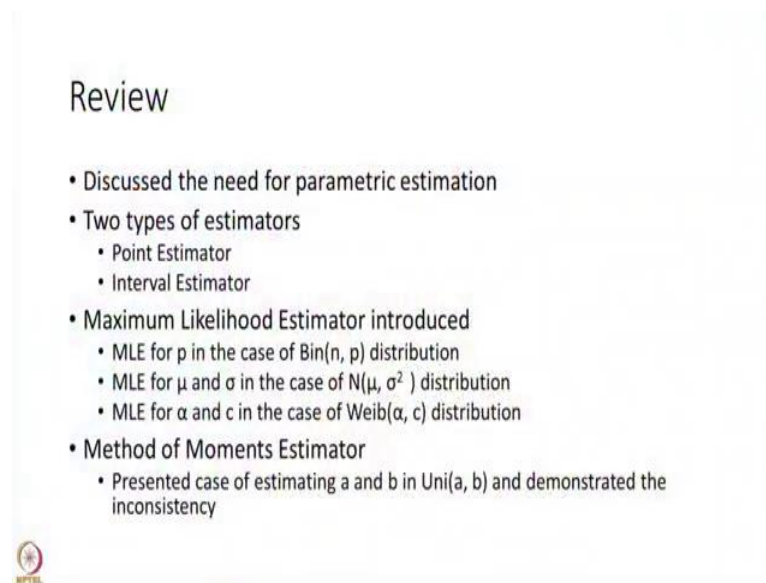


Dealing with Materials Data
Professor M P Gururanjan
Professor Hina A Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 52
Parameter Estimator 3

Hello and welcome to the course on Dealing with Materials Data. In the present sessions we are considering the case of a parametric estimation of population parameters.

(Refer Slide Time: 00:37)



The slide is titled "Review" and contains a bulleted list of topics discussed in the course. The list includes: "Discussed the need for parametric estimation", "Two types of estimators" (with sub-points for Point Estimator and Interval Estimator), "Maximum Likelihood Estimator introduced" (with sub-points for MLE for p in the case of Bin(n, p) distribution, MLE for μ and σ in the case of $N(\mu, \sigma^2)$ distribution, and MLE for α and c in the case of Weib(α, c) distribution), and "Method of Moments Estimator" (with a sub-point for the case of estimating a and b in Uni(a, b) and demonstrating inconsistency). A small logo is visible in the bottom left corner of the slide.

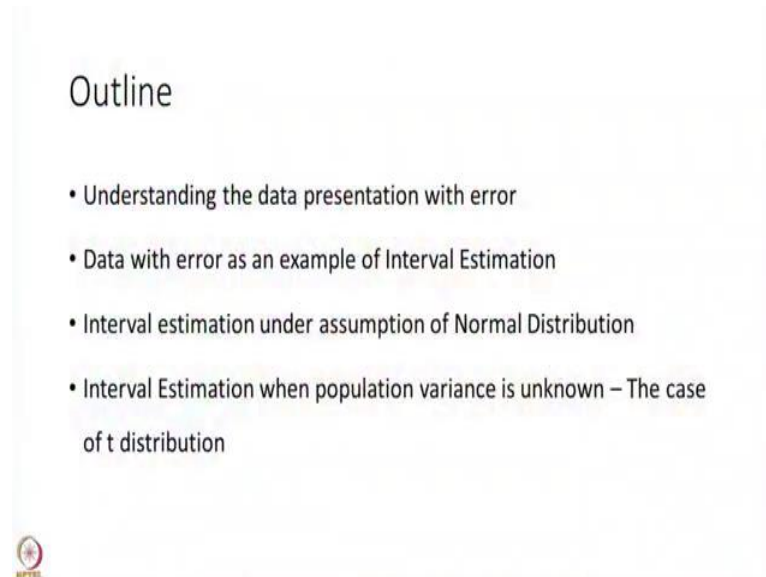
Let us recall that we discussed in details the need for parametric estimation. The need is that, you have an unknown population, you know the population to the extent that it follows a certain distribution with some unknown parameters.

Say follows a distribution f with an unknown parameter θ then if we know θ , we know the distribution therefore we know the population. And therefore our effort is to try to estimate the parameter θ . We said there are 2 types of estimators, point estimator and interval estimator and we discussed point estimators in the previous few sessions in which we talked about a maximum likelihood estimator of a parameter.

And we also talked about the method of movements estimator of parameter. We showed in the case of MLE which is maximum likelihood estimator that sometimes when you try to find a maximum likelihood estimator you may have to go and find a solution numerically. While, in the case of method of movements which is a MME estimator we found that one has to be careful because at times it can give an inconsistent result.

And this we showed by estimating the parameter a and b of a uniform, continuous uniform distribution a and b and with a typical one particular sample of size n we show that it can lead up to inconsistency.

(Refer Slide Time: 02:27)



Now, we are coming to the next session in which we would like to talk about interval estimation. So, let me tell you the interval estimation is not something an unknown area to you. It is something that has been used very regularly in the representation of scientific data that we collect through experiment. So, first we will start by understanding the data representation with the error or with the accuracy as whichever word you use it.

Error is more negative, accuracy is more positive but it talks about the same thing. Data with error as an example of interval estimation then we will do the interval estimation assuming a normal distribution and then we will see that how it also leads to a t distribution. If you want to have an interval estimator with population variance is also unknown, so let us begin.

(Refer Slide Time: 03:32)

Data presentation with Error

- Here is a laboratory data on Linear Thermal Expansion vs. Temperature
- Linear Thermal expansion values have two components
 - Value
 - Range
- What does this signify?

Temp (K)	LinearthermExp
77	-0.38±0.05
200	- 0.28±0.05
535	+0.44±0.14
668	+0.86±0.14
785	+1.458±0.14
889	+2.02±0.14
1008	+2.66±0.14
1061	+2.96± 0.14
1137	+3.60±0.14
1205	+3.90±0.14
1289	+4.40±0.14

Here, you can see on one part of this slide I have a data of a coefficient of linear thermal expansion and the temperature versus the temperature in terms of Kelvin. So, you see that temperatures are straight forward given the (lin) the coefficient of linear thermal expansion has two components in it. For, example here it reach minus 0.38 plus or minus 0.05. When the first value minus 3.0, minus 0.38 is called the value and this is called the range, plus or minus 0.05 represents the range.

(Refer Slide Time: 04:22)

Data presentation with Error

-0.38-0.05 < Data(77) < -0.38+0.05

- The first component "Value" shows the value of average linear thermal expansion measured when several experiments were carried out at the same temperature
- The second component "range" informs that the actual measures in the several experiments carried out at the same temperature varied in this range.
- In general, the second component represents what is called $\pm 1\sigma$ limits

Temp (K)	LinearthermExp
77	-0.38±0.05
200	- 0.28±0.05
535	+0.44±0.14
668	+0.86±0.14
785	+1.458±0.14
889	+2.02±0.14
1008	+2.66±0.14
1061	+2.96± 0.14
1137	+3.60±0.14
1205	+3.90±0.14
1289	+4.40±0.14

Now, the first component actually tells you that this is the average value of linear thermal expansion they have got and the second value which gives you the range it actually represents the one sigma limits that the data has given, if you take sigma as a standard deviation of the

data. Then it generally represents plus or minus 1 sigma limit, it means that the data lies what this really representation says.

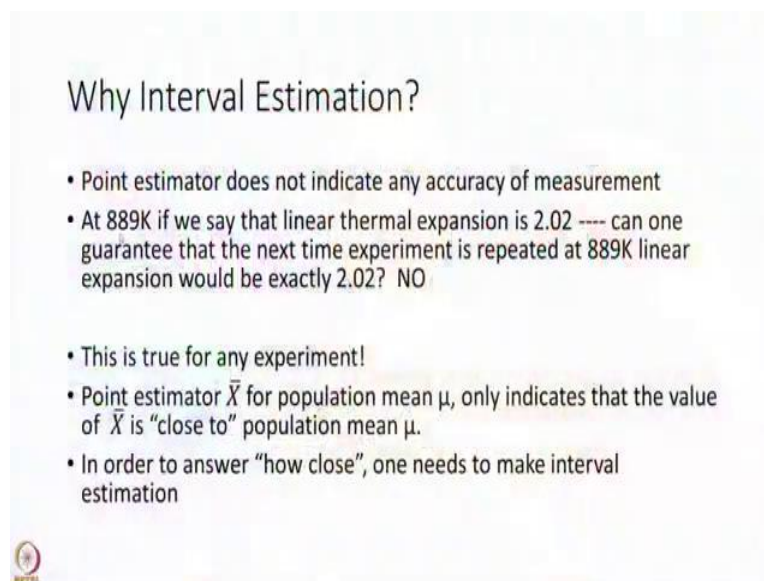
For example, what it says is that your data lies, your data at temperature 77 lies between minus 0.38, minus 0.05 and it is less than minus 0.38 plus 0.05. This I have described more carefully in the next slide, so let us go through it.

(Refer Slide Time: 05:39)

So, here for example, you take a case of 889 Kelvin temperature then at 889 Kelvin temperature, the coefficient of thermal expansion is 2.02 plus or minus 0.14. In statistical terms it means that at temperature 889 Kelvin 68 percent of your data would fall in the range of 1.88 and 2.16. This is called an interval estimation of linear thermal expansion given at a temperature 889 Kelvin. This is what is called an interval estimator.

Remember that if you talk only about 2.02 it is a point estimator when you add a range to it becomes an interval estimator. So, why to be need an interval estimator?

(Refer Slide Time: 06:50)



Why Interval Estimation?

- Point estimator does not indicate any accuracy of measurement
- At 889K if we say that linear thermal expansion is 2.02 ---- can one guarantee that the next time experiment is repeated at 889K linear expansion would be exactly 2.02? NO
- This is true for any experiment!
- Point estimator \bar{X} for population mean μ , only indicates that the value of \bar{X} is "close to" population mean μ .
- In order to answer "how close", one needs to make interval estimation

Well, if you take in this case at 889 Kelvin if we say that the thermal expansion is 2.02, this is what I got when I did my experiment. Suppose someone else does the experiment can you guarantee that it will come to 2.02? No.

So, this is true for any experiment and therefore we say that a point estimator says sample mean, for a population mean only indicates that the value of sample mean is close to the population mean. It gives you an idea where does the population mean lie. But how close it is

or how many times it be closer to this, what all values that can population mean take? These questions are answered through the interval estimator.


(Refer Slide Time: 07:50)

Interval Estimation

- Let X_1, X_2, \dots, X_n be n random measurements of an experiment.
- Assume that these measurement come from a Normal population with mean μ and standard deviation σ .
- $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$, hence $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Therefore, $P[-1.96 < Z < +1.96] = 0.95$

$$\Rightarrow P\left[-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$
- $\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$ is called 95% interval estimator of μ



So, again we start with an example is to go with the normal distribution, let X_1, X_2, X_3, X_n be n random measurements of an experiment and assume that this measurement come from the normal population with mean μ and standard deviation σ

- Let X_1, X_2, \dots, X_n be n random measurements of an experiment.
- Assume that these measurements come from a Normal population with mean μ and standard deviation σ .

$$E(\bar{X}) = \mu \text{ and } Var(\bar{X}) = \frac{\sigma^2}{n}, \text{ hence } Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

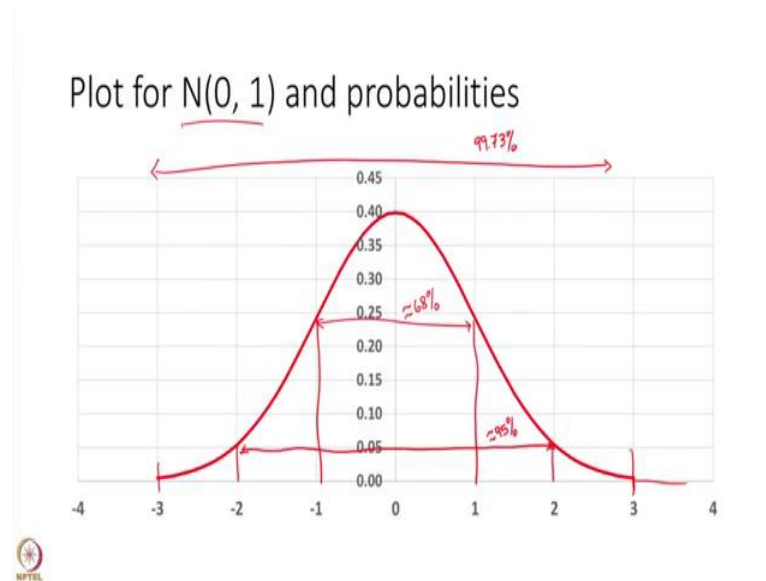
- Therefore, $P[-1.96 < Z < +1.96] = 0.95$

$$\Rightarrow P\left[-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

- $\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$ is called 95% interval estimator of μ

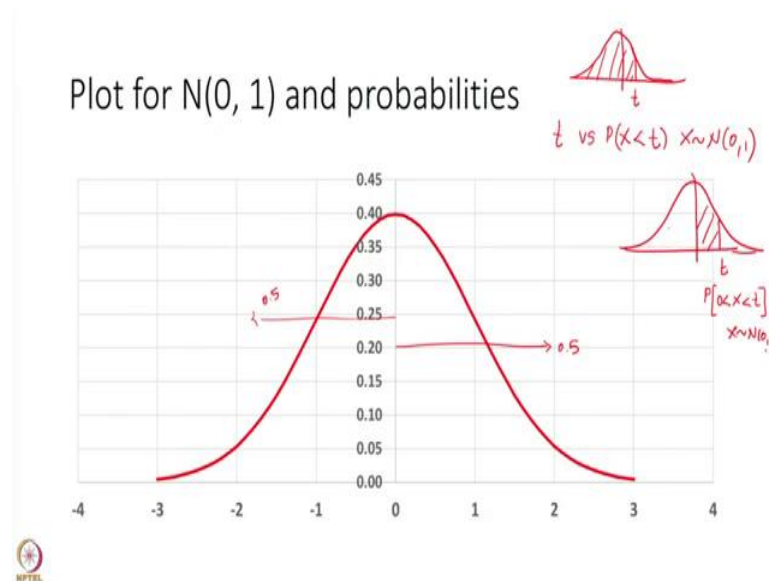
And it simplifies to say that the mean value lies between samples mean minus 1.96 standard deviation divided by square root n and sample mean plus 1.96 standard deviation divided by square root of n. And this probability is 95 it means that, the μ will lie 95 percent of the time between these two limits and this is how it is calculated. Now, where does this 1.96 and come from? Let us try to understand this. (Refer Slide Time: 09:45)



So, here I have a normal probability density plot. This is the normal, standard normal density plot and what we already know that, what we used in the previous case is that, if you take the area under this curve, this area under the curve it represents approximately 68 percent of the data. Okay the probability is 0.68, if you take between minus 2 and plus 2 limit this whole area under the curve is approximately the 95 percent.

It is not a very good approximation but it is little less than 95 percent and if you take between minus 3 and plus 3 limit you can see that it covers almost the whole data. Because the tails are very thin here, so it covers 99.73 percent of the data. This is something we have done in the past I thought we better recall it.

(Refer Slide Time: 11:27)

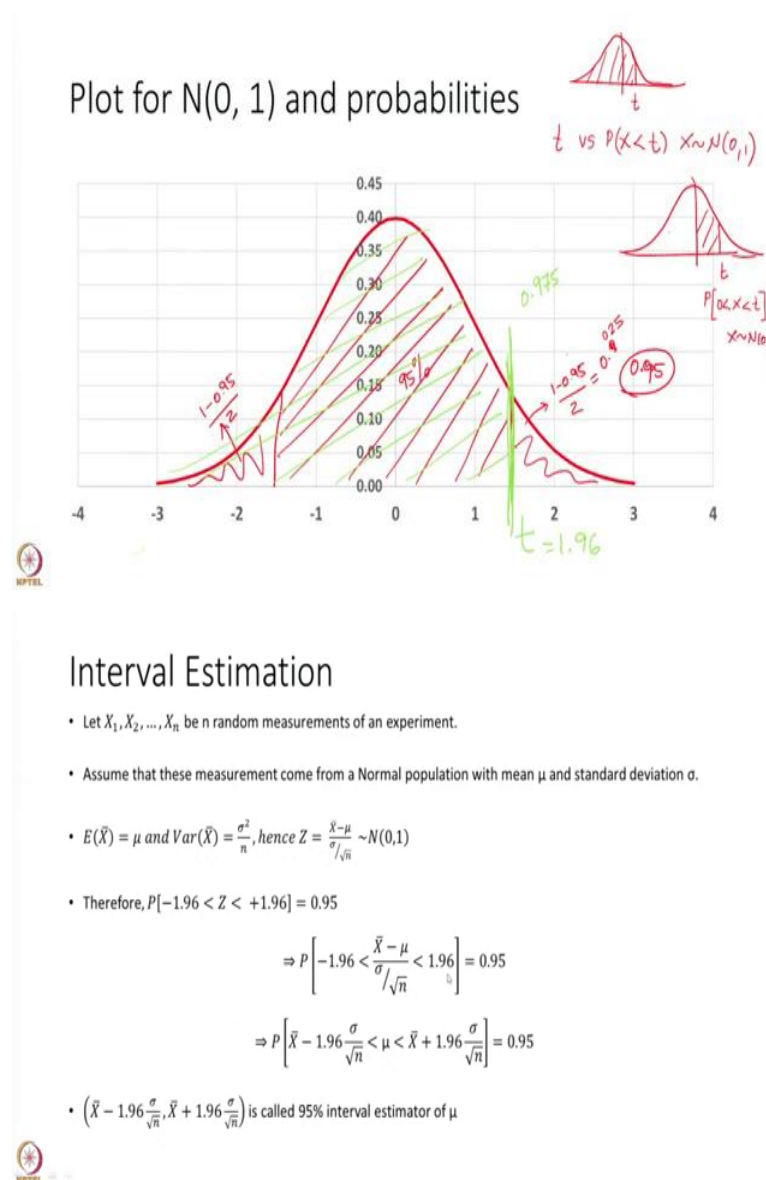


Now, let us consider present case, the normal probability plot or normal probability tables generally give you probability of, sorry-sorry we are going up and up, so if we are in the tables there is normal probability tables. The tables generally have a graph this I have explain to you in the past. Suppose this is the normal, standard normal probability density function then, it gives you the value at a value t .

It has a tabulated the value t versus the probability under this curve of X less than t where X is distributed normal $0, 1$ okay, this is what is given. Sometimes you have to be careful and I believe in are they do this because this both the thing are half-half probability. It divides exactly into two halves the probability this side is half, this probability is 0.5 and this probability is also 0.5 .

So, sometimes they take the value t as this. It means that probability that 0 is less than X is less than t and then that value where X is normal $0, 1$. This is what gets tabulated in the tables, so you have to be careful what you see in the tables. So this 1.96 the question is where did it come from?

(Refer Slide Time: 13:55)



So let us start a fresh. So, the question is where did 1.96 come from? So want to have 95 percent area under the curve. So, let us say that this is the area which we would like it to be 95 percent or 0.95 okay. Now, if the table is like this then we have to we do not have directly the value of there.

So, what we realise is that, this area and this area together that is if you call this area then this area is 1 minus 0.95 divided by 2 and this area is also 1 minus 0.95 divided by 2. This is one is the total area you take out the 0.95 these are symmetric so both areas same, so I have divided them by 2. If you look at this, this value this comes to 0.9 sorry this will come to 0.025 okay.

And therefore, the area now let me change the colour of the pen we make it green, then if you look at this point onwards, this area under the curve the green colour that I am showing which

is, all the area below this particular point is going to be 0.975 okay and therefore, it is this data point that we are looking for and this value turns out to be 1.96. And this how, this value is calculated as 1.96 because you want a 0.95 in the centre area.

(Refer Slide Time: 16:33)


Data Representation

- As in the example of linear thermal expansion data is represented with error of $\pm 1\sigma$ limit. Here σ represents standard deviation of the distribution
- This implies in the present case it would be $Z \sim N(0,1)$

$$P[-1 < Z < 1] \doteq 0.68$$

$$\Rightarrow P\left[\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.68$$

- Linear thermal expansion is given as $\bar{X} \pm \frac{\sigma}{\sqrt{n}}$
- Note that it is also referred as "Data Accuracy"



So, let us move on, so if you look at the data representation it says that as in the example of linear thermal expansion, you have a plus or minus sigma value. What I mean in this case is that, the sigma actually represents the standard deviation of the distribution okay. So, here we will have to say that in that case we would like to have the data that lie between minus 1 and plus 1 limit of the standard normal variate, Z is a standard normal variate.

So you would like to Z lie between this minus 1 and 1,

$$P[-1 < Z < 1] = 0.68$$

$$\Rightarrow P\left[\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.68$$

So this how the data linear thermal expansion data is given in this format. This is what we understand and note that this is also referred as a data accuracy. I hope you have understood let us go through it once again so that this concept is clear.

(Refer Slide Time: 18:06)

Interval Estimation

- Let X_1, X_2, \dots, X_n be n random measurements of an experiment.
- Assume that these measurement come from a Normal population with mean μ and standard deviation σ .
- $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, hence $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- Therefore, $P[-1.96 < Z < +1.96] = 0.95$

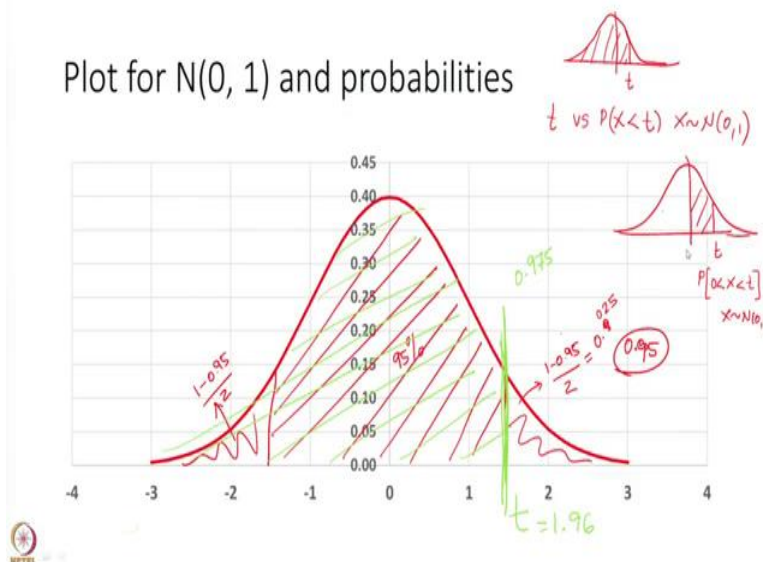
$$\Rightarrow P\left[-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

$$\Rightarrow P\left[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right] = 0.95$$

- $\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$ is called 95% interval estimator of μ



Plot for $N(0, 1)$ and probabilities



The interval estimation what we are trying to do is let us assume that the data, n data comes from a normal population with mean value μ and standard deviation σ . Then we know that the Z which is a standardised or normalised variate of \bar{X} . It is \bar{X} minus μ divided by σ over square root n . Because \bar{X} itself as is a normal variate with a mean μ and variance σ^2 over n and therefore this becomes a standardised or normalised variate, which is varying as standard normal distribution with mean 0 and variance 1.

And therefore here I explained as to how this number 1.96 has come through this particular process that you have the data which is estimated using this method. And the tabulated using this method so it the table has a t versus probability of X less than t . So, we found that if you want to have in the centre 95 percent of the data, it means that you will have 0.25 percent of

the data, 0.25 percent of the data or the two tail ends. If you add up this tail end into this 0.95, you get this green lined area which is the tabulated area.

So, in table you have to look for probability is equal to 0.975 and that t value comes to 1.96. Please remember if you are using R distribution to calculate this t value please make sure read the help. I think in all likelihood it takes a value in this manner it ignores the constant half. So, please make sure how you calculate your t value and accordingly you have to pick up the value from either the table or from the distribution. Then we said that in data representation earlier I say that it is plus or minus 1 sigma limit.

(Refer Slide Time: 20:30)

Data Representation

- As in the example of linear thermal expansion data is represented with error of $\pm 1\sigma$ limit. Here σ represents standard deviation of the distribution
- This implies in the present case it would be $Z \sim N(0,1)$

$$P[-1 < Z < 1] = 0.68$$

$$\Rightarrow P\left[\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\right] = 0.68$$

- Linear thermal expansion is given as $\bar{X} \pm \frac{\sigma}{\sqrt{n}}$
- Note that it is also referred as "Data Accuracy"

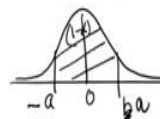
The sigma-sigma can be confusing so here, I am clarifying that sigma by sigma I represent the standard deviation of the distribution. But, in our case because we are considering the case of the standard deviation of the sample mean. And therefore the standard deviation of sample mean will turned to be sigma over square root n and therefore it is calculated in this manner or in linear thermal expansion is given by X bar plus or minus sigma square root n. Note that this is also referred as a data accuracy.

(Refer Slide Time: 21:13)

When σ is unknown

- $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$, hence $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ this is possible when population standard deviation σ is known.
- Suppose σ is unknown.
- Then note that $t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$
- Hence would like to find a and b such that $P[a < t < b] = 1 - \alpha$, where $1 - \alpha$ indicates the confidence level.
- Since t distributions are symmetric about 0, this would simplify to find a such that

$$P[-a < t < a] = 1 - \alpha$$



$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$S = \text{sample std. dev.}$

$$\frac{S^2 \sim \chi^2(n-1)}{\sigma^2} \Rightarrow \frac{(\bar{X} - \mu)^2}{\sigma^2/\sqrt{n}} \sim \chi^2(n-1)$$



When σ is unknown

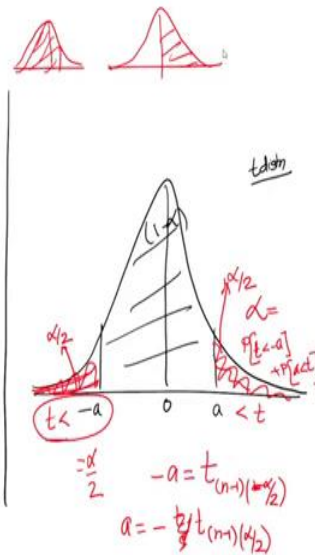
$$P[-a < t < a] = 1 - \alpha$$

$$P[t < -a] + P[t < a] = \alpha$$

Due to symmetry, $2 * P[t < -a] = \alpha$

$$\text{Or } P[t < -a] = \frac{\alpha}{2}$$

$$P\left[-t_{(n-1), \alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{(n-1), \alpha/2}\right] = 1 - \alpha$$



What happens when sigma is unknown, the standard deviation is unknown? You see in the previous case we have assumed that mean is not known and therefore you have given the interval estimation of mean using a standard normal deviate or standard normal variable Z. What if sigma is unknown?

You are all familiar in that case what we do is in standard normal variate Z is defined as

- $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$, hence $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ this is possible when population standard deviation σ is known.

Suppose σ is unknown.

- Then note that $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n - 1)$
- Hence would like to find a and b such that $P[a < t < b] = 1 - \alpha$, where $1 - \alpha$ indicates the confidence level.

S square over sigma square actually. And therefore X bar minus Mu everything divided sigma square divided by S square, over sigma square sorry, S divided by sigma divided by square root n is distributed as Chi square n minus 1 degrees of freedom.

Please recall, we have done this in the past, you can go through the previous slides and confirmed it. Thus, this is a t distribution with n minus 1 degrees of freedom and therefore we would like to find a and b such that probability of a and b, probability of a less than t less than b is 1 over alpha. Where 1 over alpha indicates the confidence level, why we are calling it 1 over alpha? You will know when we go through the session of hypothesis testing to remain consistent with all the explanation.

I am calling it 1 over minus alpha and therefore what we infer is that t distributions are a symmetric distribution around 0. And therefore this would simplify to say that probability of minus a is less than t is less than a is 1 over 1 minus alpha. In other words, these two are if t is a symmetric distribution around 0 then you want to find, the 2 values a and b such that this area is some 1 minus alpha.

$$P[-a < t < a] = 1 - \alpha$$

$$P[t < -a] + P[a < t] = \alpha$$

$$\text{Due to symmetry, } 2 * P[t < -a] = \alpha$$

$$\text{Or } P[t < -a] = \frac{\alpha}{2}$$

$$P \left[-t_{(n-1), \alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{(n-1), \alpha/2} \right] = 1 - \alpha$$

Let us try to understand this through t graph because these are very important points. Just as in normal, say this is a t distribution density function of t distribution, this is 0. We are looking for minus a and plus a such that this area is 1 minus alpha. So, obviously it means that if you

add up let me use a different colour. If I add up this area along with this area, it will be alpha right or that is the, this I call it, this is what is probability of t smaller than alpha and this is probability t greater than alpha.

So, probability t is smaller than minus a plus probability, I am sorry I have said alpha I must correct myself, I am saying it correct now alpha is equal to probability of t less than minus a plus probability of a less than t , this is what has been shown here. Now, these two are also equal so I am saying that, this is alpha by 2 and this alone is also alpha by 2 and therefore you get probability of t less than a is only alpha by 2.

Again you have to look in to the normal probability tables sorry, t probability tables. It will follow the same procedure as normal there also you will have to check, how the probabilities are calculated, is it calculated as all under this curve or is it calculated by taking half here and only this. So, depending on that you decide what should be your value and therefore we call this as, probability we can find this value, we can call the a value is equal to a value of t with n minus 1 degrees of freedom.

Which gives the, this is minus a which gives this probability as 1 sorry which gives a probability as alpha by 2. So, this is what it is given here. So, if then in that case a is minus t over t at sorry it is minus t value with n minus 1 degree of freedom and alpha by 2 probability. And therefore this is what the value it has been given here and we have found this equation that, you have to find these t values from the tables.

Again let us repeat, we find that probability of minus a less than t less than a is $1 - \alpha$. Therefore, the probability at the two tail ends together add up two alpha and therefore only one tail end would add up to only alpha by 2. And therefore the t value that we need to find refers to t at alpha by 2 and we put those two values and we find it. Now, ofcourse if you can find it in this manner (sorry) or you have to make sure which way your table is and then accordingly pick up the value of t . But I am going to represent these values in this manner.

(Refer Slide Time: 30:13)

When σ is unknown

$$\bullet P \left[\bar{X} - \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2} \right] = 1 - \alpha$$

• If $1 - \alpha = 0.95$ and $n = 5$, then $n-1 = 4$ and $t_{4, 0.025} = 2.776$

• The 95% accuracy of the data can be given by

$$\bullet \bar{X} \pm \frac{S}{2} 2.776 = \bar{X} \pm 1.388 S$$



And therefore finally we get, when sigma is unknown you get probability

$$P \left[\bar{X} - \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{(n-1), \alpha/2} \right] = 1 - \alpha$$

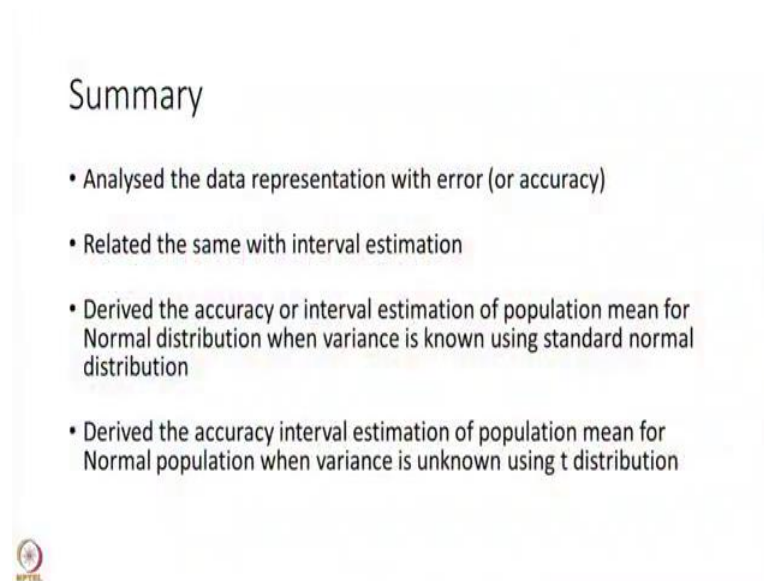
If $1 - \alpha = 0.95$ and $n = 5$, then $n-1 = 4$ and $t_{4, 0.025} = 2.776$

The 95% accuracy of the data can be given by

$$\bar{X} \pm \frac{S}{2} 2.776 = \bar{X} \pm 1.388 S$$

So, when sigma is unknown, you use the t distribution to estimate the interval in which the mean value of the population would lie.

(Refer Slide Time: 31:30)



Summary

- Analysed the data representation with error (or accuracy)
- Related the same with interval estimation
- Derived the accuracy or interval estimation of population mean for Normal distribution when variance is known using standard normal distribution
- Derived the accuracy interval estimation of population mean for Normal population when variance is unknown using t distribution

NPTEL

So let us quickly summarize, we analysed the data representation with the error or with the accuracy and we found that this is same as what we call in statistics interval estimation. In fact it has been derived from statistics only but you are more familiar with the data representation with error. So, we started from that and we said that, that is what is interval estimation.

We derived the interval estimation of population mean for a normal distribution when variance is known. Using standard normal distribution and we derived the interval estimation of population mean for normal distribution or normal population when the variance is unknown using t distribution. Thank you.