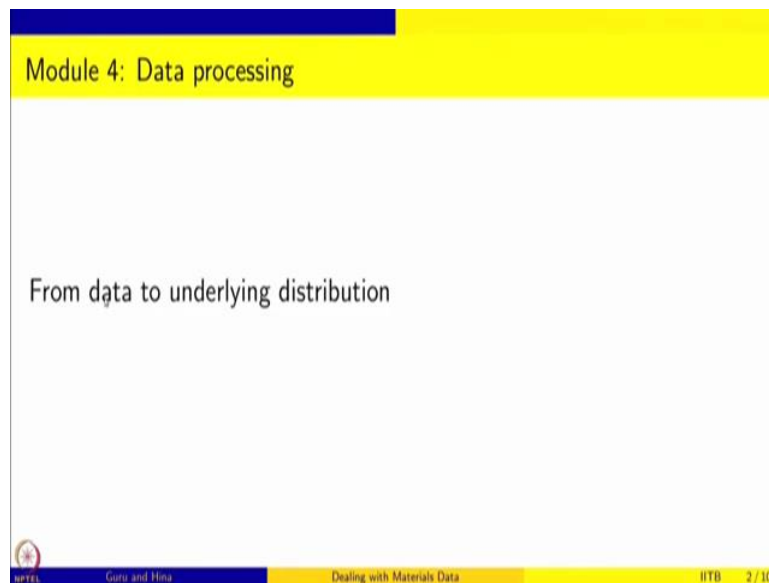


**Dealing with Materials Data: Collection, Analysis and Interpretation**  
**Professor M P Gururajan**  
**Professor Hina A Gokhale**  
**Department of Metallurgical Engineering and Materials Science**  
**Indian Institute of Technology, Bombay**  
**Lecture 65 - From Data to Underlying Distribution**

Welcome to Dealing with Materials Data, we are looking at the collection, analysis and interpretation of data for material science and engineering.

(Refer Slide Time: 0:25)



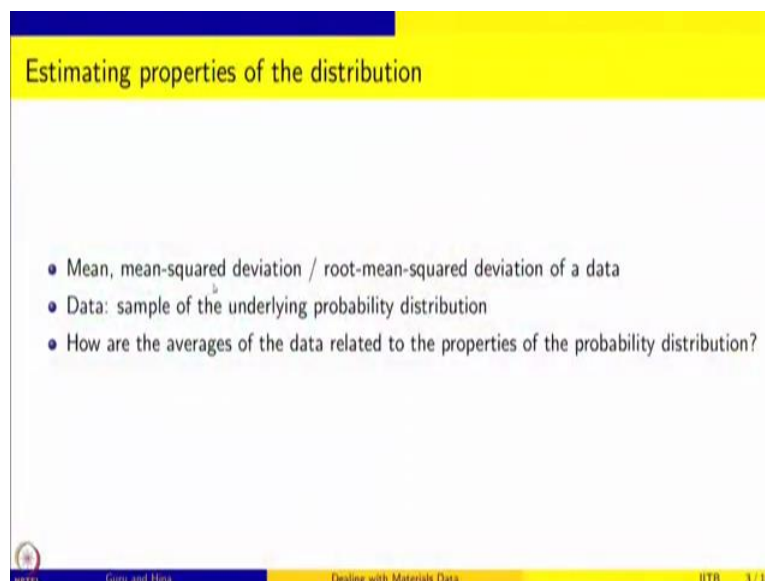
Module 4: Data processing

From data to underlying distribution

IITB 2 / 10

We are in module 4 on data processing, and we are looking at how to go from data to the underlying distribution.

(Refer Slide Time: 0:34)



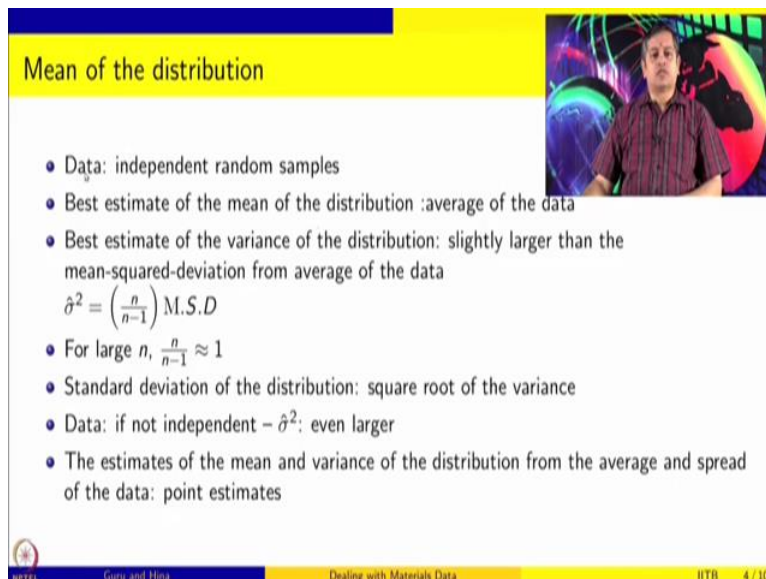
Estimating properties of the distribution

- Mean, mean-squared deviation / root-mean-squared deviation of a data
- Data: sample of the underlying probability distribution
- How are the averages of the data related to the properties of the probability distribution?

IITB 3 / 10

We have known how to estimate properties of a given data set or data series. We can calculate the average and the mean squared deviation and root mean squared deviation from the average of the data. But we know that the data is a sample of the underlying probability distribution. So, we want to get the averages of the, from the averages of the data, the properties of the probability distribution. So how to go from the quantities that we calculate here to the properties of the probability distribution is something that we are going to discuss in this session.

(Refer Slide Time: 1:13)



**Mean of the distribution**

- Data: independent random samples
- Best estimate of the mean of the distribution :average of the data
- Best estimate of the variance of the distribution: slightly larger than the mean-squared-deviation from average of the data  

$$\hat{\sigma}^2 = \left(\frac{n}{n-1}\right) M.S.D$$
- For large  $n$ ,  $\frac{n}{n-1} \approx 1$
- Standard deviation of the distribution: square root of the variance
- Data: if not independent –  $\hat{\sigma}^2$ : even larger
- The estimates of the mean and variance of the distribution from the average and spread of the data: point estimates

Small video inset showing a man speaking.

We are going to assume that the data is independent random samples. And the best estimate of the mean of the distribution is nothing but the average of the data. Best estimate of the variance of the distribution is slightly larger than the mean squared deviation from average of the data.

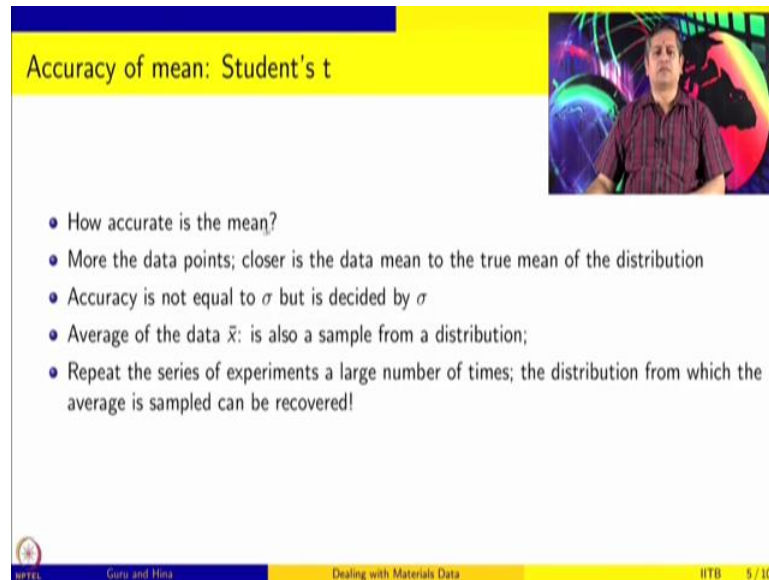
$$\hat{\sigma} = \left(\frac{n}{n-1}\right) M.S.D$$

And standard deviation of the distribution of course, is the square root of the variance. But if this assumption of data being independent is not true, then the variance that you will get will be even larger than what you estimate from this formula. And the estimates of the mean and variance of the distribution from the average and spread of the data, we call this as point estimates, because we are just calculating one number.

For example, we calculate the average of the data and we say that this is the best estimate for the mean of the distribution. We calculate the MSD and from which we can calculate the variance and we say that that is the best estimate for the variance of the distribution. So, these

kind of estimates are point estimates, but sometimes we are also interested in giving interval estimates, which we will also discuss in this session as we go along.

(Refer Slide Time: 2:43)



The slide is titled "Accuracy of mean: Student's t" and features a list of five bullet points. In the top right corner, there is a small video inset showing a man in a red shirt. The slide footer includes the IITB logo, the text "Guru and Hira", "Dealing with Materials Data", and "IITB 5 / 10".

- How accurate is the mean?
- More the data points; closer is the data mean to the true mean of the distribution
- Accuracy is not equal to  $\sigma$  but is decided by  $\sigma$
- Average of the data  $\bar{x}$ : is also a sample from a distribution;
- Repeat the series of experiments a large number of times; the distribution from which the average is sampled can be recovered!

Sometimes we want to know how accurate is the mean that we have estimated. If you have more and more data points, of course, the average that you will get from the data will be closer to the true mean of the distribution. The accuracy of the mean is given by the standard deviation, but it is not equal to the standard deviation. The average of the data  $\bar{x}$  is also a sample from the distribution.

So, if you generate lots of data and lots of such averages, that will actually help you recover the distribution from which the average itself is sampled. So, it is possible to do large number of experiments and to get better estimate for the mean by getting the average from several datasets.

(Refer Slide Time: 3:37)


Variance:  $n$  independent measurements

- $n$  independent series of measurements
- Variance of the average:  $\sigma_x^2 = \frac{\sigma^2}{n}$
- 

$$\hat{\sigma}_x = \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\overline{\Delta x^2}}{n-1}} \quad (1)$$

where  $\overline{\Delta x^2} = \overline{x^2} - \bar{x}^2$

- Note: only when the statistical variations in the measurements are all independent
- Not independent: individual fluctuations do not add quadratically; error becomes larger
- What happens if the data are correlated? Discuss later in terms of correlation length  $n_c$ .



MPYU Guru and Hina Dealing with Materials Data IITB 6 / 10

Now, variance again if you assume that the measurements we are making are all independent, is the, variance of the average

$$\sigma_x^2 = \frac{\sigma^2}{n}$$

$$\hat{\sigma}_x = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\overline{\Delta x^2}}{n-1}}$$

This is true only when the statistical variations in the measurements are all independent. If they are not independent, individual fluctuations do not add up like this and so there becomes larger. What happens if the data is correlated? So, correlation also should be accounted for when we are calculating the variance and how to do it in a particular scenario is something that we will discuss later in the case studies.

(Refer Slide Time: 4:44)

Student's t

- If the measurements are samples from a normal distribution; since  $\hat{\sigma}$  has a spread,  $t = \frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}} \sim T_\nu$  where  $\nu$  is the degrees of freedom of the t-distribution
- Using the t-distribution, now, we can give a confidence interval: the interval within which the true mean will lie with a given probability (or confidence level, say, 50%, 75%, 90%, 99% and so on)
- Interval estimates

npTEL Guru and Hina Dealing with Materials Data IITB 7/10

Now, let us go back to the estimating the mean. If the measurements are samples from a normal distribution, so, we are making here an assumption about the distribution from which this data is sampled. And since the variance has a spread, we have to look at this quantity  $\frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}}$  or  $\bar{x} - \mu$  by  $\hat{\sigma}$  by root n.

$$t = \frac{\sqrt{n}(\bar{x}-\mu)}{\hat{\sigma}} \sim T_\nu$$

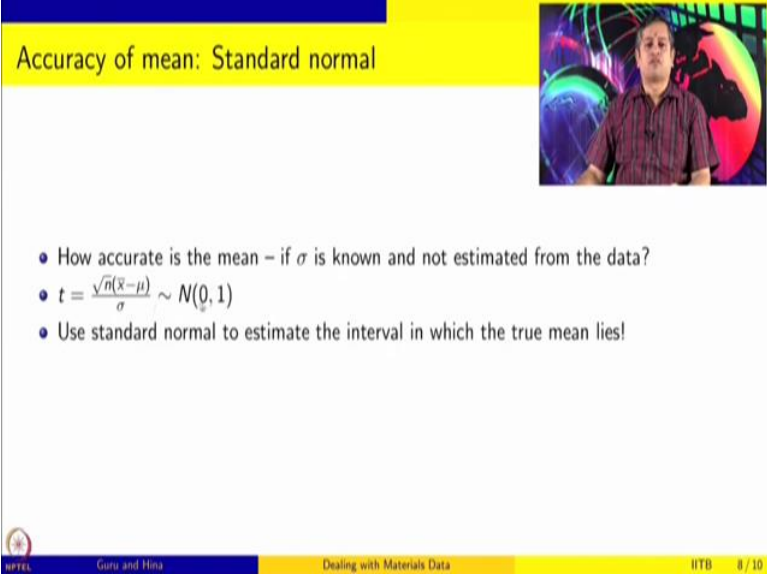
And that goes as t distribution with new degrees of freedom. And this degrees of freedom is determined by the total number of observations n minus 1, because we have already calculated the 1 quantity from the data which is the average.

Using t distribution we can give a confidence interval. By that what we mean is that we can say, with 50 percent probability the mean will lie in this range or with 95 percent probability the mean will lie in this range or with 99 percent probability the mean will lie in this range and so on. This kind of estimates where we are not giving one number for the mean but we are saying what is the range in which the mean will fall is known as interval estimates.

So, you can either give point estimates, you can take the data, you can for example average and give that as a maximum likelihood estimate for the mean of the distribution. Or you can say that okay, the mean will lie in this range with so much of certainty, 90 percent probability that the mean will lie only in this range. So, those kind of things you can make, this is from the t distribution. Assuming normal this is t distribution, because there is the Sigma also which has a spread.

But suppose if the sigma is known exactly, and you do not have to estimate it from the data, then you can put sigma here and you can see that that distribution is actually standard normal.

(Refer Slide Time: 6:48)



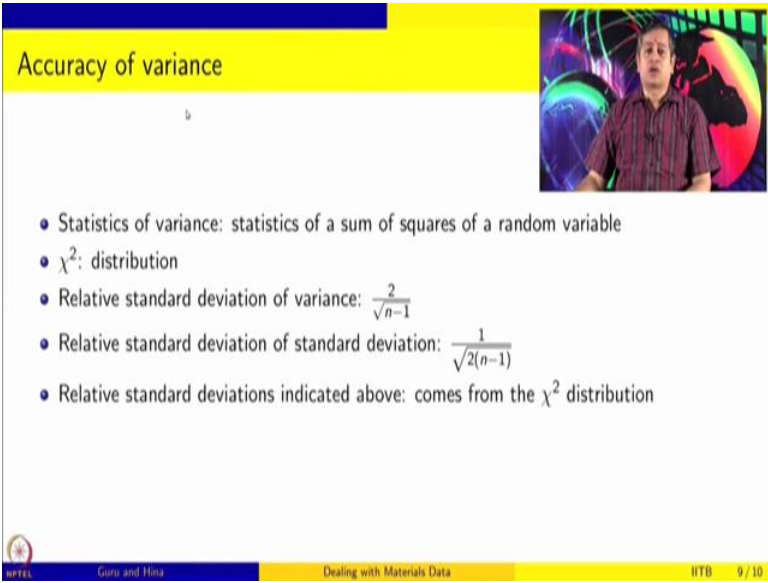
Accuracy of mean: Standard normal

- How accurate is the mean – if  $\sigma$  is known and not estimated from the data?
- $t = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} \sim N(0, 1)$
- Use standard normal to estimate the interval in which the true mean lies!

NPTEL Guru and Hina Dealing with Materials Data IITB 8 / 10

So it is also possible to estimate from the standard normal distribution, the intervals. So you can say, okay, with 90 percent probability, the true mean will lie in this range. And for that, we have to use the standard normal distribution. So whether it is t or standard normal is determined, whether it is sigma hat or sigma. Sigma hat meaning we estimated it from the data, Sigma meaning we knew it, and we did not calculate it from the data.

(Refer Slide Time: 7:14)



Accuracy of variance

- Statistics of variance: statistics of a sum of squares of a random variable
- $\chi^2$ : distribution
- Relative standard deviation of variance:  $\frac{2}{\sqrt{n-1}}$
- Relative standard deviation of standard deviation:  $\frac{1}{\sqrt{2(n-1)}}$
- Relative standard deviations indicated above: comes from the  $\chi^2$  distribution

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

$$\text{Relative standard deviation of variance: } \frac{2}{\sqrt{n-1}}$$

$$\text{Relative standard deviation of standard deviation: } \frac{1}{\sqrt{2(n-1)}}$$

And these were obtained from the chi squared distribution. So, it is possible to estimate also the accuracy of the variance.

(Refer Slide Time: 7:51)

**Accuracy of variance**

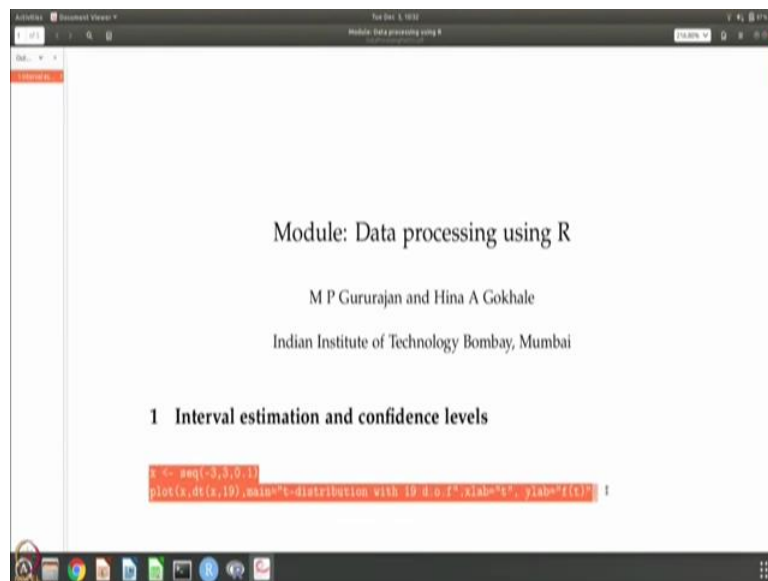
- Independent measurements: relative standard accuracy of  $\hat{\sigma}$  is  $\frac{1}{\sqrt{2(n-1)}}$
- Consider out conductivity case
- Mean: 101.3
- Standard deviation: 0.1
- We have made 20 measurements
- That means, relative inaccuracy of 0.1 is  $1/\sqrt{(38)} = 0.16$
- So, including the error in standard deviation, we should report the conductivity as is  $101.3 \pm 0.12$
- Note, however, given the significant digits, we will still report only 0.1

NPTEL Guru and Hema Dealing with Materials Data HITB 10 / 10

Just by knowing these numbers, one can get these numbers fairly easily. For example, let us say that we have this conductivity data that we are looking at, the mean of which is 101.3 and the standard deviation is 0.1. We have made 20 measurements, so that means a relative inaccuracy in the variance, the 0.1 is 1 by square root 38 and that is 0.16. So the actual number could be 0.1 plus or minus 0.16. So because of which, if we include the error, then we should report the number as 101.3 plus or minus 0.12.

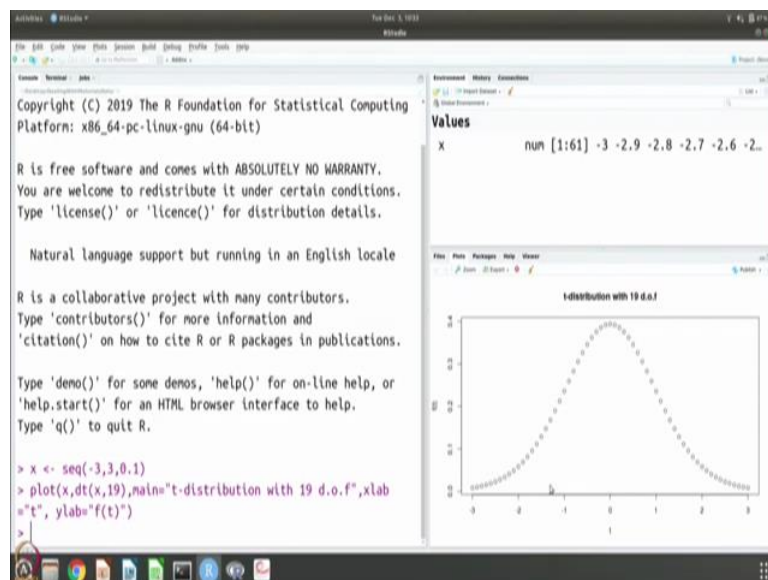
But because of the significance of the digits, we are not willing to go beyond 1 significant digit, because of which we will still report it as 0.1. But if the significant digits are higher, or if this number is not 2 but something like 6 or 7, then the numbers would actually change. So, one way of looking at accuracy of variance is to look at the relative accuracy and report numbers accordingly.

(Refer Slide Time: 9:01)



So we will just go to the data that we have and which is the data on the conductivity. And let us look at the point and interval estimates. And let us also try to understand where these estimates are coming from. So the first thing to do is to, we are going to use the t distribution, because we are going to assume that the data is normally distributed. So let us just plot the t distribution and see how it looks like.

(Refer Slide Time: 9:25)



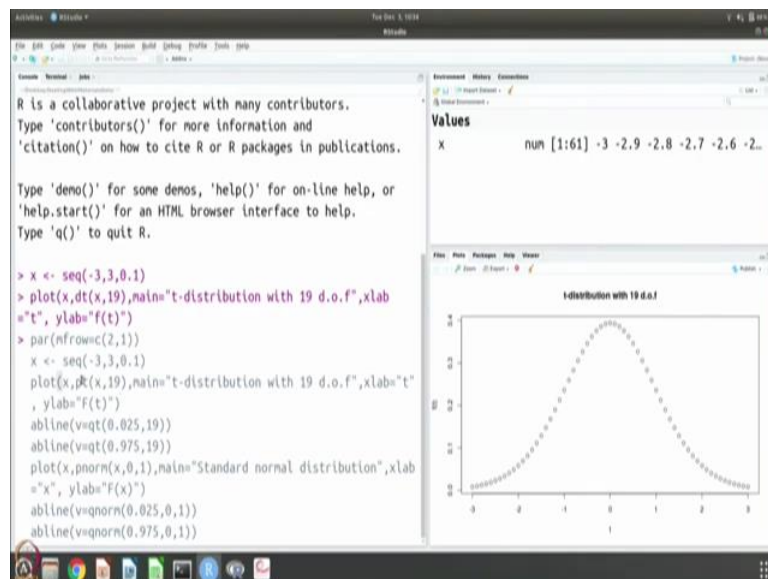
So this is the code. So we are very familiar with this. So we have a sequence minus 3 to plus 3. And we are going to plot this sequence and the probability density function for that sequence with 19 degrees of freedom using the t distribution, so that is what is plotted. And you can see that the t distribution looks like this. Now from this probability distribution function, we know



that the area under this curve is 1, which means for the data to lie anywhere between minus infinity to plus infinity in this curve is 1, that is 100 percent probability that the data will lie.

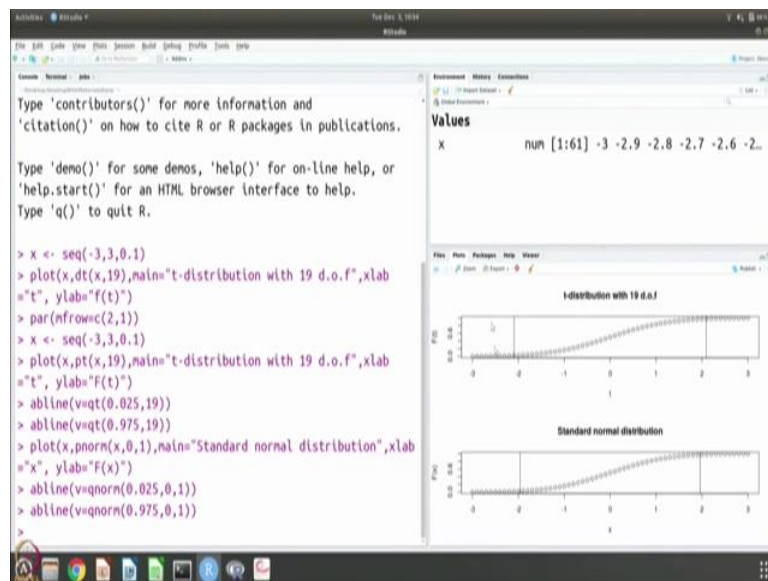
But that is not very useful. So, for example, if you wanted to know 95 percent of data, for example, will lie in what range and it is symmetric about 0. So, if you take off 0.025 on this side and 0.025 on that side, the remaining region will give you the probability that 90 percent of the time the data will fall in this range. So, that is what we are using to give the interval estimates. To know it a little bit better, let us do the other plotting.

(Refer Slide Time: 10:41)



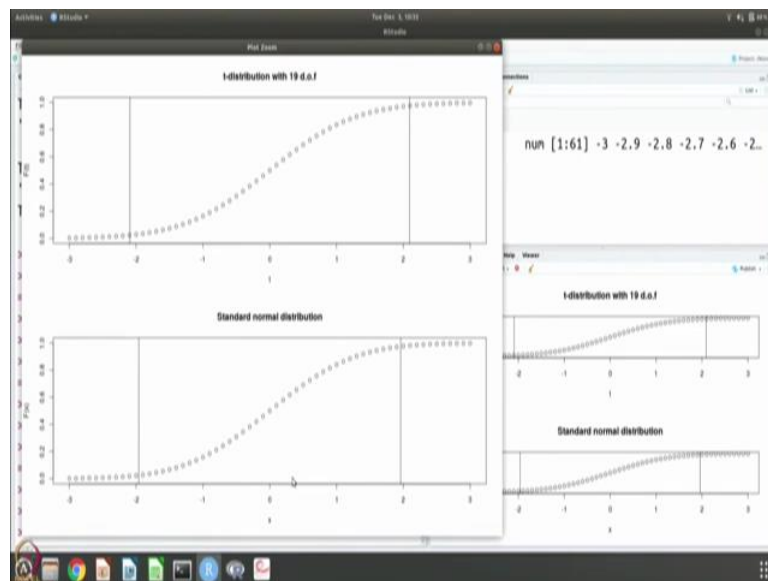
So, we have, we are going to have 2 plots, both are from minus 3 to plus 3. One is a t distribution with 19 degrees of freedom, the other one is normal distribution with, standard normal with the 0 mean and 1 standard deviation. And what we are plotting is the cumulative probability distribution. So pt and pnorm, that is what we are plotting. And we are going to draw 2 vertical lines, one is 0.025, other one is 0.975. Okay, so let us plot this.

(Refer Slide Time: 11:31)



So, you see that, first thing is between t distribution and standard normal distribution, there is a small difference. So let us zoom this and see.

(Refer Slide Time: 11:41)

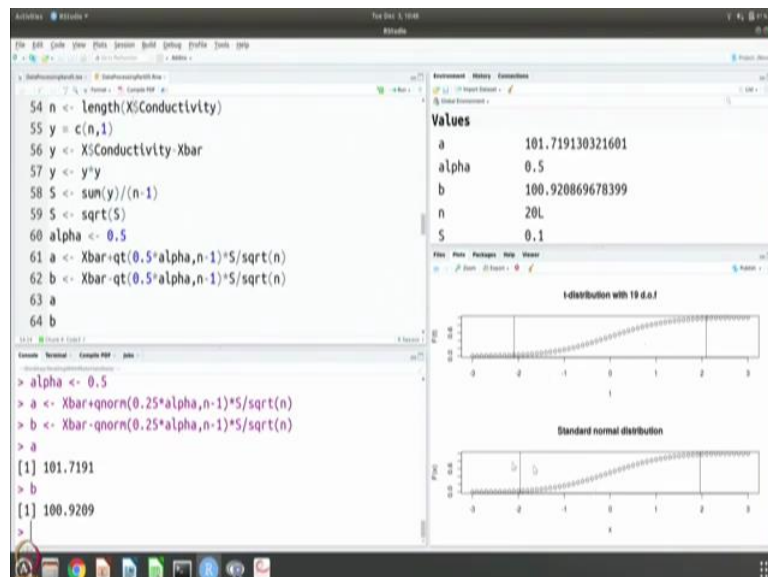


So, you can see that in the case of t distribution 2.5 percent of the data will fall up to this value, whereas in the case of normal it is slightly greater than minus 2. Similarly, on the other side, 97.5 percent of the data will fall from minus infinity to this value, which is slightly greater than 2, but it is slightly smaller than 2 for the normal distribution. In other words, in t distribution and in standard normal distribution, between these two lines, 95 percent of the data will fall because remaining 2.5 is in this tail and in that tail.

Similarly, for standard normal, 95 percent of the data will fall in this range and 2.5 is falling here and 2.5 is falling here. This is how we determine the confidence interval. When we say the probability with 95 percent confidence that the mean will lie in this range. So, this is the range that we are trying to calculate from the given data and then we are giving that range saying that okay, so the mean should lie in this. The true mean is somewhere and our data mean is somewhere, but that distribution is because we are sampling from the distribution.

And depending on whether we know the standard deviation or we do not know the standard deviation we use either normal or t distribution. So, that is what the idea behind determining the interval estimate is.

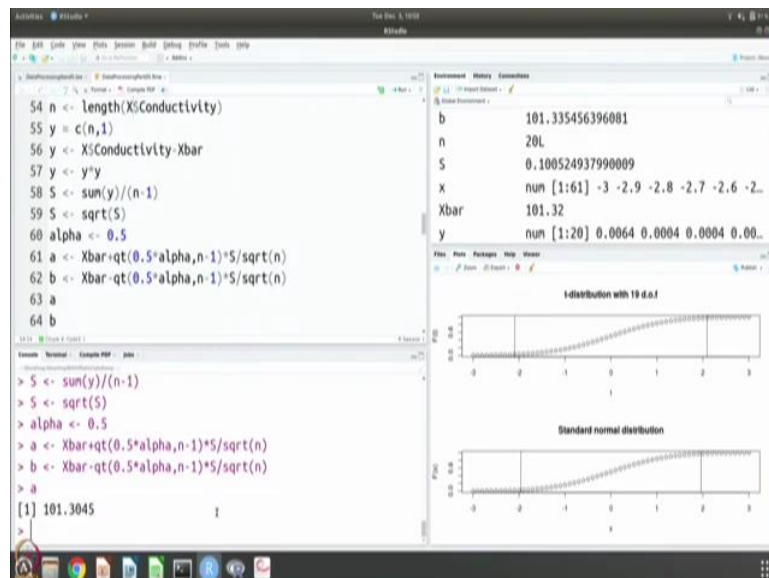
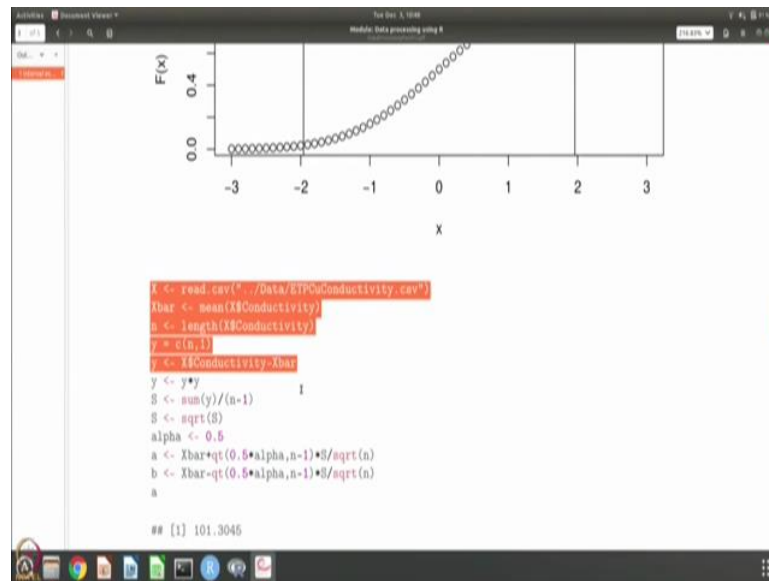
(Refer Slide Time: 13:13)



So, we are looking at the conductivity data and using the t distribution and standard normal distribution, it is possible to estimate the confidence intervals for the true mean for all the conductivity data, in what range will it lie. And to do that, we are going to use the same idea. So, if we are given some confidence intervals. Let us say we want to make sure that 90 percent probability the data should lie in this range or 95 like we have taken.

Which means there is a 2.5 percent probability here and 2.5 percent probability here and same is here. So, which means 5 into 0.5, so, that should be the value that you should calculate to know what this point is. And 0.025 on this side, you should calculate to know where this point is, in the t distribution and in the standard normal distribution. So, and if you do, then if that alpha is 5, for example, 100 into 1 minus alpha, so alpha is 0.05, so that is 95. So 95 percent confidence level you can give this.

(Refer Slide Time: 14:27)



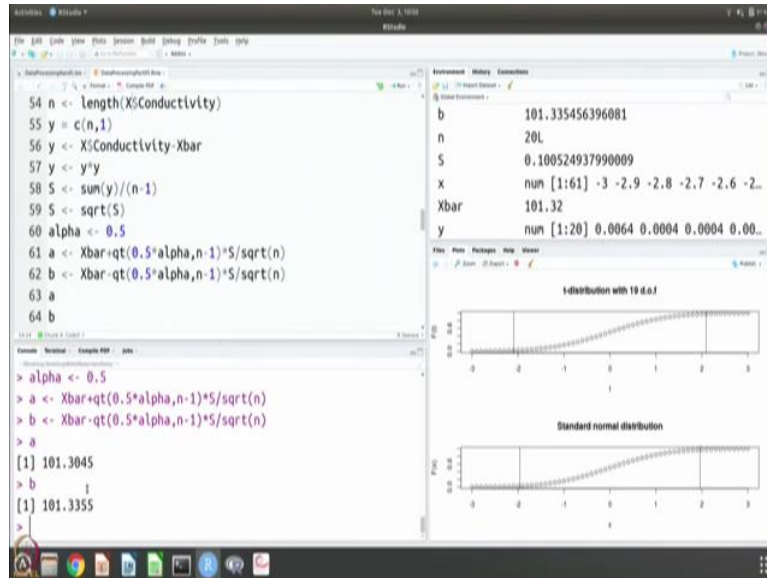
So, in order to do that, let us take the first the conductivity copper data and let us do it here. So, you can see what this computation does. We are going to read the data and we are going to calculate the average value of the data. We know how many data points are there and then we are going to calculate the standard deviation from the data itself. So, this is the standard deviation from the data.

Once we have the standard deviation, let us say we want to know what is the 50 percent probability that the mean will lie, what is the range in which the mean will lie, if we say 50 percent of the times we have to be right. That means the alpha is 0.5 and we will multiply this by 0.5, because we want to have half on this side half on that side. 50 percent is in the middle,

so there is 25 percent on the lower end and 25 percent on the upper end of these stale portions. So, that is why this 0.5 multiplication is there.

And we are using t distribution with n minus 1 degrees of freedom because we have calculated the average, so 1 degree of freedom is gone. And t distribution because we are using the standard deviation which we have estimated from the data itself.

(Refer Slide Time: 15:43)



So, you can get and you will see that the a and b values, so, 101.3045 to 101.3355, in between this the mean will lie and that is with 50 percent probability. Of course, you can make this higher by giving other values. Suppose if you wanted to have 90 percent probability, then 10 percent is the alpha value and so, 0.05, 0.05 on either side. So, this 0.5 multiplying 0.1 will take care of that.

And if we do that, so, if you want to have higher confidence that you want to make sure that not just 50 percent of the times but 90 percent of the times the data should fall, obviously, if 90 percent of the times it has to fall, these points are going to become wider. And that is what you see 101.28, 101.35 as compared to the previous value where it was 30 and 33, right, so 28 to 35. So, the range has expanded and you can do it for other values also.

(Refer Slide Time: 16:58)

```

54 n <- length(X$Conductivity)
55 y <- c(n,1)
56 y <- X$Conductivity-Xbar
57 y <- y*y
58 S <- sum(y)/(n-1)
59 S <- sqrt(S)
60 alpha <- 0.5
61 a <- Xbar+qt(0.5*alpha,n-1)*S/sqrt(n)
62 b <- Xbar-qt(0.5*alpha,n-1)*S/sqrt(n)
63 a
64 b

```

```

b      101.384308258247
n      20L
S      0.100524937990009
X      num [1:61] -3 -2.9 -2.8 -2.7 -2.6 -2...
Xbar   101.32
y      num [1:20] 0.0064 0.0004 0.0004 0.00...

```

t-distribution with 19 d.o.f

Standard normal distribution

```

54 n <- length(X$Conductivity)
55 y <- c(n,1)
56 y <- X$Conductivity-Xbar
57 y <- y*y
58 S <- sum(y)/(n-1)
59 S <- sqrt(S)
60 alpha <- 0.01
61 a <- Xbar+qt(0.5*alpha,n-1)*S/sqrt(n)
62 b <- Xbar-qt(0.5*alpha,n-1)*S/sqrt(n)
63 a
64 b

```

```

> b <- Xbar-qt(0.5*alpha,n-1)*S/sqrt(n)
> a
[1] 101.273
function (...) .Primitive("c")
> b
[1] 101.367

```

```

b      101.367047119184
n      20L
S      0.100524937990009
X      num [1:61] -3 -2.9 -2.8 -2.7 -2.6 -2...
Xbar   101.32
y      num [1:20] 0.0064 0.0004 0.0004 0.00...

```

t-distribution with 19 d.o.f

Standard normal distribution

```

54 n <- length(X$Conductivity)
55 y <- c(n,1)
56 y <- X$Conductivity-Xbar
57 y <- y*y
58 S <- sum(y)/(n-1)
59 S <- sqrt(S)
60 alpha <- 0.5
61 a <- Xbar+qt(0.5*alpha,n-1)*S/sqrt(n)
62 b <- Xbar-qt(0.5*alpha,n-1)*S/sqrt(n)
63 a
64 b

```

```

> alpha <- 0.01
> a <- Xbar+qt(0.5*alpha,n-1)*S/sqrt(n)
> b <- Xbar-qt(0.5*alpha,n-1)*S/sqrt(n)
> a
[1] 101.2557
> b
[1] 101.3843

```

```

b      101.384308258247
n      20L
S      0.100524937990009
X      num [1:61] -3 -2.9 -2.8 -2.7 -2.6 -2...
Xbar   101.32
y      num [1:20] 0.0064 0.0004 0.0004 0.00...

```

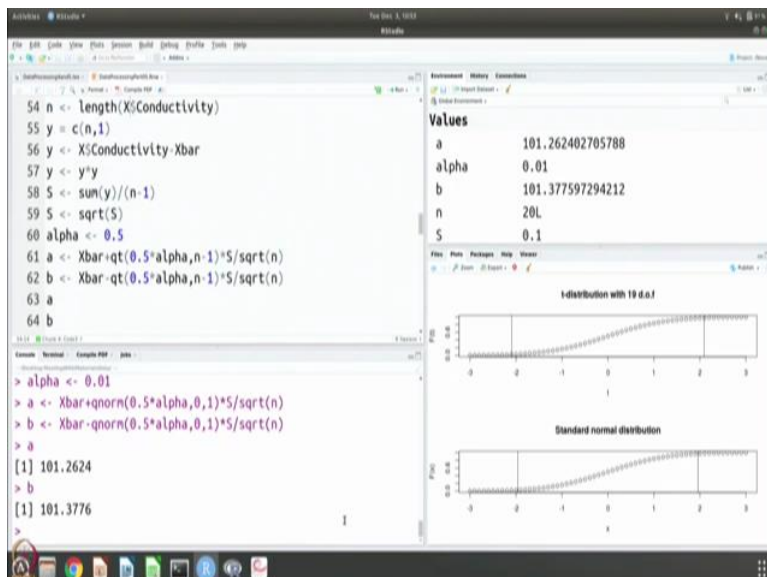
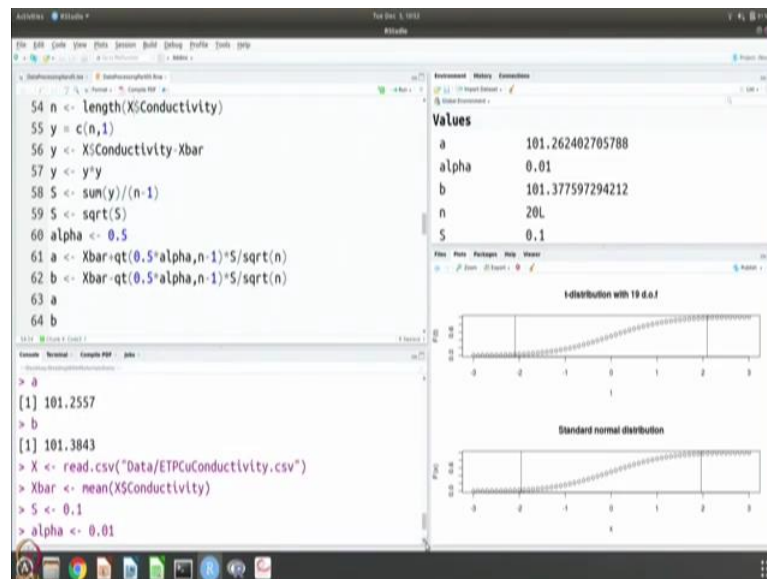
t-distribution with 19 d.o.f

Standard normal distribution



Suppose, if you take 0.05 that is basically the value that we have calculated here. So that is the 95 percent confidence interval that will be between 101.27 and 101.367. And you can of course, calculate still further. Suppose 99 percent confidence interval you want to get, and you will find that that is between 101.2557 and 101.3843. This is all assuming that it is t distribution, but you can also calculate by assuming that it is normal distribution. In which case we do not need all these calculations, we are just going to assume that s is 0.1, let us say.

(Refer Slide Time: 17:52)

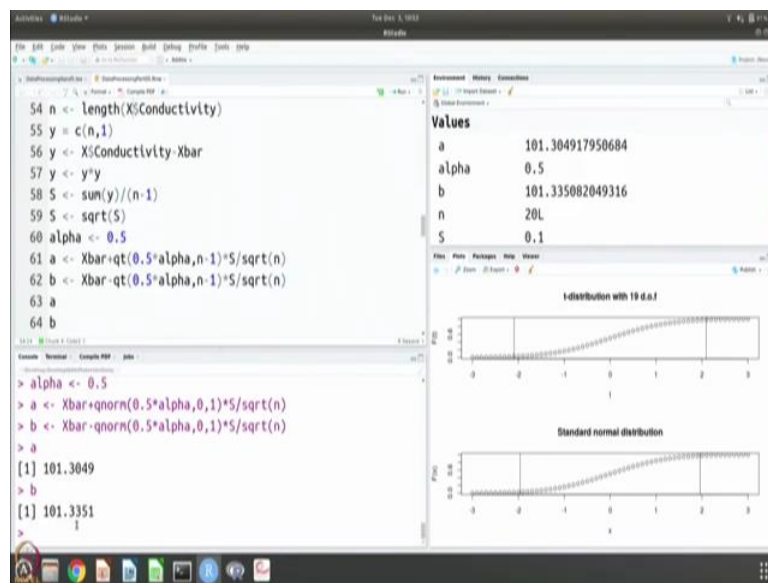


And so, we do not need any of this. We have the mean we have the standard deviation, mean is from the data, standard deviation is assumed to be known. And then in that case, we should not use qt we should use qnorm. And you know, this is standard normal distribution. So it is 0

mean and 1 standard deviation. So you can see this is 101.26 to 101.37, right it is 25 and 38, so obviously t is slightly larger interval than normal.

So 38 has become 37, and 25 has become 26. So normal obviously has a much shorter interval, a little bit shorter, it is not too much short, but it is a bit shorter. So you can do the same thing for normal now with 95 percent or you can do for 90 percent. And as you can see, 95 to 90 if you go 2736 becomes 2835.

(Refer Slide Time: 19:20)



So, if you are okay with 50 percent confidence interval for example, let us say we want 0.5, then we will be between 101.3 and 101.33. So, in this way, we can estimate the interval in which the mean will lie with any given level of confidence. So, to summarize, we have looked at getting estimates from the data for the probability distribution. And there are two that you can get, point estimates, which is a mean and standard deviation. And they can be obtained from the average of the data and spread of the data.

But in addition, if you assume that you know the distribution from which the data is coming, you can also give confidence levels for the value that you are estimating. You can tell with so much probability, the true mean will lie in this range. Specifically, we have looked at the case of standard normal and t distribution and standard normal when the variance is known, t distribution when the variance itself is also calculated from the data.

Finally, there is also a way to estimate the relative error in standard deviation, which is useful when we are reporting the numbers. Because typically we report the numbers as  $\mu$  plus or minus standard deviation. And if there are errors in standard deviation which we may know



from the data, we should accommodate that also when reporting the value and we have seen one example.

So, we have used this copper conductivity data throughout and we have done all these calculations to know how the point and interval estimation works. So, we are going to continue with robust estimation where we do not want to assume anything about the underlying distribution of the data and they are ranked based, there could also be bootstrapping methods. So that we will discuss as part of this session, as part of this module in a different session. Thank you.