretical cumulative distribution functions and the Q-Q plot and P-P plot they are all okay and as compared to the previous case the Q-Q plot is slightly off, but it is still okay. It is fitting most of the data.

And so that is what we are realizing. Now that we have done this exercise, we have been looking at also the electrical conductivity data of ETP copper, and we noticed that, that data was fitting or looking like normal distribution. Is it so, can we check if it is indeed the normal distribution? So, for doing that let us do this. So, we are going to read the data that is ETP copper conductivity data and we are going to say describe that distribution, that data.

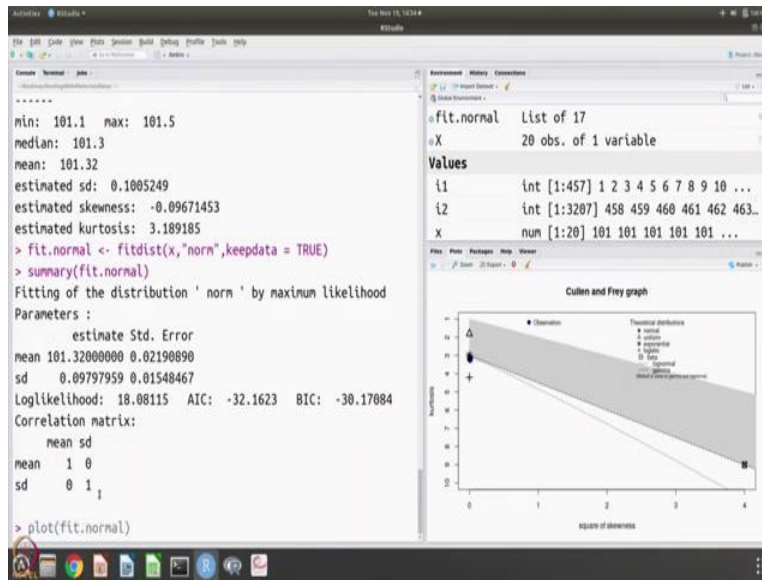
(Refer Slide Time: 20:09)



And we find that our observation of course lies along this star, which is normal distribution. So, this is what we have been noticing and that this is minimum is 101.1, maximum is 101.5, median is 101.3 and mean is 101.32 and standard deviation was 0.1. So, these we have already seen and the skewness, you can see is quite close to 0 and kurtosis is quite close to 3.

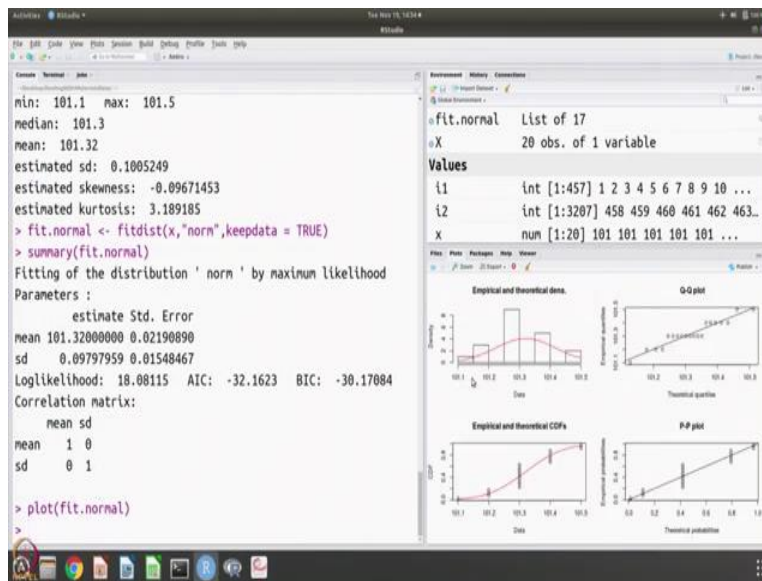
So, this shows you that this is a very nice, normal distribution. Of course, we can check that indeed is so. How do we do that? So we try to fit this to normal distribution and here is the. So, we say that okay fit the distribution, take the X data and fit it to normal distribution and give the summary of the fit.

(Refer Slide Time: 21:10)



So, we again find that of course it fits and it used the maximum likelihood method and this is the mean and the standard error and the standard deviation. So, it is like 0.1 and this time you can see that the loglikelihood the AIC, BIB etc. are not infinites. So, it is giving you some numbers and it also gives you what is known as correlation matrix. So, we will at some point look at what it is, of course, let us plot the normal fit we have made.

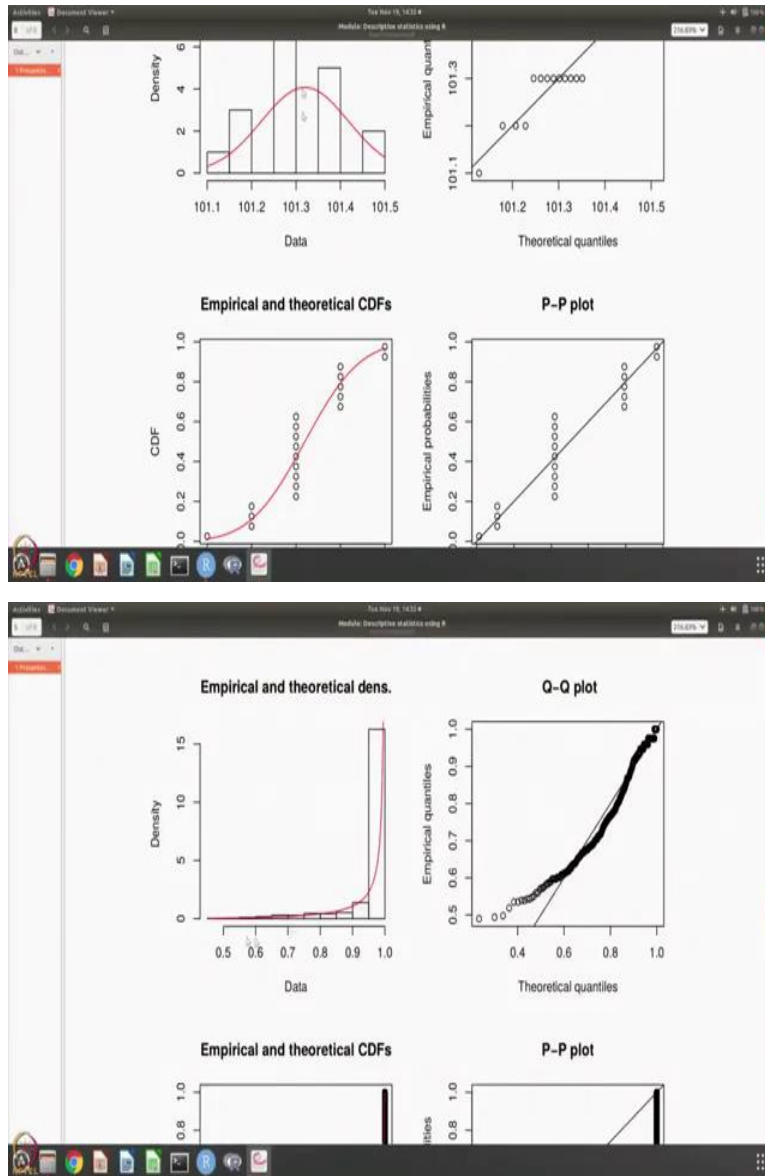
(Refer Slide Time: 21:55)



So, you can see that the experimental and the empirical and theoretical densities match and the cumulative distribution functions match and the Q-Q plot is a nice line, as also the P-P plot. So,

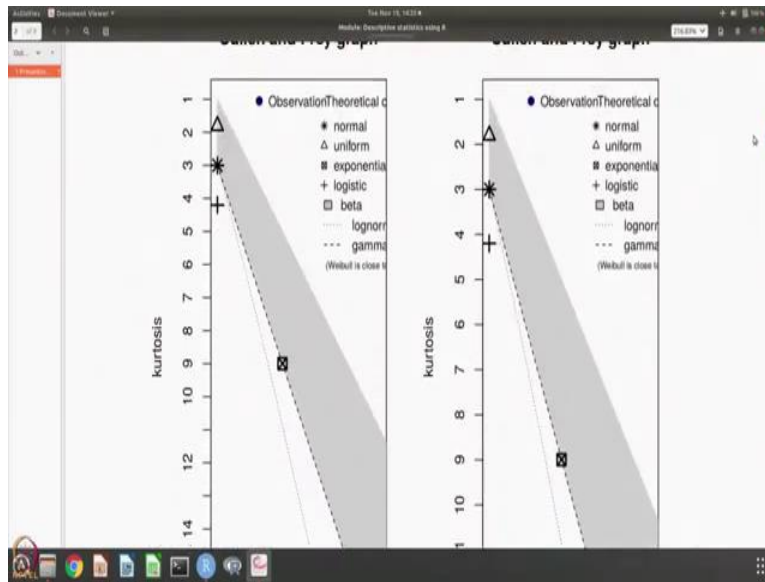
we can see that in this case everything is nicely following the normal distribution. So, to summarize so we have been looking at data and sometimes we find that the data can be better described by distributions.

(Refer Slide Time: 22:33)



For example, in the case of conductivity, this is repeated measurements, which give you values about some mean and the distribution is there because of random noise and that is why it is a normal distribution, but on the other hand every single measurement gives you a set of distribution for grain sizes and this obviously is not a normal distribution or a bell shaped curve. So, to describe this kind of distributions, you can use this library fitdistrplus and you can get information.

(Refer Slide Time: 23:09)



And generally, the methodology here is that by looking at where the skewness and kurtosis values lie, we decide what could be the best theoretical distribution that will fit the given empirical data. So, that is the exercise that we have done. And we will come back to some aspects of this fitting exercise after we go through the probability distribution. Thank you.