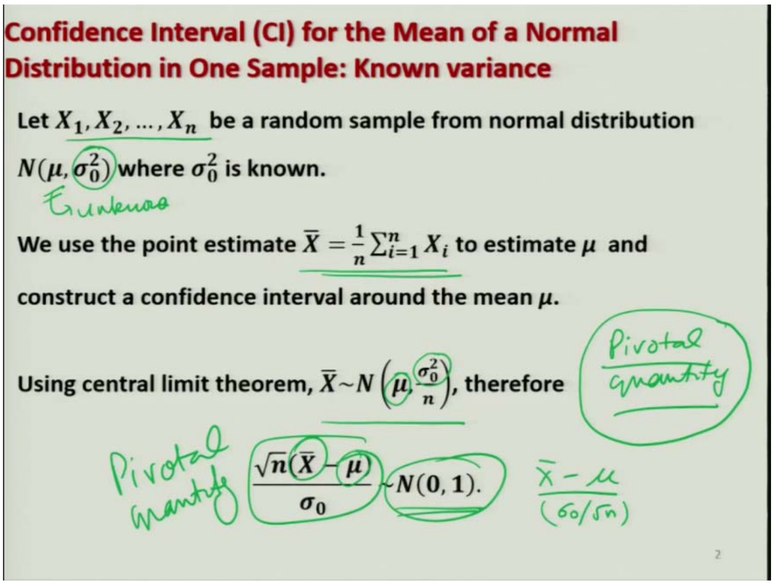
Essentials of Data Science with R Software – 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 63
Confidence Interval for Mean in One Sample with Known Variance

Hello friends welcome to the course Essential of Data Science with R Software 1 in which we are trying to understand the basic concepts of probability theory and statistical inference. So, you can recall that in the last lecture, we had initiated a discussion on the Confidence Interval Estimation and I had simply given you the basic concepts, basic definitions, fundamentals. And now, in this lecture, we are going to take one example.

The example is that we are going to consider a sample from our normal population and we will try to obtain the confidence interval for the parameter μ. Now, for σ^2 you have two options, whether σ^2 is known or σ^2 is unknown. So, in this lecture, we are going to consider σ^2 to be known. And in the next lecture, I will try to consider the σ^2 to be unknown.

So, now, in this lecture whatever we had learned in the last lecture that we are going to implement it. So, what you have to do that you simply have to follow the methodology that how are we going to obtain the confidence interval? What are the steps involved? Once you understand these steps, then after that finding out the confidence interval for most of the probability distributions or any parameter of those probability distribution will not be difficult that will be pretty straightforward. So, my request you in this lecture is that that please try to follow the steps. So, let us begin our lecture.

(Refer Slide Time: 02:05)



So, now, we are considering the situation that we have got a random sample from a normal population sample is $X_1, X_2, \ldots, X_n$ and this normal population the mean here is μ, which is unknown and σ_0^2 is the variance which is known. Now, you have learned how are you going to estimate the parameter μ through the point estimation and we had obtained that the point estimate of μ was $\bar{X}$.

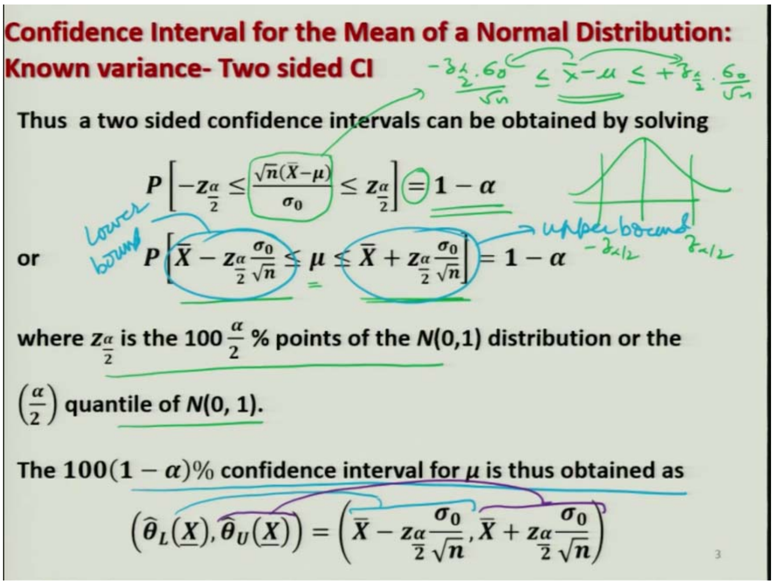
Either in the case of method of moments or the maximum likelihood estimation. And now, we want to construct the confidence interval around the mean μ. So, what we try to do here now, you have to first understand that you need to somehow find the pivotal quantity first. Once you can obtain this thing, your job is nearly done after that there is only simple algebra that you have to follow in a very systematic way.

So, the main concern lies that somehow under the given circumstances, you have to find a pivotal quantity, which is depending on the values of $X_1, X_2, \ldots, X_n$ the parameters of the distribution, but its probability distribution is independent of any of the parameter. So, now, you can recall the central limit theorem, we had proved there, that $\bar{X}$ follows our normal distribution with the same parameter μ same mean μ and variance σ^2/n.

So, now, here the variance is σ_0^2. So, the variance becomes here σ_0^2/n which is a known quantity and we also had learned that the $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} \sim N(0,1)$. So, this is essentially $\bar{X}$ - μ upon the standard deviation of $\bar{X}$ which is $\sigma_0/\sqrt{n}$. So, we know that this is going to follow our N(0, 1) and you can now see that this is your pivotal quantity.

This itself depend on $X_1, X_2, \ldots, X_n$ and μ, but its a distribution is depending on N(0, 1) which does not involve any unknown parameter μ.

(Refer Slide Time: 04:48)



So, now we have to just construct the confidence interval and now you just try to observe that how simple and easy it is. So, we know what we need to find. Now, using this pivotal quantity whose distribution is N(0, 1) it will be like this one. So, we need to find that this quantity is $\sqrt{n}\,(\bar{X} - \mu)/\sigma_0$ should lie between $-Z_{\alpha/2}$ and $+Z_{\alpha/2}$ like this $-\alpha/2$ and $+\alpha/2$ all the symbols and notation.

Now, they are known to you. And the probability of such an event is exactly equal to $1-\alpha$. You see, when you are trying to use the continuous distribution in which you are trying to find out the confidence interval we can very easily assume the equality sign, but in case if you are trying to construct it for the discrete random variable, then you have to be careful. At any rate, let us try to concentrate on this issue.

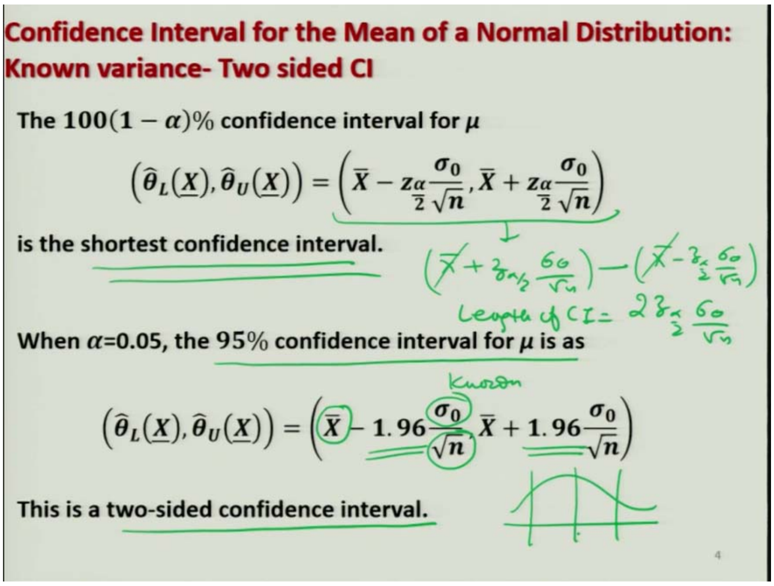
Now, you can see here that you simply have to solve this equation. So, if you try to solve equation here this will become here $-Z_{\alpha/2}$ into $\sigma_0/\sqrt{n}$ which is lying between $\bar{X} - \mu$ and this will be $+Z_{\alpha/2}$ into $\sigma_0/\sqrt{n}$ and you want to know what is the range for μ.

So, now, you have to simply solve this thing this $\bar{X}$ will go on the left hand and right hand side and this will become that μ lies between $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ and you know that this is α/2 is the 100 α/2% point on the N(0, 1) distribution or the or the α/2 quantiles of N(0, 1).

So, now, I can see here that this is the value of lower and upper limits of μ. So, this is the lower bound and this is the upper bound. So, I can say now, here the 100(1 – α)% confidence interval for μ is like this the $\hat{\theta}_L(\underline{X})$ in your earlier notation is given by like this $\bar{X} - Z_{\alpha/2}\ \sigma_0/\sqrt{n}$ and $\hat{\theta}_U(\underline{X})$ this will become $\bar{X} + Z_{\alpha/2}\ \sigma_0/\sqrt{n}$.

So, now you have obtained and you can see that these two limits they are depending on the value of capital $\bar{X}$ and that is what we said that, this is the statistic. These are two statistic θ, $\hat{\theta}_U$ and $\hat{\theta}_L$ because they are going to be random.

(Refer Slide Time: 08:16)



And in case if you try to prove that whether this interval is good or bad, then we can prove statistically also that the length of the interval is the shortest. So, the length of this interval is found as $\bar{X} + Z_{\alpha/2}\ \sigma_0/\sqrt{n} - \bar{X} - Z_{\alpha/2}\ \sigma_0/\sqrt{n}$. So, this this $\bar{X}$ gets cancelled out and the length of the interval here is simply here $2\ Z_{\alpha/2}\ \sigma_0/\sqrt{n}$.

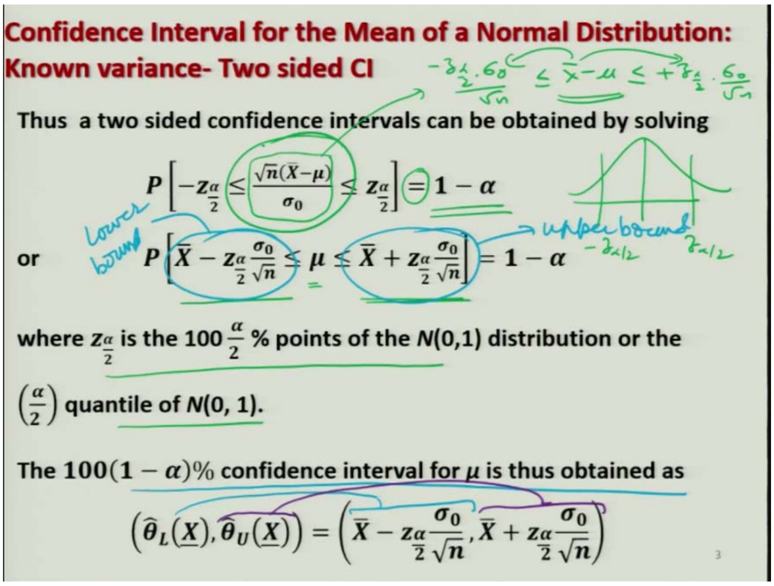
So, this is the length of a confidence interval. Briefly we will write many times confidence interval. So, this interval that you have obtained here this is the shortest confidence interval.

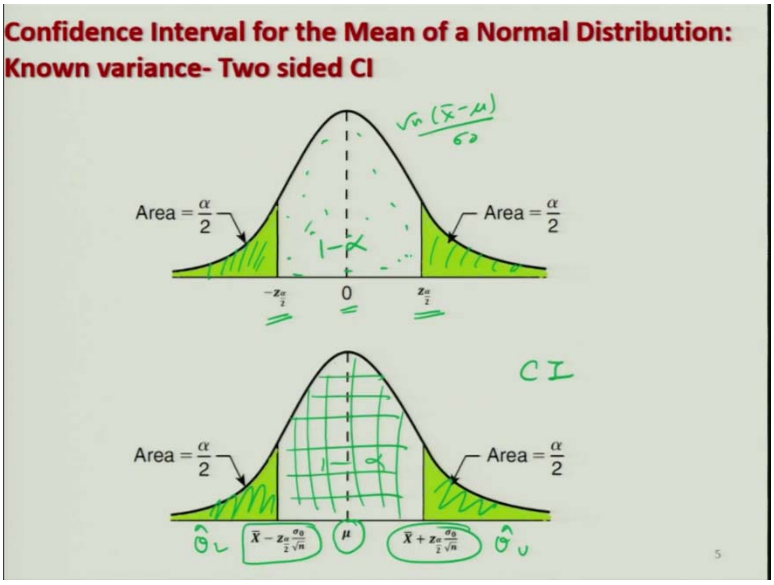
So, for example, if you try to take α is equal to 0.5 then this interval will be called as 95% confidence interval for μ.

And in case if you try to see the value of $Z_{\alpha/2}$ for α is equal to 0.05 this will come out to be 1.96 from the table, I will try to show you that if we can obtain the same value from the R software also, that we actually have now learned because this is simply your here quantiles and the value of $\bar{X}$ can be obtained from the sample is small n is the value of the sample size, σ_0 is known.

So now for a given sample $X_1, X_2, \ldots, X_n$ you can find out the value of lower and upper limits of the confidence interval and you can see here this is a two sided confidence interval and in case if you try to plot it also this will look like this. That there are two levels and somewhere it is in the, there is the parameter.

(Refer Slide Time: 10:19)

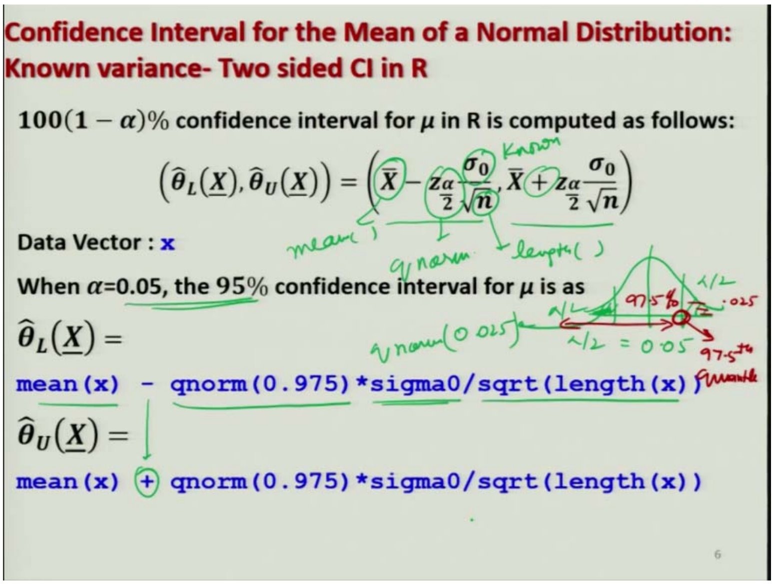




For example, like this you can see here that in case if you try to see the curve for the contribution of your pivotal quantity that we try to show you here this quantity. So, this could, this is the distribution of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$. So, you can see here it will look like this that the mean here is 0 and here this is $-Z_{\alpha/2}$ and here on the right hand side this is $+Z_{\alpha/2}$ and this area which is here shaded in green color, this is actually $\alpha/2$.

So, obviously, this area which is here in the dotted, this is $1-\alpha$. And when you are trying to find out the confidence interval then it will look like this that μ here is here and on the left hand side of the μ, we have here the value of $\hat{\theta}_L$ and on the right hand side of the μ we have $\hat{\theta}_U$ which are given by $\bar{X} \pm Z_{\alpha/2} \ \sigma_0 / \sqrt{n}$ and you can see here this shaded area here and here this is $\alpha/2$ and this area here in the middle, this is $1-\alpha$. So, this is what we mean that this is the area which is the confidence band.

(Refer Slide Time: 11:52)



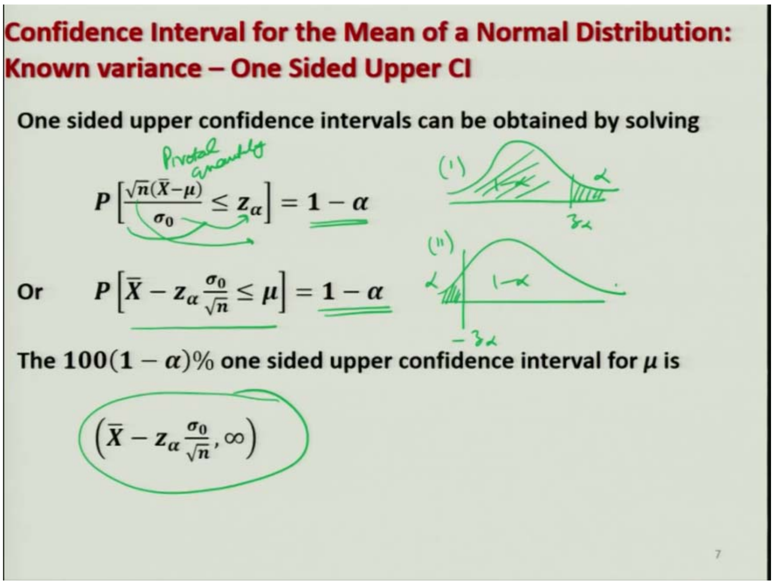
Now, in case if you try to see if you want to compute this interval in the R software, well, there is no direct command in this case, but when we are trying to do the case when σ^2 is unknown to us, then we have our direct command and we can obtain the confidence interval directly. So, if you try to see the structure of this interval, the lower limit and upper limit then means, you know how to compute each of the factor, each of the term.

For example, this can be obtained by see here mean and this $Z_{\alpha/2}$ can be obtained by here q norm that you have used when you did the normal distribution and this here n, n is the number of observations. So, this can be obtained as a length of the data vector and σ_0 is known. So, we can very easily write one line a command for obtaining the lower limit and similarly, just by changing the – sign to + sign, we can write a one-line command for the upper limit also.

So, for example, if I say that α is equal to 0.05 so, you want here 95% confidence interval. So, this area here is α/2, this area here is α/2 So, α/2 is going to be here 0.05. So, this area here is going to be 0.025 quantiles. So, it can be obtained by here q norm 0.025 and guess point here, this is this area here is 0.025. So, this point is actually here, this side, this is the point here, which is 97.5% of the area.

So, this is here 97.5th quantiles. So, you can obtain any of this value and then you have to use it appropriately. So, now, I can write down the lower limit of the confidence interval as a mean of X, suppose X is that data vector, then qnorm 0.975 into σ_0 which is given to us some known value and a square root of length of X data vector. And similarly, if you try to change the sign from – to + you can obtain here the upper limit of the confidence interval and I will try to show you all these things as an example also. So, you can see here this is essentially a two-sided confidence interval.

(Refer Slide Time: 14:37)

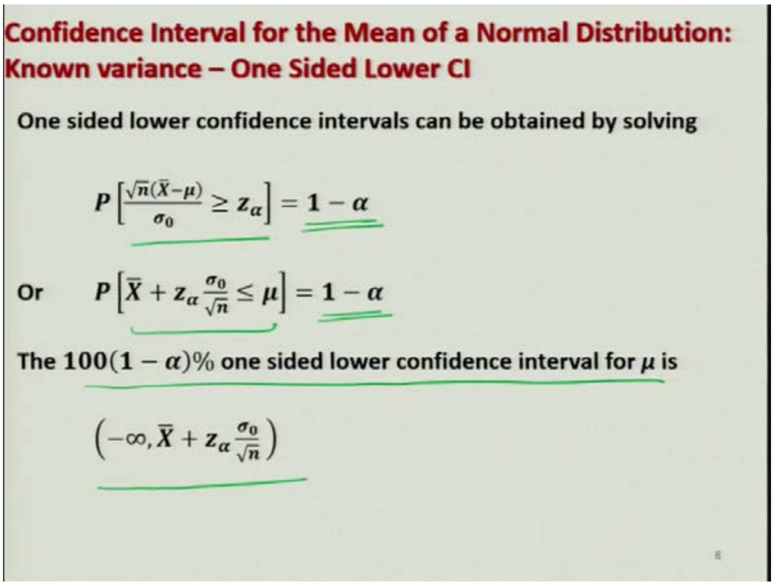


And similarly, you can also find one sided confidence interval. For one sided what will happen? Like this, there are two options that you this α is on the right hand side or on the right tail of the distribution or α is on the left-hand side or the left tail of the distribution and the remaining area will be $1 - \alpha$ in both cases. So, in the case number one, this point is going to be here is Z α and in the case number here to this is going to be – Z of α.

So, now, we can define the one sided upper confidence interval, which is obtained by this quantity, pivotal quantity is $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ is less than equal to Z_α is equal to $1 - \alpha$. That the probability that this area is equal to $1 - \alpha$ and if you try to solve it, you can just put this σ_0 on the right hand side bring a $\sqrt{n}$.

We are here simply solve it and then you will get here that $P\left[\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0} \le z_\alpha\right] = 1-\alpha$. So, I can say that $100(1 - \alpha)\%$ one sided upper confidence interval for μ is like this $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$, as simple as that.

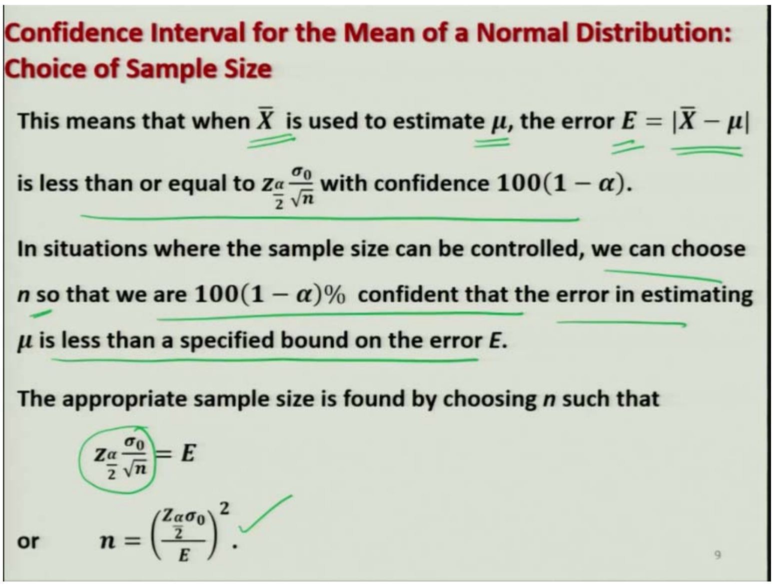
(Refer Slide Time: 16:22)



And similarly, if you want to find out the one-sided lower confidence interval for μ this can be obtained by solving that $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma_0}$ is greater than or equal to Z_α and the probability of such an event is $1 - \alpha$ and if you simply solve it, this will solve that μ is greater than or equal to $\bar{X} + Z_\alpha \sigma_0 / \sqrt{n}$ and the probability of such an event is $1 - \alpha$.

So, I can say that the $100(1 - \alpha)\%$ one sided lower confidence interval for μ is $-\infty$ to $\bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}}$.

(Refer Slide Time: 17:04)



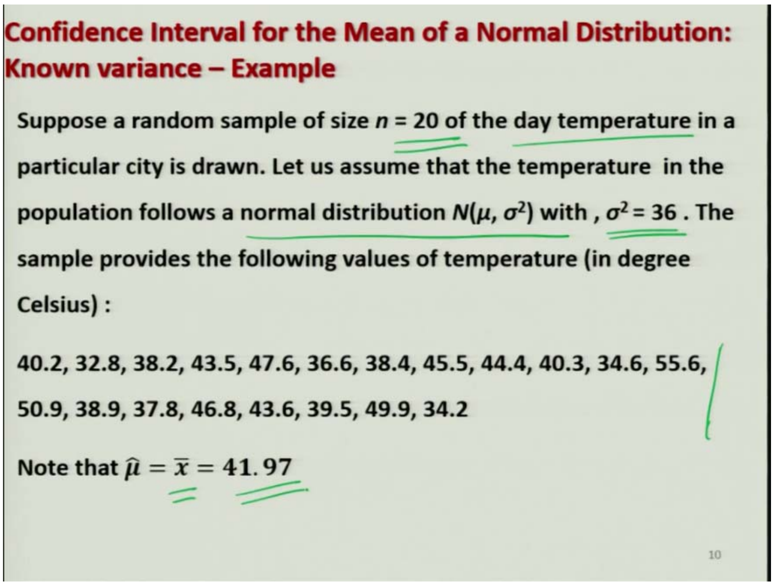
So, now, you have obtained all these things and I will just try to show you here that this concept of confidence interval also helps in finding out the value of sample sizes and it helps us in choosing the sample size. So, how that is a very popular utility of this confidence interval. So, here if you try to see we are essentially using this $\bar{X}$ to estimate the unknown parameters μ.

So, the error what is happening that some values are lower than μ and some values will be upper than μ. So, I can write down this deviation or this error as capital E which is the absolute value of the division between $\bar{X}$ and μ. So, we can make our condition, we can impose a condition that this error is less than or equal to $Z_{\alpha/2}$ σ_0 divided by root n with confidence 100(1 – α).

So, in this type of situation where the sample size can be controlled, so, that we can choose the sample size is small and so, that we are 100(1 – α)% confident that the error in estimating the parameter μ is less than a specified bound on the error. Show that you are sure that you are not making the error more than this for in this case.

What we can do this $Z_{\alpha/2}$ $\sigma_0 / \sqrt{n}$ can be substituted as capital E and this can be solved and we get here the value of a small n to be $Z_{\alpha/2}$ σ_0 divided by E whole square and this is the appropriate sample size.

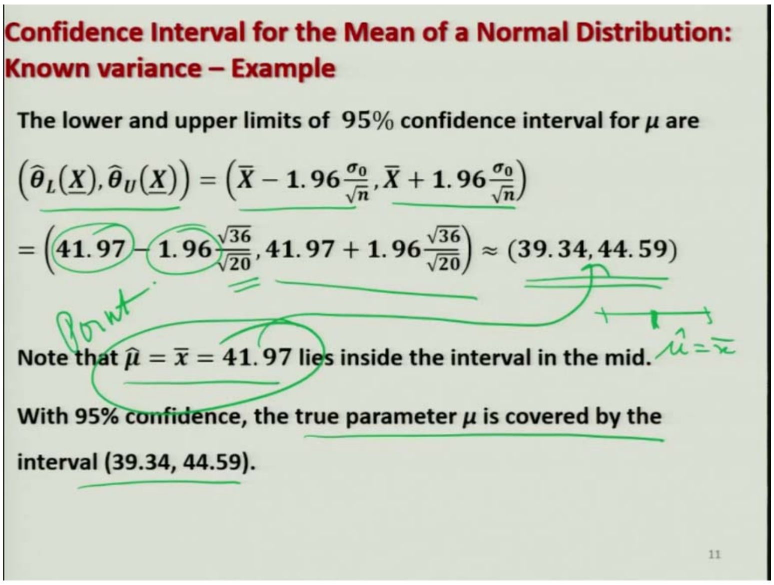
(Refer Slide Time: 18:54)



So, so, now, that we try to take an example and try to show you all these things. So, suppose a sample of size 20 on the day temperature in a particular city is collected. So, temperature of 20 days in a city that is collected and we assume that this temperature is following a normal distribution μ σ^2 and suppose σ^2 is known too. Now, the sample values of the temperature are obtained here like this in degrees Celsius.

Now, that we know that if you want to find out the estimate of μ as a point estimate, then we can compute the sample mean and if you try to compute the sample mean of all this value, this will come out to be 41.97 degree Celsius.

(Refer Slide Time: 19:45)

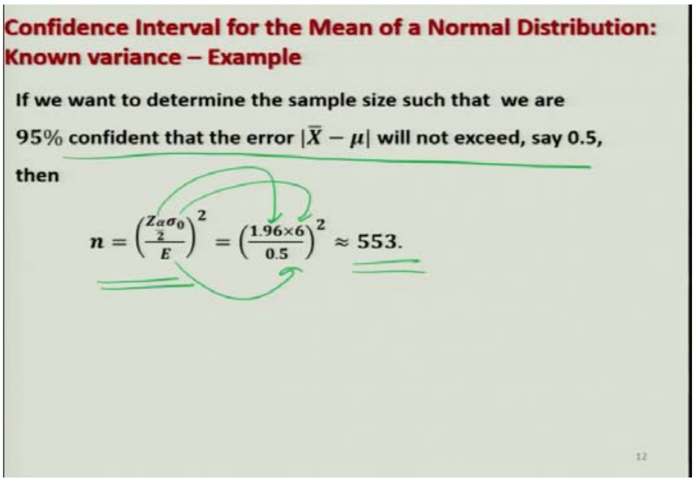


But now, we are interested in finding out the interval estimate of μ. So, we try to use the earlier obtained expression and we try to obtain the 95% confidence interval for μ whose lower and upper limits are obtained has like their $\hat{\theta}_L$ and $\hat{\theta}_U$ based on $X_1, X_2, \ldots, X_n$ which are operated by these two limits.

So, now we can obtain the value of sample mean it is here 41.97, the value of $Z_{\alpha/2}$ from the table can be obtained as 1.96 or that can be obtained directly from the R software also and then σ_0 is the square root of 36 and small n is 20 and similarly, you can write down here the upper limit also and you can solve it and this will come out to be approximately 39.34 to 44.59 that means the μ is expected to lie in the range 39 degrees to so, almost 44 degrees.

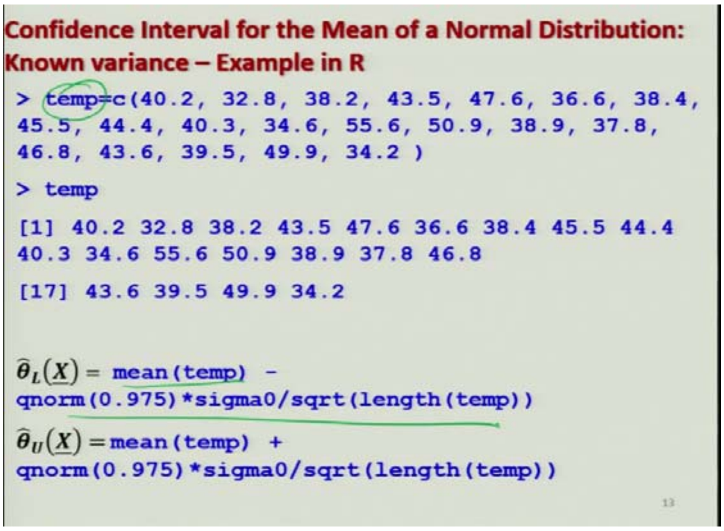
And you can see here that this sample mean $\overline{X}$ which is which was the point estimate this lying inside this interval and it is lying in the mid means, if you try to take this value somewhere in the middle this will be your here the value of $\hat{\mu}$ is equal to $\overline{X}$. So, now, we can say with 95% confidence that true parameter μ is covered by the interval 39.34 to 44.59 this is the interval estimate. And whereas this thing this is the point estimate that we have to just keep in mind. So, you can see it is not difficult to obtain them.

(Refer Slide Time: 21:25)



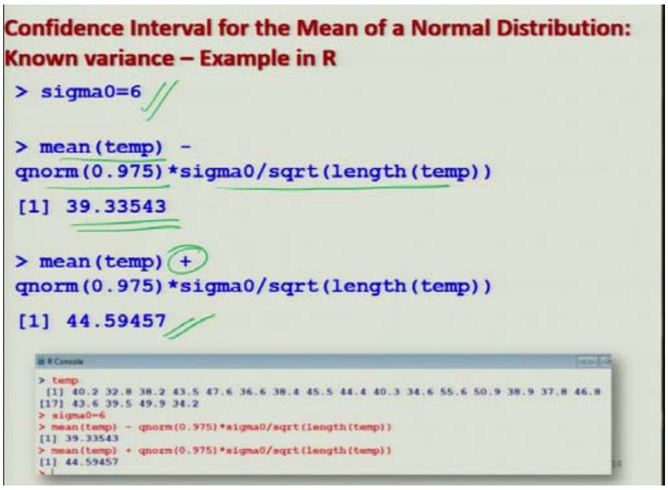
And in case if you want to determine the value of n that is the sample size says that we are 95% confident that error $\overline{X}$ - μ that is the absolute value of error will not exceed is say 0.5 then I can use this expression of n and I can simply substitute here the value of $Z_{\alpha/2}$ as 1.96 σ_0 as 6,an error here as say 0.5 and we can often hear that the approximate sample size is going to be 553. So, that means you need the 553 number of observations.

(Refer slide Time: 22:04)



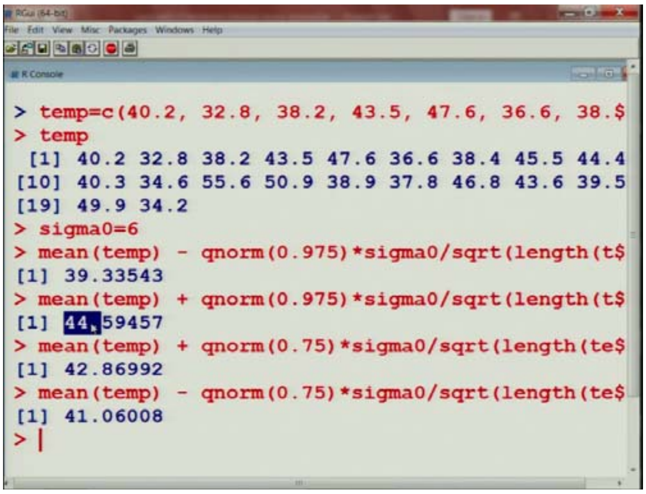
So, now that we try to do the same thing in the R console also so I have just entered this data in that data vector see here temp, and then I am simply trying to compute the $\hat{\theta}_L$ and $\hat{\theta}_U$ by these two expressions.

(Refer slide Time: 22:23)



And you can see here that these value come out to me like that mean says σ_0 is given. So, I have to enter it separately and then the lower limit will come out to be here like this mean temperature – qnorm 0.975 and σ_0^2 root of length and temperature this really will come out with 39.33 and similarly, you can obtain the upper limit just by changing the – sign to + sign and you can obtain here these two limits 39.33 through 44.59.

(Refer Slide Time: 22:53)



So, now, I have entered the data on here temperature as temp data vector you can see here I need to give here the value of σ_0 as 4 you can see here like this and if I try to compute here the lower confidence interval this will come out to be like this the same expression that you have just used and if I try to just change this – sign to + sign this will be upper confidence limit.

And surely, if you try to change here your this α that means suppose if I take here only the seventy fifth quantiles, you will see here this value is going to decrease. So, it depends on the value of α that how are you going to obtain the confidence interval for example, if you try to take the seventy fifth quantiles then it is going to be the interval is going to be shorter the length is not between 41 to 42.

Whereas in the earlier case, the length here is 39 to 44. So, now, let me come to an end to this lecture, I have given you the details that how are you going to construct the confidence interval in a normal distribution for the mean when the variants is known to us. Now, in practice this variants may not be known to us in that situation.

What we are going to do, we will simply try to estimate it, replace it there by the question come how are you going to find out the pivotal value. So, that I will try to show you in the next lecture, but it is important for you in this lecture that you try to do the same algebra with your own hand manually and try to understand how can you obtain the confidence interval. This methodology you will see that it is applicable in most of the cases whenever you want to find out the confidence intervals for different parameters in different probability functions. So, you try to practice it and I will see you in the next lecture till then, goodbye.