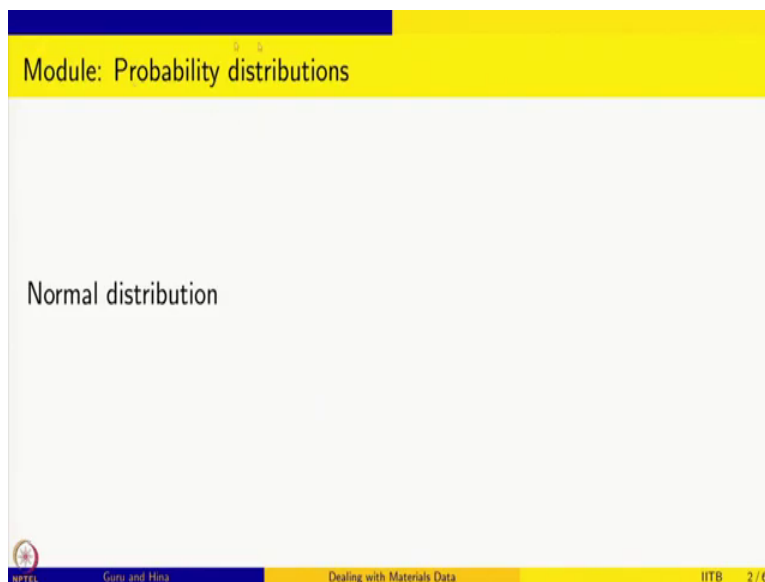


**Dealing with Materials Data: Collection, Analysis and Interpretation**  
**Professor M P Gururanjan**  
**Professor Hina A Gokhale**  
**Department of Metallurgical Engineering and Materials Science**  
**Indian Institute of Technology, Bombay**  
**Lecture 44**  
**Normal distribution**

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 third module, this is a module on probability distributions, and we have looked at several discrete probability distributions and we are now going to consider one of the continuous probability distributions which is called as the normal distribution.

(Refer Slide Time: 0:31)



As the name indicates, it is a very common one and it is also a very important one for several reasons. So, we are going to spend quite a bit of time understanding normal distribution. And so in this session, we are going to look at one aspect of normal distribution namely that many times if you make measurements; because of random errors or thermal noise, the value that you would measure repeatedly if you make the experiment will not be the same number.

The error will actually make it get distributed as a normal distribution. So, this is the context in which we are going to understand normal distribution. First, we will look at the other context and

also the importance of normal distribution and we have been using normal distribution in some cases earlier. So, we will even revisit some of those and try to understand what we did and why that makes sense in the context of understanding normal distributions.

(Refer Slide Time: 1:46)

The slide is titled "Data with random errors" and contains the following text:

- Consider any measurement  $X^*$
- Let us say that the mean of the measurements is  $\mu$
- Let us say that the standard deviation of the measurements is  $\sigma$
- Assume that the reason for deviations from the mean is random noise; note that this rules out data such as grain size distribution
- $X \sim N(\mu, \sigma)$
- $Z \sim N(0, 1)$
- $Z$ : standard normal distribution;  $z = \frac{x - \mu}{\sigma}$

At the bottom of the slide, there is a footer with the IITB logo, the text "Guru and Nina", "Dealing with Materials Data", and "IITB 3 / 6".

Let us consider any measurement that you are making. Let us say that  $X$  is the measurement and let us say that the mean value of the measurements after you repeated the experiment some number of times is  $\mu$  and let us say that the standard deviation of the measurement is  $\sigma$ . We are assuming that the reason for deviations from the mean is the random noise. So, this rules out data such as grain size distribution, because we saw that you can make a measurement and the measurement can give you a mean and a standard deviation.

But that might be because of some other distribution that is there in the system it is not because of random noise. So, it is not related to the normal distribution. It need not be related to normal distribution, we have also seen in the in the previous example, for example, that mixing 2 or convoluting 2 distributions can result in some other distribution we saw that normal plus hypergeometric actually gives you binominal and so on.

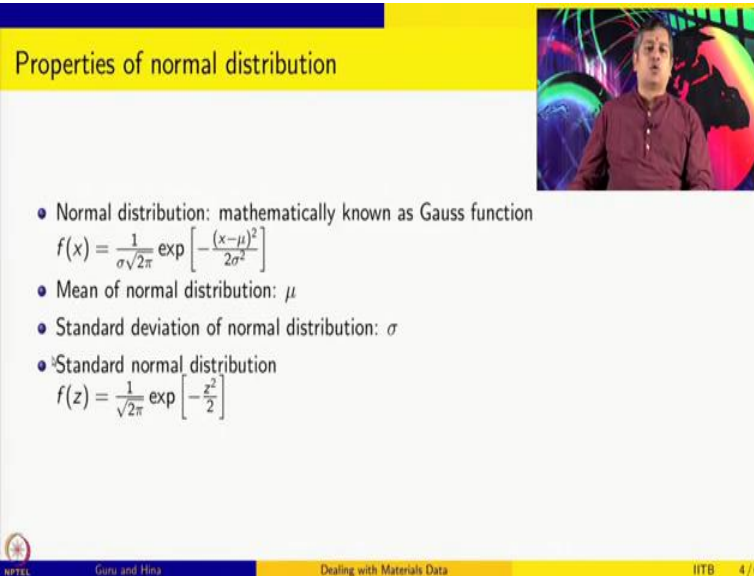
So, normal can also be a result of some other distribution that is not the case that we are looking at this moment. We are saying that if you have some random error or noise or thermal noise that

will lead to normal distribution in the measurements that you make and we say that  $x$  goes as normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

$$z = \frac{x - \mu}{\sigma}$$

And we also define what is known as standard normal distribution for which the mean becomes 0 and the standard deviation is one and you obtain from  $x$ , the  $z$  by doing the transformation that is you take all the measurements and subtract the mean and divide by the standard deviation. The result and variable will actually follow the standard normal distribution.

(Refer Slide Time: 3:47)



The slide is titled "Properties of normal distribution" and features a presenter in the top right corner. The content includes the following points:

- Normal distribution: mathematically known as Gauss function  
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$
- Mean of normal distribution:  $\mu$
- Standard deviation of normal distribution:  $\sigma$
- Standard normal distribution  
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$$


At the bottom of the slide, there is a footer with the following text: "NPTEL Guru and Hina Dealing with Materials Data IITB 4/6".

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$$

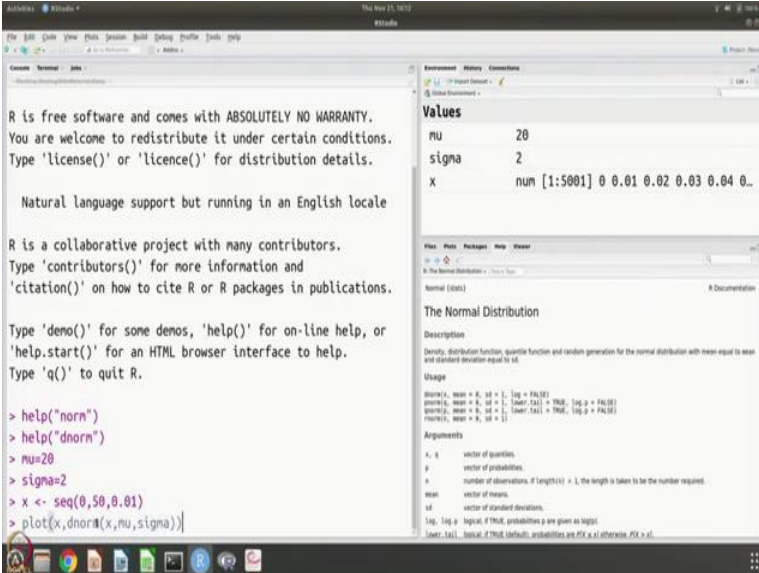
(Refer Slide Time: 4:38)

## Normal distribution using R



- `norm`
- `dnorm`, `pnorm`, `qnorm`, and `rnorm`
- Can you plot the probability density, cumulative distribution function and quantile function for the normal distribution?
- Can you generate 20 random variates from a normal distribution?
- Assume  $\mu = 20$ ;  $\sigma = 2$

NIPTEL    Guru and Hina    Dealing with Materials Data    IITB    5 / 6



```
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.

> help("norm")
> help("dnorm")
> mu=20
> sigma=2
> x <- seq(0,50,0.01)
> plot(x, dnorm(x,mu,sigma))
```

**Values**

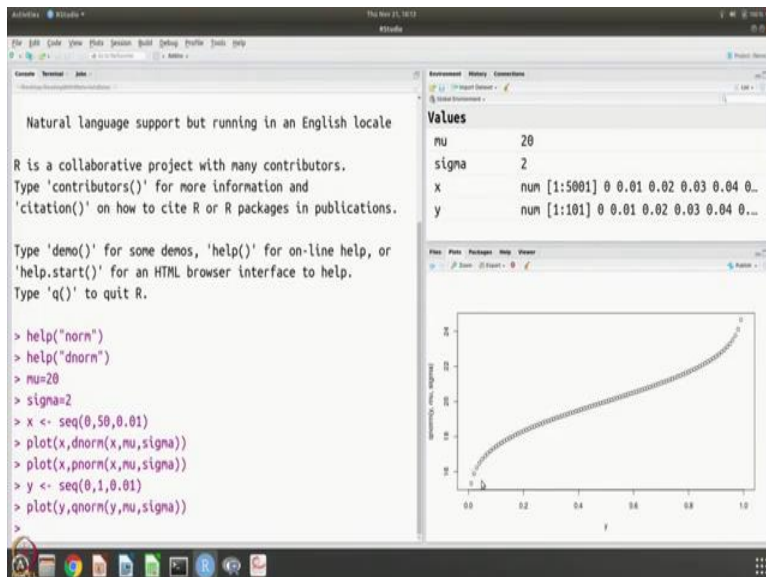
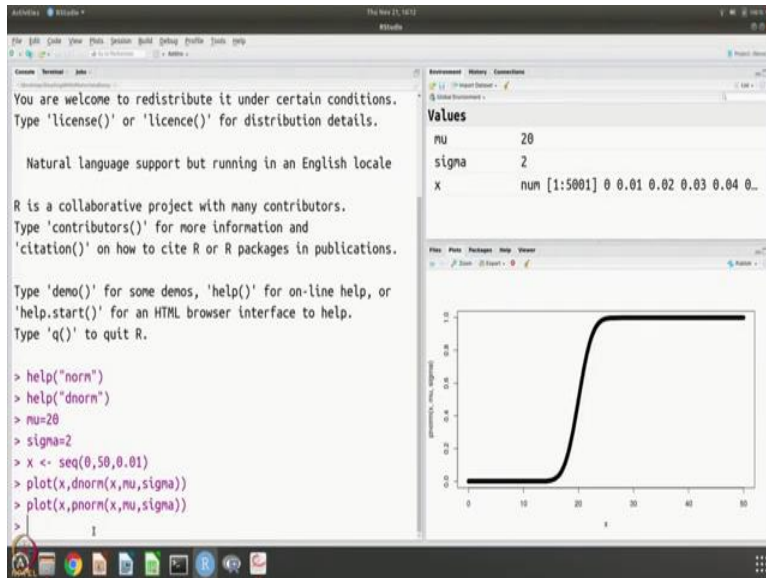
mu	20
sigma	2
x	num [1:5001] 0 0.01 0.02 0.03 0.04 0...

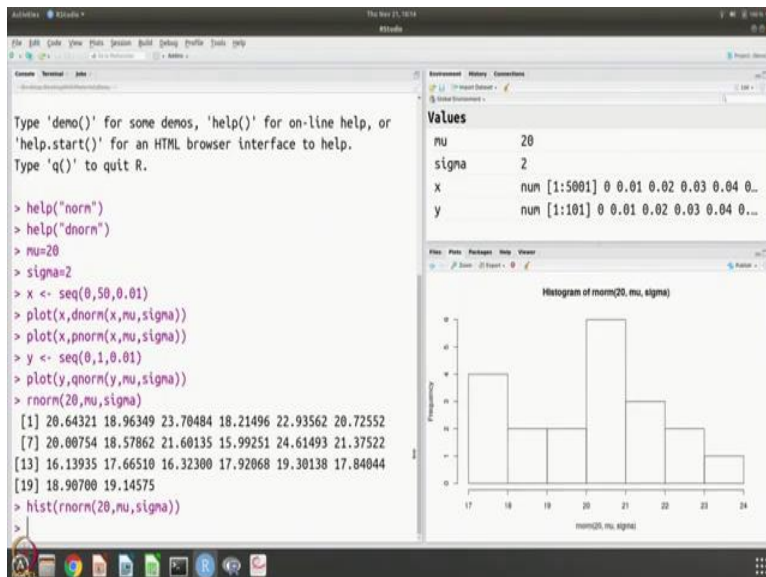
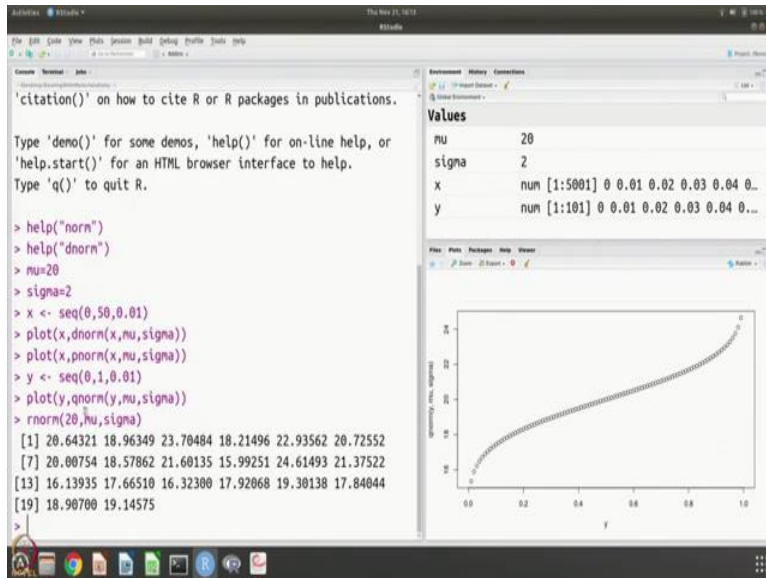
**The Normal Distribution**

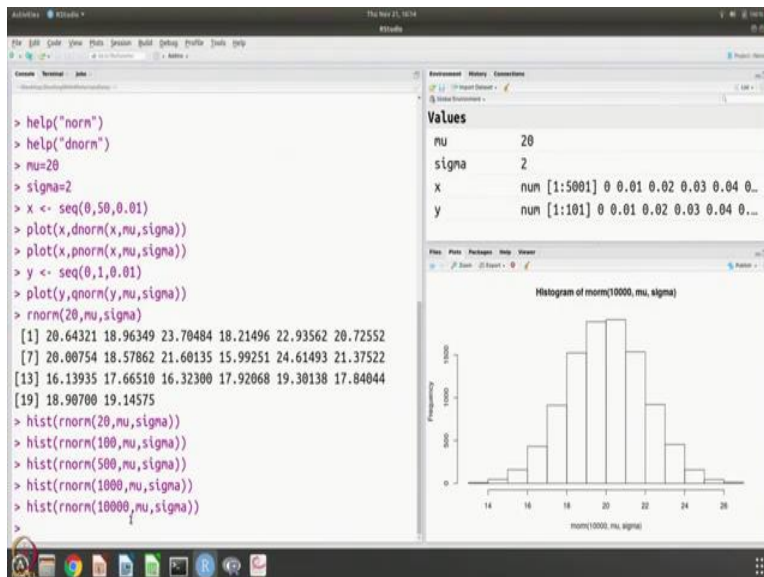
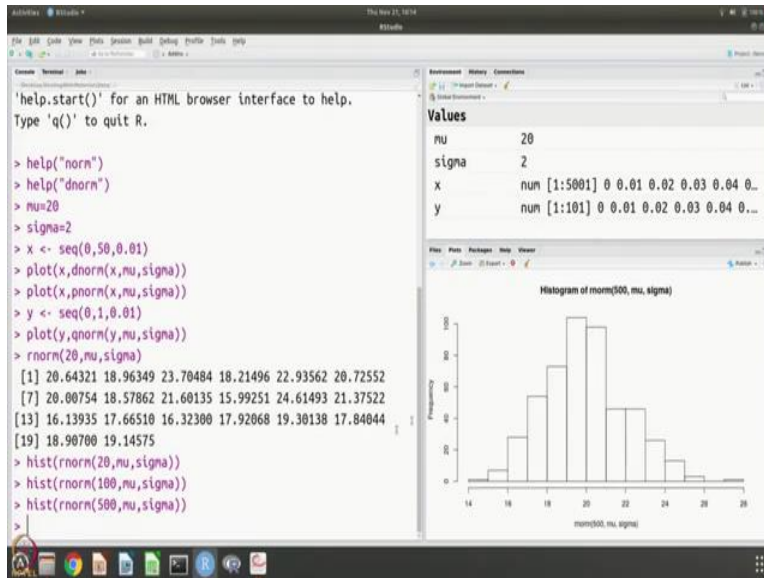
**Description**  
Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

**Usage**  
`dnorm(x, mean = 0, sd = 1, log = FALSE)`  
`pnorm(x, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`  
`qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`  
`rnorm(n, mean = 0, sd = 1)`

**Arguments**  
`x`, `x` vector of quantiles.  
`p` vector of probabilities.  
`n` number of observations. If `length(x) > 1`, the length is taken to be the number required.  
`mean` vector of means.  
`sd` vector of standard deviations.  
`log`, `log.p` logical; if TRUE, probabilities `p` are given as log(p).  
`lower.tail`, `lower.tail` logical; probabilities are `P(X ≤ x)` (otherwise, `P(X > x)`).







And how do we work with normal distribution using our norm is the key word. So, dnorm, pnorm, rnorm will give you the probability density cumulative distribution function and quantile function and random variates and let us assume the mu to be 20 and sigma to be 2 can we get the, so as we did earlier, so let us look up norm. So, ok that is not the norm we want the norm.

So means standard deviation is what you have to give, right? So if you want to have mean to be 20 and the standard deviation to be 2, and so you can plot so let us make x let us say it goes from, it is a sequence and it goes from 0 to some 50 thing. So, you can say x being x comma, mu comma sigma, right. So this is the standard, the normal distribution and you can of course, make the cumulative distribution function.

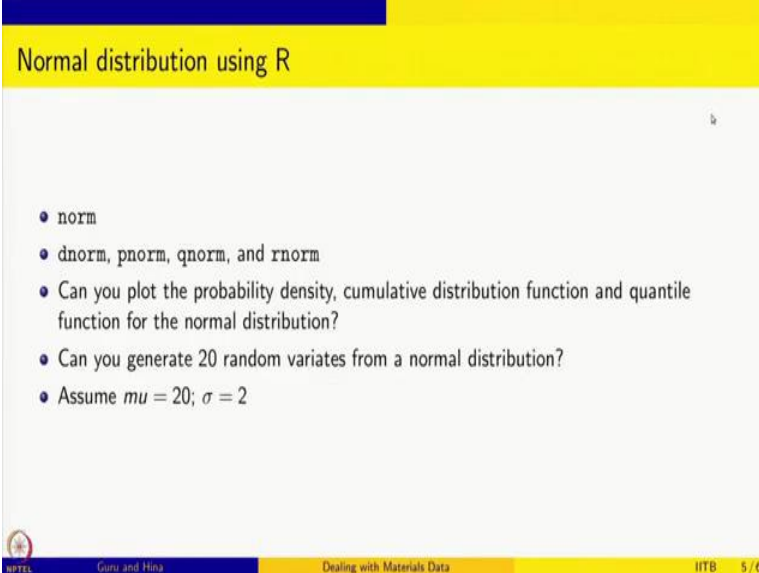


That is just by changing it up pnorm and the here is the cumulative distribution function. So for quantiles we have to make sure that the sequence runs from 0 to 1. So, let us say that we want to say qnorm. That is in terms of y this is y. So, here is the, the quantile function and of course, if you want to generate random variates. So, you have to say rnorm and you have to tell how many random variates you want.

Let us say we want 20 and the mean is mu and standard deviation is sigma. So, we have rnorm. So, you can see that they these are values which are centered around 20 and the sigma is 2, so, it will be between 18 and 22 is 1 sigma and 16 and 24 is 2 sigma. So, that is how these values will be distributed and of course, you can make a histogram plot of this rnorm and see that, so you see that, but if you generate more and more random numbers, and you will see that it is becoming a nice bell shaped curve, right.

So let us say that we want to make some thousand. So you can see now that it is a very symmetric curve, about 20, and a standard deviation of 2. So things are between 18 to 22 and 16 to 24, most of the measurements would fall. So you can make more and more nice looking the normal plots by generating more and more random numbers. So this is 1 aspect.

(Refer Slide Time: 9:01)



Normal distribution using R

- norm
- dnorm, pnorm, qnorm, and rnorm
- Can you plot the probability density, cumulative distribution function and quantile function for the normal distribution?
- Can you generate 20 random variates from a normal distribution?
- Assume  $\mu = 20$ ;  $\sigma = 2$

NPTEL Guru and Hina Dealing with Materials Data IITB 5 / 6

So, now let us go back to normal distribution. So, like I mentioned normal distribution in the context in which we are talking about is also because of errors that you see in measurement.



(Refer Slide Time: 9:11)

Example 1

- Conductivity of ETP copper

NIPTEL Guru and Hina Dealing with Materials Data IITB 6 / 6

And we have already looked at 1 example of conductivity of ETP copper and we actually plotted and we also tried to figure out what distribution it is and we showed that it is normal distribution. Now, that we know the probability distribution function for normal distribution, can we take that data and plot it and also get the do a simulation of random variants from normal distribution and compare the 2. So that is what we want to do in this exercise.

(Refer Slide Time: 9:41)

module: Descriptive statistics using R

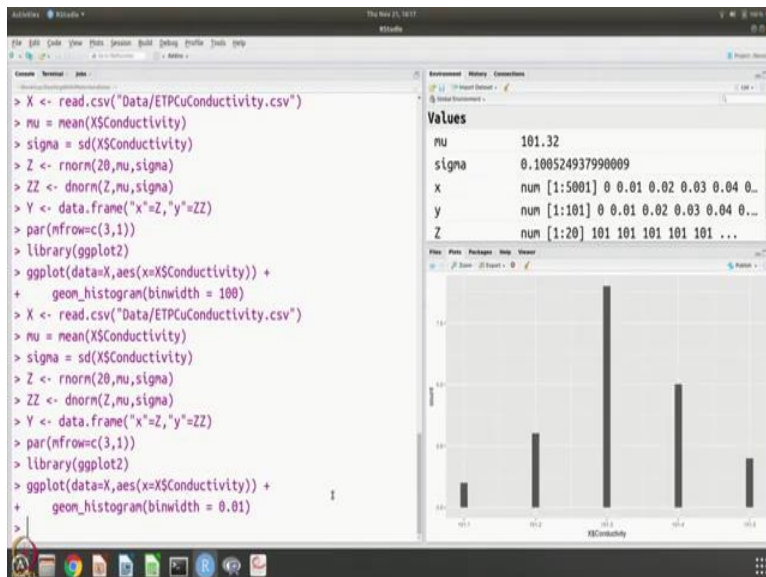
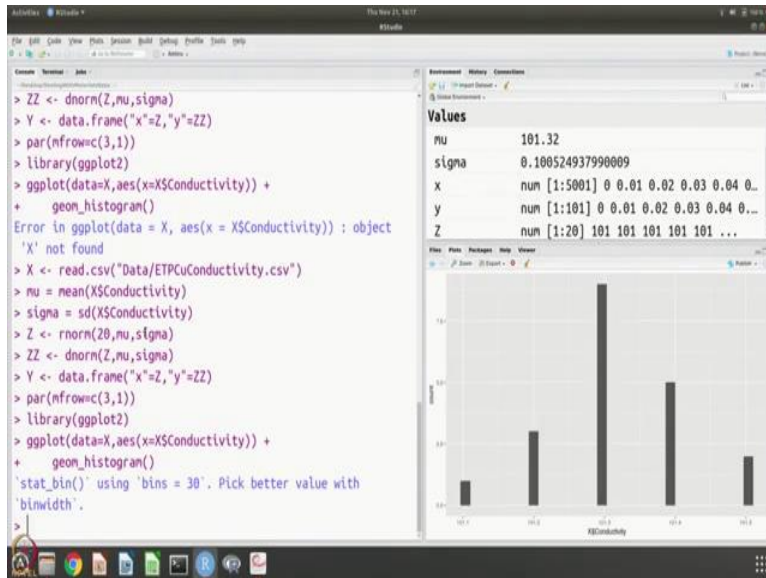
M P Gururajan and Hina A Gokhale  
Indian Institute of Technology Bombay, Mumbai

### 1 Normal distributions using R

```
X <- read.csv("../Data/ETPCuConductivity.csv")
mu = mean(X$Conductivity)
sigma = sd(X$Conductivity)
Z <- rnorm(20, mu, sigma)
ZZ <- dnorm(Z, mu, sigma)
Y <- data.frame("z"=Z, "y"=ZZ)
par(mfrow=c(3,1))
library(ggplot2)
ggplot(data=Y, aes(z, ZZ)) +
  geom_histogram()

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

ggplot(data=Y, aes(z, ZZ)) +
  geom_line()
```



```

1 Normal distributions using R

X <- read.csv("../Data/ETPCuConductivity.csv")
mu = mean(X$Conductivity)
sigma = sd(X$Conductivity)
Z <- rnorm(20,mu,sigma)
ZZ <- dnorm(Z,mu,sigma)
Y <- data.frame("x"=Z,"y"=ZZ)
par(mfrow=c(3,1))
library(ggplot2)
ggplot(data=X,aes(x=X$Conductivity)) +
  geom_histogram()

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

ggplot(data=Y,aes(Z,ZZ)) +
  geom_line()
ggplot(data=X,aes(x=X$Conductivity)) +
  geom_histogram() +
  geom_line(data=Y,aes(Z,ZZ))

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```

```

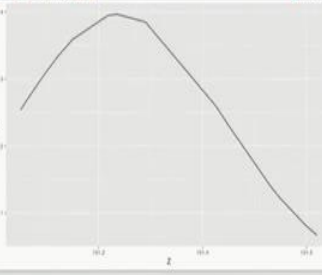
> sigma = sd(X$Conductivity)
> Z <- rnorm(20,mu,sigma)
> ZZ <- dnorm(Z,mu,sigma)
> Y <- data.frame("x"=Z,"y"=ZZ)
> par(mfrow=c(3,1))
> library(ggplot2)
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram(binwidth = 100)
> X <- read.csv("Data/ETPCuConductivity.csv")
> mu = mean(X$Conductivity)
> sigma = sd(X$Conductivity)
> Z <- rnorm(20,mu,sigma)
> ZZ <- dnorm(Z,mu,sigma)
> Y <- data.frame("x"=Z,"y"=ZZ)
> par(mfrow=c(3,1))
> library(ggplot2)
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram(binwidth = 0.01)
> ggplot(data=Y,aes(Z,ZZ)) +
+   geom_line()

```

Environment History Connections

Values

mu	101.32
sigma	0.100524937990009
x	num [1:5001] 0 0.01 0.02 0.03 0.04 0...
y	num [1:101] 0 0.01 0.02 0.03 0.04 0...
Z	num [1:20] 101 101 101 101 101 ...



```

1 Normal distributions using R

X <- read.csv("../Data/ETPCuConductivity.csv")
mu = mean(X$Conductivity)
sigma = sd(X$Conductivity)
Z <- rnorm(20,mu,sigma)
ZZ <- dnorm(Z,mu,sigma)
Y <- data.frame("x"=Z,"y"=ZZ)
par(mfrow=c(3,1))
library(ggplot2)
ggplot(data=X,aes(x=X$Conductivity)) +
  geom_histogram()

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

ggplot(data=Y,aes(Z,ZZ)) +
  geom_line()
ggplot(data=X,aes(x=X$Conductivity)) +
  geom_histogram() +
  geom_line(data=Y,aes(Z,ZZ))

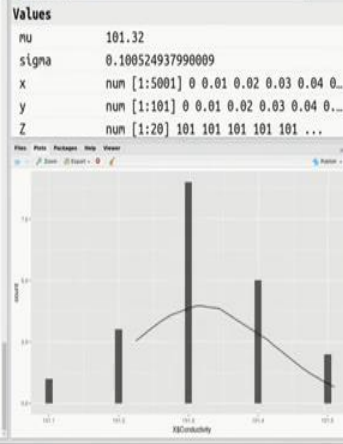
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```

```

> ZZ <- dnorm(Z,mu,sigma)
> Y <- data.frame("x"=Z,"y"=ZZ)
> par(mfrow=c(3,1))
> library(ggplot2)
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram(binwidth = 0.01)
> ggplot(data=Y,aes(Z,ZZ)) +
+   geom_line()
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram() +
+   geom_line(data=Y,aes(Z,ZZ))
'stat_bin()' using 'bins = 30'. Pick better value with
'binwidth'.
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram(0.01) +
+   geom_line(data=Y,aes(Z,ZZ))
Error: 'mapping' must be created by 'aes()'
> ggplot(data=X,aes(x=X$Conductivity)) +
+   geom_histogram(binwidth=0.01) +
+   geom_line(data=Y,aes(Z,ZZ))

```



So for that, I am going to use this set of commands. So here is the script. So, we are going to read the data on ETP, copper conductivity and mu is basically the mean. So, let us calculate from the data and sigma is the standard deviation. So, we are calculated from the data and that is random variate. From a normal distribution, we are pulling out some 20 numbers with this given mu and sigma and z, z is basically the probability density function using z that we have calculated and so we have made a data frame out of this simulated values that we have got.

So we are going to make three plots. First is of course to take the data and plot it as a histogram. So we are using ggplot. So this data is x and aesthetics is take x as conductivity and make a histogram, right. So, that is what case there is a problem in reading the data. So, this is the

distribution you get to and it says something about bins, pick a better value with bandwidth and we can do that.

So you can be a bandwidth let us say sorry. So, you can make this and you see that this is the distribution that you get. Of course, next you can do the plotting of the so, this is the same mean and same standard deviation but we generated the data so it is a basically a simulation and we are going to plot it as a line. So, this is the distribution that you find. Of course, we can put them both together in the same plot.

That is ggplot allows you to do that. So let us do this. So, you can see that these are the data that we measured and this is the simulated distribution that we are getting and, of course here again it is complaining about. So, you can see that we have plotted both on the same plot, and if you generate more and more random numbers, so you will also get this distribution which looks like this. Of course, the total number of data points we have is very small, we have only 20 data points and it still shows a nice normal distribution.

That is because in this case truly the mean value or the actual value of conductivity is somewhere quite close to where the main value is, and all this distribution that you see is because of the errors. So, this is one of the most important reasons why we are interested in normal distribution, random errors are always distributed normally and that is what is going to allow us to do lots of analysis as you will see and we are going to derive lots of distributions from normal and we are going to use them also to understand data.

But there are other uses for a normal distribution in material science and engineering, in general everywhere you will get to see normal distribution. But there are certain specific things which are very relevant to us. One of them for example, is that normal distribution is related to error function and diffusion typically gets solutions which are error function. So, so diffusion problems is one of the problems and it is a stochastic problem.

So, random walk of atoms is what leads to diffusion and so, it is it is a naturally stochastic problem like nucleation it is 1 of the other important stochastic problems and so, we will look at error function for example, in the subsequent sessions. We also had this probability scale that is based on normal distribution. So, we will try to understand probability scale. We tried to identify the

distribution, empirical distribution of a data, whether it is normal or log normal or weibull things like that, so that also is based on the normal distribution at some level.

So, we always use normal as the sort of benchmark for understanding other distributions. Skewness and Kurtosis, for example, is basically to tell how much the distributions deviate from some of the properties of normal distribution in terms of symmetry and in terms of tails and so on. So, we will look at all these aspects one by one and understand normal distribution in greater depth and we will do all that using R in the sessions to come. Thank you.