

**Statistics for Experimentalists**  
**Prof. Kannan. A**  
**Department of Chemical Engineering**  
**Indian Institute of Technology – Madras**


**Lecture - 17**  
**Confidence Intervals (Part B)**

Fine, let us continue.

**(Refer Slide Time: 00:17)**

**... Interval Estimates When  $\sigma^2$  is Known**

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$
$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$


 NPTEL

We left at this particular point, probability of  $\bar{X} - z_{\alpha/2} \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma/\sqrt{n}$  that is  $= 1 - \alpha$ . A few things to note, we are using capital X here and this is small z representing the upper  $\alpha / 2\%$  points of the standard normal distribution. Sigma is assumed to be known, n is the sample size. This is the unknown population parameter mu and here we are having  $1 - \alpha$ , that alpha divided by 2 is used in the subscript for z.

**(Refer Slide Time: 01:12)**

### ... Interval Estimates When $\sigma^2$ is Known

❖ If  $\bar{x}$  is the sample mean of a random sample of size  $n$  obtained from a normal population with known variance  $\sigma^2$  then the 100(1- $\alpha$ )% confidence interval (CI) on  $\mu$  is given



$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Hence, if  $\bar{x}$  is the sample mean, small  $\bar{x}$  is the sample mean of a random sample of size  $n$  obtained from a normal population with known variance  $\sigma^2$ , then the 100\*1- $\alpha$ % confidence interval on  $\mu$  is given by  $\bar{x} - z_{\alpha/2} \sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2} \sigma/\sqrt{n}$ . So, this is based on the earlier definition. So, the values are so identified on either side of  $\mu$  such that the probability that  $\mu$  lies between these two values is 1- $\alpha$ .


So, what are those values and after identifying those values, we can project them as the lower limit and the upper limit for  $\mu$  and that is what we have exactly done here. We will now look at  $z_{\alpha/2}$ .

**(Refer Slide Time: 02:28)**

### ... Interval Estimates When $\sigma^2$ is Known

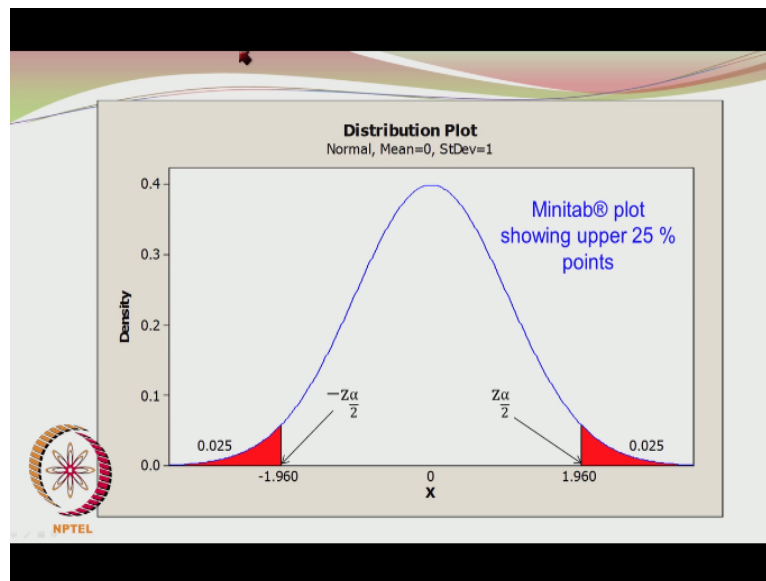
$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Here  $z_{\alpha/2}$  is the upper 100  $\alpha/2$  percentage point of the standard normal distribution.



$z_{\alpha/2}$  is the upper 100\* $\alpha/2$ % point of the standard normal distribution.

(Refer Slide Time: 02:36)




You have the normal distribution sketched here. What is this? This is the standard normal distribution with mean 0 and standard deviation 1. And the upper  $100 \cdot \alpha / 2\%$  points are shown here and also here. This is  $-z_{\alpha/2}$ . We are defining with respect to  $z_{\alpha/2}$  and if we choose  $\alpha$  as 0.05, then  $1 - \alpha$  will be 0.95. So, that would represent the 95% confidence interval and  $\alpha/2$  will then become 0.025 so that 0.025 is the area under the curve beyond the point  $z_{\alpha/2}$ .

So, the probability of the standard normal variable taking a value  $> z_{\alpha/2}$  is 0.025, when  $\alpha$  is 0.05. So, that is what is done here and according to symmetry, if you have located  $z_{\alpha/2}$  here for this particular case of  $\alpha$  0.05  $z_{\alpha/2}$  takes the value of 1.96. So, probability of the standard normal variable  $> 1.96$  is 0.025. Similarly, here you are having -1.96 and that represents  $-z_{\alpha/2}$  and the probability of the random variable  $z$  taking a value  $< -1.96$  is 0.025.

(Refer Slide Time: 04:45)

**... Interval Estimates When  $\sigma^2$  is Known**

It is useful to understand that just as  $X$  is a random variable,  $\bar{X}$  the point estimator is also a random variable.




$X$  is a random variable  $\bar{X}$  the point estimator is also a random variable.

**(Refer Slide Time: 04:54)**

**... Interval Estimates When  $\sigma^2$  is Known**

The **confidence interval (CI)** we have constructed is bounded by two random variables.

Hence the confidence interval itself is also a random variable and hence can theoretically have different bounds.



And the confidence interval we have constructed is bounded by two random variables. We can see here  $\bar{x} - z \alpha/2 \sigma/\sqrt{n} \leq \mu \leq \bar{x} + z \alpha/2 \sigma/\sqrt{n}$ . So, this came from the definition of the confidence interval and  $\bar{x}$  is a random variable and  $\bar{x} - z \alpha/2 \sigma/\sqrt{n}$  is also a random variable.  $\bar{x} + z \alpha/2 \sigma/\sqrt{n}$  is also a random variable. So, the confidence interval is based on two random variables.


Hence the confidence interval is also a random variable and it can theoretically have different bounds.

**(Refer Slide Time: 05:57)**

**... Interval Estimates When  $\sigma^2$  is Known**

The results below show the probabilities for different sample means ( $P(X \leq \bar{X})$ ) when

- ❖  $\mu^* = 50$  (assumed for illustration),
- ❖  $\sigma = 15$  (standard deviation of the normal population)
- ❖  $n = 16$  (sample size).




To demonstrate what is meant by a confidence interval, we will take a small example. Probability of  $X < \bar{X}$ , we will find after assuming that  $\mu^* = 50$ , I am putting star because it is assumed. The  $\mu$  is assumed. We really do not know the exact value of the population mean but for the purpose of demonstration, let us see what happens if  $\mu$  is 50.  $\sigma = 15$ , standard deviation of the normal distribution and  $n = 16$ , which is the sample size.

**(Refer Slide Time: 06:45)**

**... Interval Estimates When  $\sigma^2$  is Known**

The probabilities of various sample means  $\bar{X}$  belonging to the population of mean 50 may be computed as shown in table below. Note that



$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{15}{\sqrt{16}} = 7.35$$

So, the probabilities of various sample means  $\bar{X}$  belonging to the population of mean 50 may be computed as shown in the table below. What we are doing is we are going to take different sample means and we are going to see what is the probability of choosing that sample mean value or lower from a distribution of sample means centered around 50. So, given  $\mu$  assumed  $\mu^* = 50$  what is the probability of a random variable  $\bar{X}$  taking the specified value of  $\bar{X}$  or lower.

So, we are going to find that probability using the assumed  $\mu^*$  and the known  $\sigma/\sqrt{n}$ . So, we are also going to fix the sample size and for example, we have  $n=16$ , we know  $z_{\alpha/2}$  from the previous graph corresponding to the  $\alpha$  value of 0.05,  $z_{\alpha/2}$  had a value of 1.96. So, we take  $Z_{\alpha/2}$  as 1.96 for the 95% confidence interval. This becomes 1.96,  $\sigma$  we assume or take that value as 15. We have taken a sample of size 16.

So, if you plug in these numbers in that formula we will get  $Z_{\alpha/2} \sigma/\sqrt{n}=7.35$ . So, this number, you please remember.

**(Refer Slide Time: 08:54)**

$\mu^* = 50, \sigma = 15, n = 16$

Sample Mean (SM)	$z = (SM - \mu) / (\sigma / \sqrt{n})$	Probability ( $Z \leq z$ )
40	-2.667	0.0038
41	-2.4	0.0082
42	-2.133	0.0164
$50 - 7.35 = 42.65$	-1.96	0.025
44	-1.6	0.0548
46	-1.067	0.1431
48	-0.533	0.2970
50	0	0.5
52	0.533	0.7031
54	1.067	0.857
56	1.600	0.945
$50 + 7.35 = 57.35$	1.960	0.975
58	2.133	0.984

Right, this is a very interesting table. You have sample means here, you can have infinite number of sample means because it is a continuous distribution. You can have 40, 40.001, 40.0001 etc. So, there can be infinite sample means for the purpose of demonstration, I have chosen increments of 1 up to 42. Then you have 42.65 and then increments of 2 from 44, 46 and so on to 56, then you have 57.35 and then 58. We calculate  $Z$ ,  $Z$  is  $\bar{X} - \mu / \sigma / \sqrt{n}$  and that comes as  $40 - 50 / 15 / 4$ .

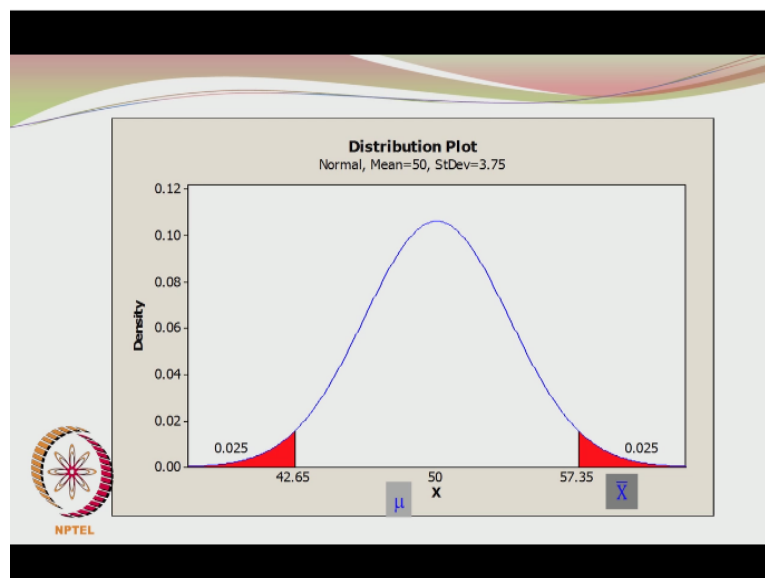
So, I compute these  $Z$  values because I have to convert it into the standard normal variable form. So, these are the values taken by the standard normal variable. So, I have values from -2.67 to 2.13 and then I am finding the probability that  $Z \leq z$ . What is the probability that the sample mean of 40 or lower could have come from the sampling distribution with mean 50 and standard deviation  $15/4$ ,  $15/4$  is 3.75. So, we are having a probability distribution centered around 50 and having a standard deviation of 3.75.

So, what is the probability of picking up a sample mean of 40 or lower from that distribution. And that comes as 0.0038. So, that is a very small chance of picking up a random sample with mean value of 40 from a sampling distribution of means centered around 50 and standard deviation of 3.75. Then you have 41, the probability increases slightly and when you come to 42.65, then the Z value is -1.96 and the probability is 0.025 okay.

And then the probability values keep increasing and then we have 57.35, which is  $50 + 7.35$ , which is  $\mu + z \alpha/2 \sigma/\sqrt{n}$ . We saw that value  $Z \alpha/2 \sigma/\sqrt{n}$  as 7.35. So, we have  $\mu + Z \alpha/2 \sigma/\sqrt{n}$  as 57.35. Here you have  $50 - 7.35$  or  $\mu - Z \alpha/2 \sigma/\sqrt{n}$  and that comes as 42.65. So, at 42.65 the probability of  $Z \leq z$  is 0.025 and the probability of  $Z \leq z$  at a sample mean of 57.35 is 0.975.

So, you are having 0.025 here so, 42.65 and 57.35 you are having 0.975. the probability difference is  $0.975 - 0.025$  which is 0.95. In other words, we can say what is the probability of picking up a sample mean between 57.34 and 42.65 when the sampling distribution is centered around 50 with a standard deviation of 3.75. So, that probability is  $0.975 - 0.025$  which is 0.95.

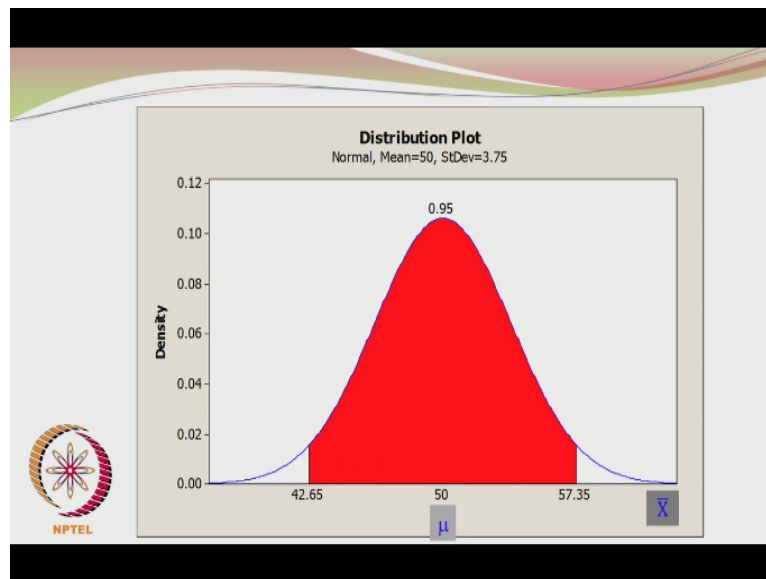
**(Refer Slide Time: 13:25)**



So, I have represented the information given in the table graphically. Here, we have 42.65, here we have 57.35. the probability of the random sample having a value  $\leq 42.65$ , when the distribution is centered around 50 and the standard deviation is 3.75 is 0.025 here. And this entire probability for  $\bar{X} < 57.35$  is 0.975. So, the region between 42.65 and 57.35 will have

an area under the curve of 0.95 or the probability of  $\bar{X}$  lying between these two values will be 0.95.

**(Refer Slide Time: 14:22)**




That is what I have done here, I have shaded that portion and that comes as 0.95. This plot and the previous one were generated using Minitab.

**(Refer Slide Time: 14:37)**

**... Interval Estimates When  $\sigma^2$  is Known**

The plots generated with MINITAB® show that the sample means with values of 42.65 and 57.35 are the 95% confidence lower and upper bounds on the population mean  $\mu$  of 50.



So, the plots generated with Minitab show that the sample means with values 42.65 and 57.35 are the 95% confidence lower and upper bounds on the population mean  $\mu$  of 50.


**(Refer Slide Time: 15:02)**



**... Interval Estimates When  $\sigma^2$  is Known**

The probability of this

i.e.  $P(42.65 < \mu < 57.35) = 0.95$




The probability is  $\mu$  lying between 42.65 and 57.35 is 0.95.

**(Refer Slide Time: 15:19)**

**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ However, during the actual sampling process, you get **only one sample** and hence only one sample mean  $\bar{X}$ . Then how do you know that your sample's CI bounds  $\mu$ ?
- ❖ Note that we assumed  $\mu^* = 50$  for illustration only. Hence  $\mu$  is also not known.



During the actual sampling process, we get only one sample and hence we get only one sample mean  $\bar{X}$ . So, how do you know that your sample's confidence interval bounds  $\mu$ ? We have assumed  $\mu^* = 50$  for illustration purposes only.  $\mu$  also may not be 50, it may be some other value.


**(Refer Slide Time: 15:52)**

### CI Approach

- ❖ But we are only working on the CI length around  $\mu$ .
- ❖ The results below show the confidence interval length given by

$$|(\bar{X} - \mu)| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Here we do not need to know  $\mu$ .



We are working on a confidence interval length around  $\mu$ . What is that confidence interval length? That is what we are going to see. The confidence interval length is defined as the absolute value of  $\bar{X} - \mu$ . The distance between  $\bar{X}$  and  $\mu$  is termed as the length and we are taking the positive value only. So, we put a modulus on  $\bar{X} - \mu$ . If  $\bar{X} > \mu$ , it is positive, if  $\bar{X} < \mu$ ,  $\bar{X} - \mu$  is negative. But, modulus of  $\bar{X} - \mu$  is positive.


But, modulus of  $\bar{X} - \mu$  is positive only. So, the confidence interval length  $|\bar{X} - \mu| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ .

**(Refer Slide Time: 16:50)**

### CI Approach

- ❖ Using this we may find the 95% CI for each of the listed sample means and hence their upper and lower limits.
- ❖ Since  $\bar{X}$  is continuous there may be infinite confidence intervals.

Let us say we can draw a large number of them.



So, what we can do, we can construct an infinite number of confidence intervals. So, we can generate a large number of confidence intervals okay.

**(Refer Slide Time: 17:18)**

## CI Approach

- ❖ If the population mean is indeed 50, then among the possible CIs **generated based on randomly drawn sample means**, those confidence intervals overlapping the region between 42.65 and 57.35 will constitute 95% of the selected confidence intervals.




They will encompass the population mean  $\mu$ .

Continuing, if the population mean is indeed 50, then among the different possible confidence intervals generated based on the randomly drawn sample means, those confidence intervals overlapping the region between 42.65 and 57.35 will constitute 95% of the selected confidence intervals okay. So, we can generate infinite number of 95% confidence intervals based on the selected or chosen random samples.

So, let us say, we have a large number of confidence intervals, if the population mean was 50, I am putting a big if there, then among the so many different confidence intervals we have, the confidence intervals that overlap the region between 42.65 and 57.35 will constitute 95% of them, 95% of the generated confidence intervals. And of course, the confidence intervals overlapping the region between 42.65 and 57.35 will encompass the population mean  $\mu$  of 50.

Let us see what this means. I will put a table here.

**(Refer Slide Time: 19:11)**



Sample Mean	$ \bar{X} - \mu  = z_{\alpha/2}\sigma/\sqrt{n^{0.5}}$	UL	LL	Popln. Mean of 50 Included
40	7.35	47.35	32.65	NO
41	7.35	48.35	33.65	NO
42	7.35	49.35	34.65	NO
42.65	7.35	50	35.3	Yes
43	7.35	50.35	35.65	Yes
45	7.35	52.35	37.65	Yes
47	7.35	54.35	39.65	Yes
49	7.35	56.35	41.65	Yes
51	7.35	58.35	43.65	Yes
53	7.35	60.35	45.65	Yes
55	7.35	62.35	47.65	Yes
57	7.35	64.35	49.65	Yes
57.35	7.35	64.7	50	Yes
58	7.35	65.35	50.65	No

So, here we have different sample means. We can choose infinite number of them but, for purpose of demonstration I have chosen only a few typical sample means starting from 40 going up to 58. Then,  $\bar{X} - \mu = Z \alpha/2 \sigma / \sqrt{n}$ , which we know as 7.35. So, we can generate our confidence intervals with the upper limits given by this column and lower limits given by this column okay.

So, each represents a confidence interval and the upper limit for a sample mean or the confidence intervals upper limit based on a sample mean of 40 is 47.35 and the lower limit is 32.65 and whether the population mean of 50 is included? No. Similarly, 41 will not include the population mean of 50, because its upper limit is 48.35 and lower limit is 33.65. 42 again will be having an upper limit close to 50 but not quite 50 yet. So, 49.35 and 34.65, it is not having the population mean of 50.

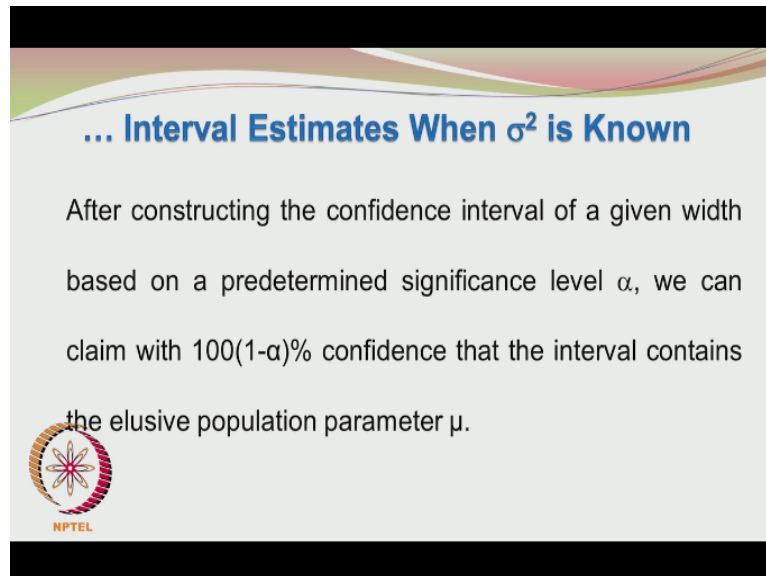
But, if you take a sample mean of 42.65, the upper limit is just touching 50 and the lower limit is 35.3 and it is including the population mean. Similarly, you go all the way up to 57.35 and the lower limit will just touch 50 and will include the population mean. But, any value  $>57.35$  will not include the population mean  $\mu$ .

So, out of so many confidence intervals, the confidence intervals which are present between 42.65 and 57.35 will encompass the population mean of 50 and the percentage of such confidence intervals falling between 42.65 and 57.35 will be 95% of all the chosen confidence intervals from the randomly chosen samples of the population. So, we have to

choose a large number of random samples from the population and calculate the random sample means. Using them, we can construct the confidence intervals.


So, we have a large number of confidence intervals and if your confidence interval specification is 95%, then 95% of the generated confidence intervals will encompass or surround the population mean  $\mu$  of 50.

**(Refer Slide Time: 22:22)**



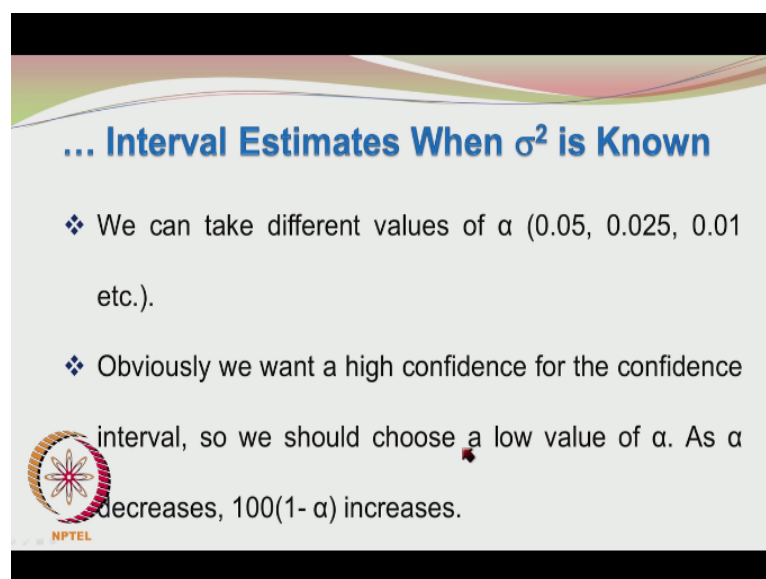
**... Interval Estimates When  $\sigma^2$  is Known**

After constructing the confidence interval of a given width based on a predetermined significance level  $\alpha$ , we can claim with  $100(1-\alpha)\%$  confidence that the interval contains the elusive population parameter  $\mu$ .




So, after constructing the confidence interval of a given width based on a pre determined value of alpha, we claim with  $100 \cdot 1 - \alpha\%$  confidence that the interval contains the elusive population parameter  $\mu$ .

**(Refer Slide Time: 22:38)**



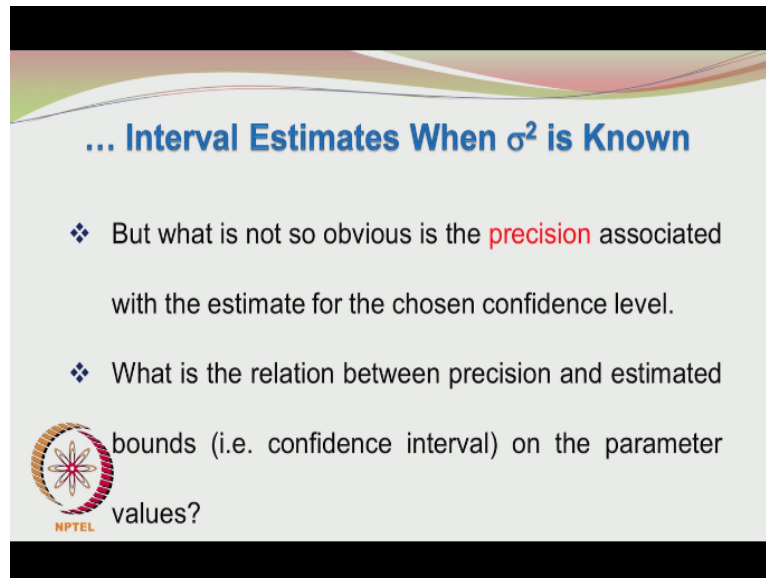
**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ We can take different values of  $\alpha$  (0.05, 0.025, 0.01 etc.).
- ❖ Obviously we want a high confidence for the confidence interval, so we should choose a low value of  $\alpha$ . As  $\alpha$  decreases,  $100(1-\alpha)$  increases.




Alpha need not be fixed at 0.05, it may also take 0.025, 0.01 etc. If alpha is 0.01, we are constructing a 99% confidence interval. So, you can see that, if the alpha of alpha decreases, our confidence level increases. When the alpha value was 0.05, we had a 95% confidence interval. If alpha is 0.01, we have a 99% confidence interval. So, when alpha value decreases, our confidence level increases. So, it looks like, we have to choose a low value of alpha. As alpha decreases,  $100 \times (1 - \alpha)$  increases.

**(Refer Slide Time: 23:38)**



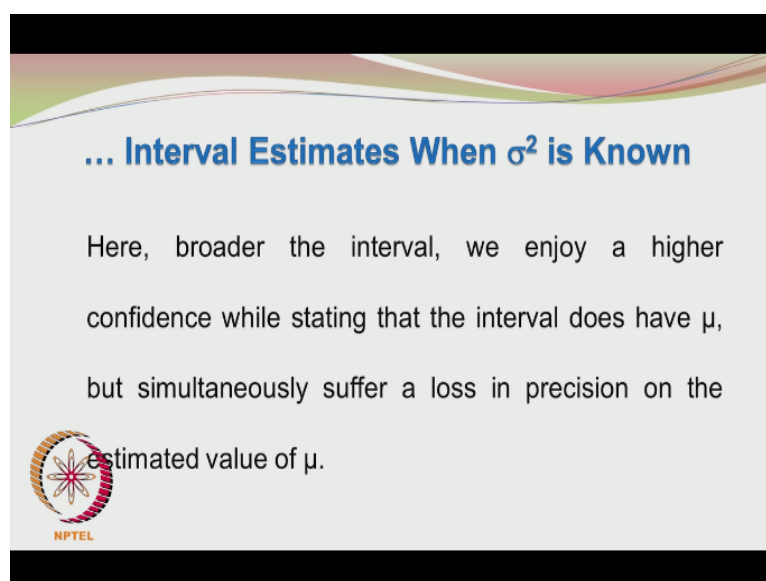
**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ But what is not so obvious is the **precision** associated with the estimate for the chosen confidence level.
- ❖ What is the relation between precision and estimated bounds (i.e. confidence interval) on the parameter values?




But, what about precision, precision means accuracy. What is the precision associated with the estimate for the chosen level of confidence? What is the relation between precision and the confidence interval? So, that is what we are going to look at now.

**(Refer Slide Time: 24:03)**



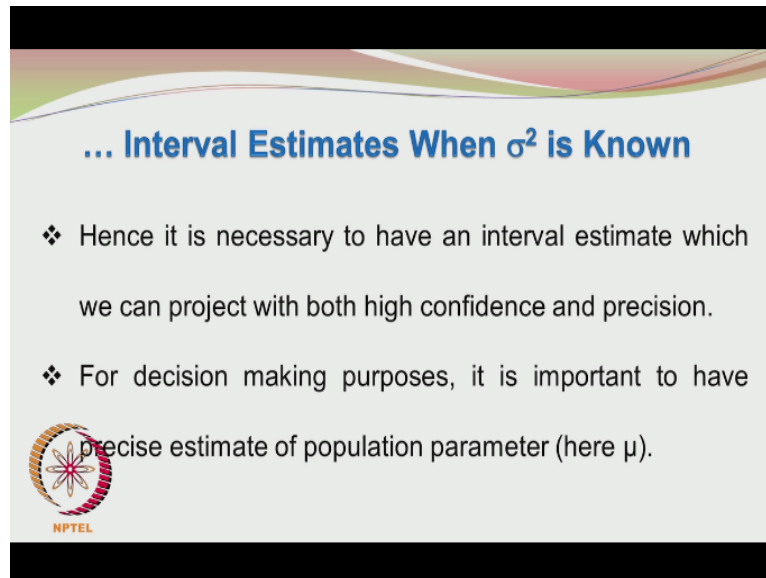
**... Interval Estimates When  $\sigma^2$  is Known**

Here, broader the interval, we enjoy a higher confidence while stating that the interval does have  $\mu$ , but simultaneously suffer a loss in precision on the estimated value of  $\mu$ .




Broader or wider the interval, we enjoy a higher confidence when stating that the interval does have the population parameter  $\mu$ . But, as the interval becomes wider, the precision suffers. Our interval estimates become less precise, when they become wider. Suppose you do experiments and very frequently you are told to put the error bars. If the variability around the experimental points are quiet high, the error bars are quiet wide then, we really do not know the precise value of the variable we are measuring.

**(Refer Slide Time: 24:57)**



**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ Hence it is necessary to have an interval estimate which we can project with both high confidence and precision.
- ❖ For decision making purposes, it is important to have precise estimate of population parameter (here  $\mu$ ).


 NPTEL

So, it is important for us to have an estimate of the population mean  $\mu$  in this case that can be cleaned with both high confidence and high precision. The interval estimate we are proposing should be having a high level of confidence and also a high level of precision. But, we saw that if we increase the level of confidence then, the precision becomes less. So, how to have both? It is important for us to have precise estimates of the population parameter so that our decisions can be made correctly.

**(Refer Slide Time: 26:00)**

**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ So how do we plan the random sampling to get the desired confidence as well as high precision?
- ❖ Let us look at the following equation



$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

How do we plan the random sampling? so that the desired confidence as well as desired precision are both achieved. So, look at the equation the basic equation probability of  $\bar{X} - Z_{\alpha/2} \sigma/\sqrt{n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \sigma/\sqrt{n}$  and that is  $1 - \alpha$ .

**(Refer Slide Time: 26:29)**

**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ We want high value of  $1 - \alpha$ , but at the same time the interval bound on  $\mu$  should be narrow.
- ❖ Lower  $\alpha$  leads to higher value of  $z_{\alpha/2}$  but causes the lower and upper bounds to drift apart, making the interval estimate less precise

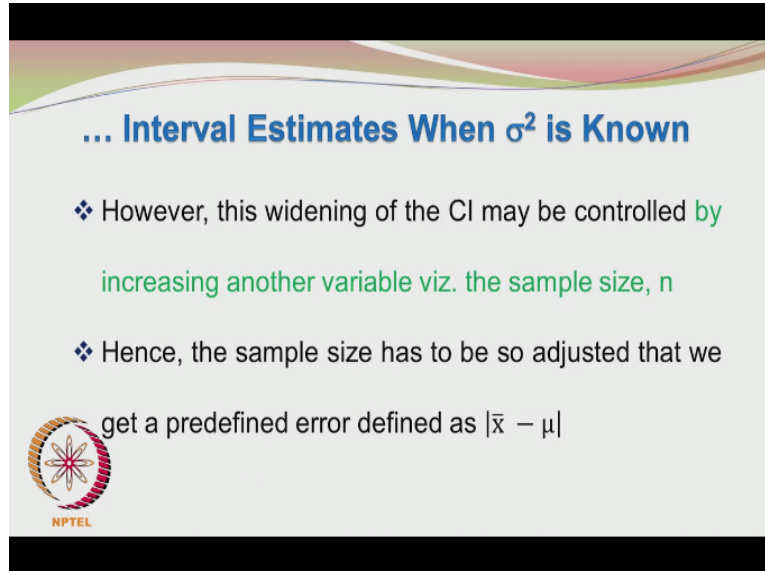
We want a high value of  $1 - \alpha$  but the same time the interval bound on  $\mu$  should be narrow. What it means is, we want to have a high value of  $1 - \alpha$  but the same time, the interval should be narrow. In other words, the lower limit should be approaching the upper limit. So that the interval is narrow and our precision improves. On other hand, we also need to have a high value of  $1 - \alpha$  so that the probability value and the confidence level increase.

If you chose a lower alpha value, it leads to a higher value of  $Z_{\alpha/2}$  and hence the lower and upper bounds will be moving away from each other. If you reduce the value of alpha, the




Z alpha/2 value based on the standard normal curve will increase. Reduce alpha, Z alpha/2 will increase. When Z alpha/2 increases, -Z alpha/2 will decrease. So, the interval between Z alpha/2 and -Z alpha/2 will widen.

**(Refer Slide Time: 28:09)**



... Interval Estimates When  $\sigma^2$  is Known

- ❖ However, this widening of the CI may be controlled by increasing another variable viz. the sample size,  $n$
- ❖ Hence, the sample size has to be so adjusted that we get a predefined error defined as  $|\bar{x} - \mu|$

 NPTEL


How to control the spreading of the interval when alpha value decreases? We have another handle here that is the sample size  $n$ , which is very well in our control. We may require a bit more investment in having a slightly larger sample size, but, that is going to pay us back in terms of increased precision. So, the sample size has to be so adjusted that we get a predefined error  $\bar{X} - \mu$ . So, if we decide how far  $\bar{X}$  should be away from  $\mu$  okay.

It does not mean that we should know  $\mu$  here. We are only telling how far  $\bar{X}$  should be away from  $\mu$  and that is a number we are going to project. Then, based on that number, we have to adjust the sample size  $n$ , such that we get both the desired level of confidence as well as the desired precision.

**(Refer Slide Time: 29:16)**

**Error E:  $|\bar{X} - \mu|$**

- ❖ The mean may be either lower or greater than  $\bar{x}$ .
- ❖ In either case we say that the error (E) should be utmost


$$|E| = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$


So, the error which is defined as  $\bar{X} - \mu$  should be of a certain limit, which should not be exceeded. And that error  $\bar{X} - \mu$  is  $= Z \alpha/2 \sigma / \text{root } n$ .

**(Refer Slide Time: 29:50)**

**... Interval Estimates When  $\sigma^2$  is Known**

- ❖ So how do we plan the random sampling to get the desired confidence as well as high precision?
- ❖ Let us look at the following equation


$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$


Again look at the fundamental expression. So, you are having this term, let us take  $\bar{X} - \mu$  here, we get  $Z \alpha/2$  okay.  $\bar{X} - \mu$  will then become  $\leq Z \alpha/2 \sigma / \text{root } n$ .  $\bar{X} - \mu$ , the absolute value of which is termed as the error E.

**(Refer Slide Time: 30:22)**

**Error :  $|\bar{X} - \mu|$**

Hence, if the error E is specified, then the sample size n that gives the desired 100(1- $\alpha$ ) confidence as well as the required ( $\leq E$ ) precision may be obtained as follows



$$n = \left[ z_{\alpha/2} \frac{\sigma}{E} \right]^2$$

So, once you have stipulated a value of E and since you already have defined error as  $Z_{\alpha/2} \sigma / \sqrt{n}$ , you square this expression, then you get E squared, then we do not have to use the modulus E squared is always positive for real values of E and so we have  $E^2 = Z_{\alpha/2}^2 \sigma^2 / n$ . So,  $n = Z_{\alpha/2}^2 \sigma^2 / E^2$ . And so, we take  $n = Z_{\alpha/2}^2 \sigma^2 / E^2$  whole squared okay.  $Z_{\alpha/2}$  is the upper  $\alpha/2\%$  point of the standard normal.

Sigma is the known standard deviation of the population and E is the stipulated error.


**(Refer Slide Time: 31:27)**

**Error :  $|\bar{X} - \mu|$**

If  $\bar{x}$  is used as an estimate of  $\mu$ , then we may be 100(1- $\alpha$ ) confident that the error given by

$$|\bar{X} - \mu|$$

will not exceed a specified amount E when the sample size is  $n = \left[ z_{\alpha/2} \frac{\sigma}{E} \right]^2$




So, if  $\bar{X}$  is used as an estimate for  $\mu$ , then we may be 100\*(1- $\alpha$ )% confident that the error given by  $\bar{X} - \mu$  will not exceed a specified amount E when the sample size is  $n = Z_{\alpha/2}^2 \sigma^2 / E^2$ . So, I will make a small correction here. If  $\bar{X}$  is used as

an estimate of  $\mu$ , then we may be  $100 \cdot (1 - \alpha)\%$  confident that the error given by  $\bar{X} - \mu$ , the absolute value of  $\bar{X} - \mu$  will not exceed a specified amount  $E$ , when the sample size is  $n = Z_{\alpha/2} \sigma / E$  whole squared.

**(Refer Slide Time: 32:19)**

**Error :  $|\bar{X} - \mu|$**

- ❖ If the value of  $n$  is not an integer, it must be rounded off to the next highest integer
- ❖ After deciding upon the sample size  $n$ , we get an interval of length such that it is twice the error (i.e.  $2E$ ).




If the value of  $n$  we compute turns out to be a non integer, it must be rounded off to the next highest integer. After deciding upon the sample size  $n$ , we get an interval of length such that it is twice the stipulated error, that is  $2E$ .

**(Refer Slide Time: 32:39)**

**Error :  $|\bar{X} - \mu|$**

Here it is assumed that

- the parent population distribution is known and it is normal
- the variance  $\sigma^2$  is known




Here, it is assumed that the parent population distribution is known and specified to be normal. The variance  $\sigma^2$  of this population distribution is also known.

**(Refer Slide Time: 32:53)**

**Error :  $|\bar{X} - \mu|$**

The sample size increases when

- ❖ the desired interval (2E) decreases, for a fixed value of  $\sigma$  and specified confidence
- ❖  $\sigma$  increases, for a fixed desired length 2E and specified confidence



The sample size required will increase when the desired interval 2E decreases for given values of sigma and the specified confidence. When the standard deviation of the population increases, that means, there is more spread in the population, then the sample size will increase for a specified value of error and a specified level of confidence okay.


So, let us say that we say that the error is so much and the confidence interval is 95%, then if the standard deviation of the population increases, there is more uncertainty there is more variability around the population mu, then we need to of course invest in a larger sample.

**(Refer Slide Time: 33:56)**

**Error :  $|\bar{X} - \mu|$**

❖ The sample size increases when

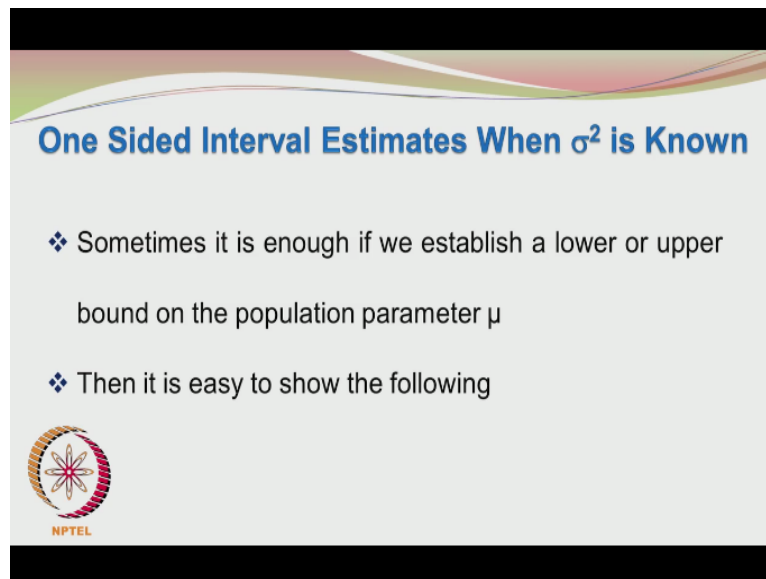
the level of the confidence increases, for fixed desired length 2E and standard deviation  $\sigma$ .



The sample size also increases when you also want a high level of confidence for a fixed desired length 2E and a fixed standard deviation sigma. So, when you want to have a higher


value of confidence level for a given precision E and a given standard deviation sigma, then you have to go for a larger sample size.

**(Refer Slide Time: 34:28)**



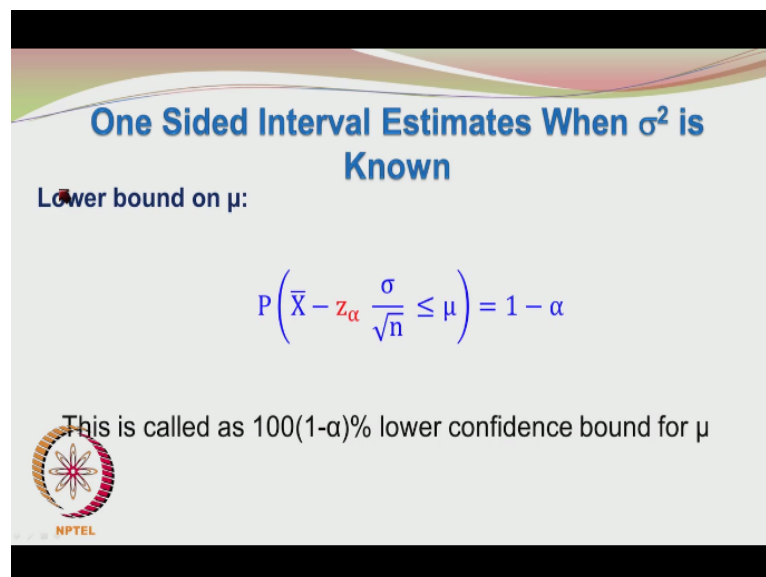
**One Sided Interval Estimates When  $\sigma^2$  is Known**

- ❖ Sometimes it is enough if we establish a lower or upper bound on the population parameter  $\mu$
- ❖ Then it is easy to show the following



In certain cases, we do not need the upper and lower bounds, it may be enough if you specify the lower bound on mu or the upper bound on mu.

**(Refer Slide Time: 34:43)**




**One Sided Interval Estimates When  $\sigma^2$  is Known**

Lower bound on  $\mu$ :

$$P\left(\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu\right) = 1 - \alpha$$

This is called as 100(1- $\alpha$ )% lower confidence bound for  $\mu$



Lower bound on mu. Probability of  $\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu = 1 - \alpha$ . Please note that, I am using Z alpha and not Z alpha/2 as I was using earlier. That is because I am now constructing a lower bound on mu. So, this is termed as a 100\*1-alpha% lower confidence bound for mu.


**(Refer Slide Time: 35:17)**

## One Sided Interval Estimates When $\sigma^2$ is Known

Upper bound on  $\mu$ :

$$P\left(\mu \leq \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

This is called as 100(1- $\alpha$ )% upper confidence bound for  $\mu$




Similarly, the upper bound on mu may be defined as probability of  $\mu \leq \bar{X} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} = 1 - \alpha$ . This is termed as 100\*1-alpha% upper confidence bound for mu. So, we are chopping of one side of mu and using the other side only and when we do that, we use alpha and not alpha/2 when finding out the upper % points.

**(Refer Slide Time: 35:52)**

## One Sided Interval Estimates When $\sigma^2$ is Known

Lower bound on  $\mu$ :  $\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu$

Upper bound on  $\mu$ :  $\mu \leq \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$




So, the lower bound on mu may be written as  $\bar{X} - Z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu$ . Upper bound on mu may be written as  $\mu \leq \bar{X} + Z_{\alpha} \frac{\sigma}{\sqrt{n}}$ .

**(Refer Slide Time: 36:08)**

## Large Sample Confidence Interval for $\mu$

- ❖ Realistically, we do not know the value of population variance  $\sigma^2$
- ❖ However, this problem is again attenuated if we take a

 large sample

NPTEL

So, let us now come to the actual situation where the population variance is not known okay. But, let say that we have chosen a large sample. How large is large? In our particular case, let us say that the sample size is  $>40$ . So, the problem of unknown population variance sigma squared is mitigated or attenuated if you take a large sample. What are the advantages of taking a large sample? Then we can relax the assumption that the parent population is normal.

The parent population can be anything, it can be normal, it need not be normal. But, our sample size is quiet large, then the central limit theorem helps us by saying that, the sampling distribution of the means is approximately normal irrespective of the shape of the parent population distribution. This is very good. So, for a large sample size, the sampling distribution of the means is normal.


**(Refer Slide Time: 37:32)**

## Large Sample Confidence Interval for $\mu$

In such cases the standard normal variable

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

may be used to find the desired probability values.

 However, we may not know the value of  $\sigma$ .

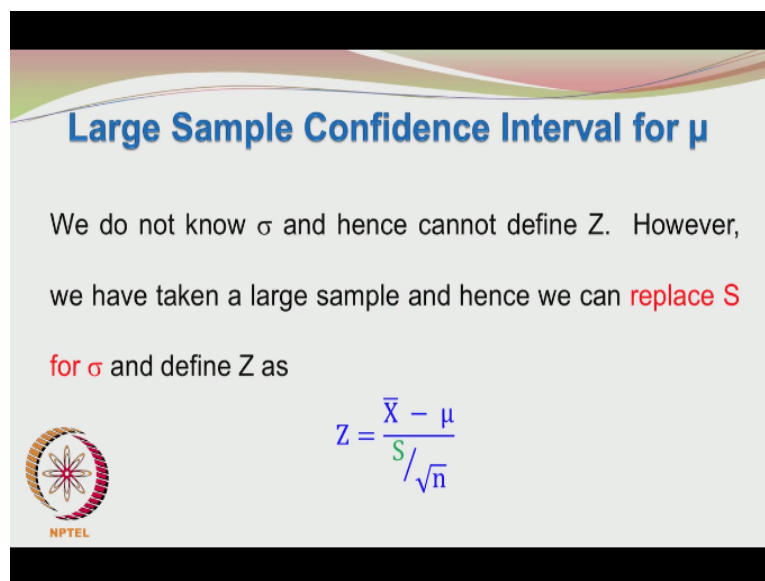
NPTEL



And previously we had used  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ . But now, we do not know the value of  $\sigma$ . What is to be done? We are having a large sample and the sampling distribution of the means is normal. So, we can normalize it and convert it into a standard normal variable  $Z$ . but, we do not know  $\sigma$ . What is to be done, what do we know? We have the sample with us. With the sample, we have the sample mean, we also have the sample standard deviation  $S$ .


So, instead of  $\sigma$ , we can put  $S$  here. And the resulting distribution will still be approximately normal. So, we can put  $S$  instead of  $\sigma$  and life will be nearly as usual as before.

**(Refer Slide Time: 38:28)**



**Large Sample Confidence Interval for  $\mu$**

We do not know  $\sigma$  and hence cannot define  $Z$ . However, we have taken a large sample and hence we can **replace  $S$**  for  $\sigma$  and define  $Z$  as


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


So, I am replacing  $S$  instead of  $\sigma$  here and so the distribution importantly is still approximately normal. So, I am able to use a standard normal variable  $Z$  and that is  $= \frac{\bar{X} - \mu}{S/\sqrt{n}}$ .

**(Refer Slide Time: 38:49)**

## Large Sample Confidence Interval for $\mu$

- ❖ Reiterating, we are now considering a large sample from a population that may be or may not be normal.
- ❖ Since the sample size is large, the condition of a **normal original population is NOT required**.




So, we are taking a large sample from a population that may or may not be normal. Since the sample size is large, the condition of a normal original population is not required.

**(Refer Slide Time: 39:06)**

## Large Sample Confidence Interval for $\mu$

According to the central limit theorem, the sample distribution is nearly normal with mean  $\mu$  and variance  $\sigma^2/n$  provided  $n$  is large.



According to the central limit theorem, the sampling distribution of the means is nearly normal or approximately normal with mean  $\mu$  and variance  $\sigma^2/n$  provided  $n$  is large.

**(Refer Slide Time: 39:20)**

## Large Sample Confidence Interval for $\mu$

Since the variance  $\sigma^2$  is unknown we assume that the sample distribution is nearly normally distributed with mean  $\mu$  and variance  $S^2/n$ .



Since the variance sigma squared is unknown, we assume that the sample distribution is nearly normally distributed with mean mu and variance S squared/n. We were able to make this assumption because of the large sample size. So, we can now appreciate the merits in investing more resources for taking a large sample okay. So, we need not take the entire population into consideration during the sampling exercise.

That is not going to be realistic but we can invest on a large sample size. So, a large sample size helps us to increase the precision of our confidence interval and it also helps us to handle situations, where the parent population is not normal. It helps us to handle situations, when the population variance is not known okay. So, when you have a large sample size and the parent population is not normal or the parent population distribution is unknown, the sampling distribution of the means become normally distributed.

If sigma squared is not known, then we have to use S squared. If the population distribution shape is not known, if the variance sigma squared is not known and we have a large sample size, we can do the following, we can still assume approximately that the sampling distribution of the means is normal. Number two: instead of sigma squared, which is not known, we can use S squared. Here S squared is the sample variance.


**(Refer Slide Time: 41:36)**

## Two Sided Interval Estimates When $\sigma^2$ is Unknown

The large sample confidence interval is now defined for  $\mu$

$$P\left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Hence we get the  $100(1-\alpha)\%$  confidence interval on  $\mu$  as




$$\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}$$

So, the large sample confidence interval is now defined for  $\mu$  as probability of  $\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} = 1 - \alpha$  and the  $100 \cdot (1 - \alpha)\%$  confidence interval on  $\mu$  is given as  $\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}$ .

**(Refer Slide Time: 42:13)**

## One Sided Interval Estimates When $\sigma^2$ is Unknown

Similarly the upper and lower bounds may be defined as shown above, taking care to replace  $\sigma$  for  $s$ .



Similarly, the upper and lower bounds may be defined taking care to replace  $\sigma$  with  $s$ , the sample standard deviation. The sample size should be preferably 40 or more. We earlier saw that for the central limit theorem to apply, we need a sample size  $>30$ . But, since we do not know  $\sigma$ , there is additional variability and hence we have to go for a larger sample size. So, this completes our discussion on confidence intervals.

This is very, very important because when you look at any statistical software output, you usually find the 95% confidence interval limits presented. The confidence intervals also have an important and interesting property. If the lower limit of the confidence interval is negative and the upper limit of the confidence interval is positive then, it is giving us some important information okay. For example, it is like a person when queried, at what time the train is going to reach the station.

He says, oh the train has left 5 minutes back or it is expected in 5 minutes' time. The person who is listening to this will get completely confused. Is the train going to come or has it already gone? So, if the lower limit is negative and the upper limit is positive okay then, the 0 value is bounded between the upper and lower limit. So, the population parameter may be 0. It has certain implications in this design of experiments and in linear regression analysis.

We may construct the confidence intervals on the model parameters and we may see that the confidence bounds may be passing through negative as well as positive values. They have special significance attached to them. And these are also very helpful to interpret the statistical estimates of these parameters. So, I will wind up the discussion now. So, I request you to think about what I have said and please remember that the confidence interval is probably bounding the value of  $\mu$  with the stated level of confidence okay. Thank you.