Essentials of Data Science with R Software-1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture 64

Confidence Interval for Mean and Variance

Hello friends. Welcome to the course essentials of data science with R software-1 in which we are trying to understand the basic concepts of probability theory and statistical inference but now we are in the part of statistical inference. So, you can recall that in the last lecture, we constructed the confidence interval for the mean under the $N(\mu, \sigma^2)$ population, when σ^2 was known to us and I had given you the all the steps in detail that how are we trying to construct the confidence interval.

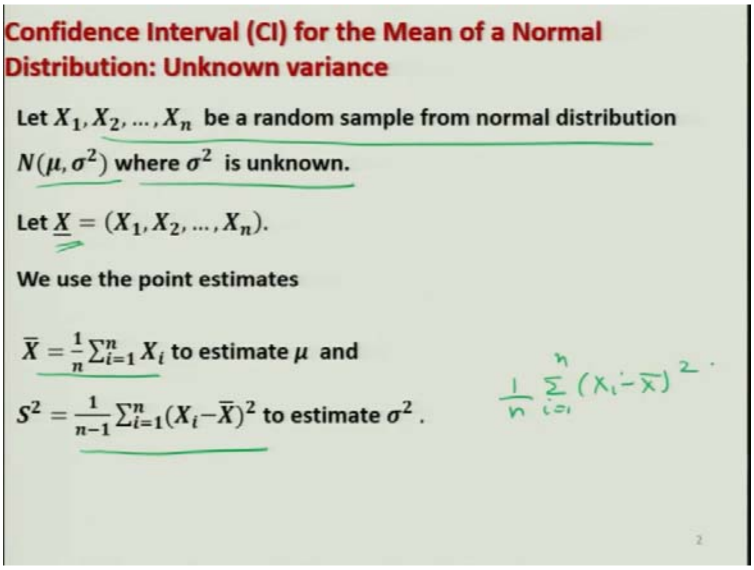
Now, in this lecture I will try to take the second case when σ^2 is unknown to us, actually that will be a more realistic situation in most of the real-life cases that you are trying to estimate a parameter and the parameters are actually unknown to us. So, if there are two parameters both the parameters are usually going to be unknown to us. So, now we want to estimate the confidence interval for mu, when σ^2 is unknown to us.

So, one simple option is this that we can estimate the σ^2 and we can replace it in the place of σ^2 and based on that we have to construct a pivotal quantity which is depending on the parameter and the random variable but the distribution of this quantity is independent of the parameter.

So, now the question is how are we going to do it? Well do you remember that sampling distributions? And do you remember that we had done three sampling distribution Chi-square, t and F. Now, in this lecture I am going to employ the properties of t distribution to construct the confidence interval for μ when σ^2 is unknown. Suppose you have a sample from $N(\mu, \sigma^2)$, now you are trying to estimate the σ^2 also.

So, do not you think that we should also construct a confidence interval for the σ^2 ? Yes we will do it in this lecture and for that we will try to use the properties of Chi-square distribution. So, let us begin our lecture and try to find out the confidence interval but remember one thing the step, the procedure, the methodologies they are going to be the same as we have done in the last lecture. So, I will not be repeating the same steps again and again in this lecture, if you have any doubt or problem my sincere suggestion is that you please try to revise the earlier lecture first and then you watch this lecture. So, we begin our lecture.

(Refer Slide Time: 03:18)



Confidence Interval (CI) for the Mean of a Normal Distribution: Unknown variance

Let $X_1, X_2, \dots, X_n$ be a random sample from normal distribution $N(\mu, \sigma^2)$ where σ^2 is unknown.

Let $\underline{X} = (X_1, X_2, \dots, X_n)$.

We use the point estimates

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ and

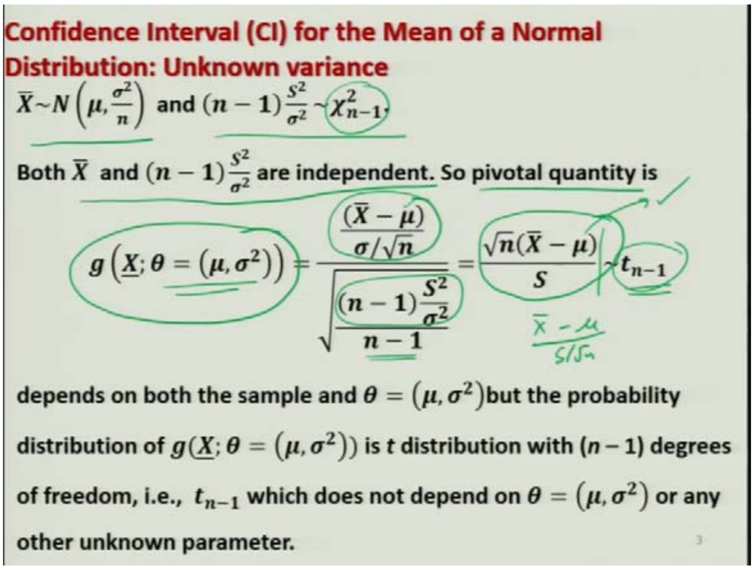
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to estimate σ^2.

$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$

2

So, now, let $X_1, X_2, \dots, X_n$ be a random sample from normal distribution, $N(\mu, \sigma^2)$ where σ^2 is now unknown to us. So, let this $X_1, X_2, \dots, X_n$ be indicated by $\underline{X}$ just for the sake of simplicity and now you know that how to estimate the parameters μ and σ^2 as a point estimates. So, the μ can be estimated by the sample mean that is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the variance can be estimated by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and means if you remember other form is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. So, these are the point estimates.

(Refer Slide Time: 04:10)



Confidence Interval (CI) for the Mean of a Normal Distribution: Unknown variance

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$

Both $\bar{X}$ and $(n-1)\frac{S^2}{\sigma^2}$ are independent. So pivotal quantity is

$$g\left(\underline{X};\theta=(\mu,\sigma^2)\right) = \frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\frac{S^2}{\sigma^2}}{n-1}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$$

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

depends on both the sample and $\theta = (\mu,\sigma^2)$ but the probability distribution of $g(\underline{X};\theta=(\mu,\sigma^2))$ is t distribution with $(n-1)$ degrees of freedom, i.e., t_{n-1} which does not depend on $\theta = (\mu,\sigma^2)$ or any other unknown parameter.

Now, we are more interested in finding out the interval estimates for μ and σ^2. So, now first we have to construct pivotal quantity. So, now this is the main task. Once you complete this task after that the entire story becomes very simple and it becomes very simple to obtain the confidence interval.

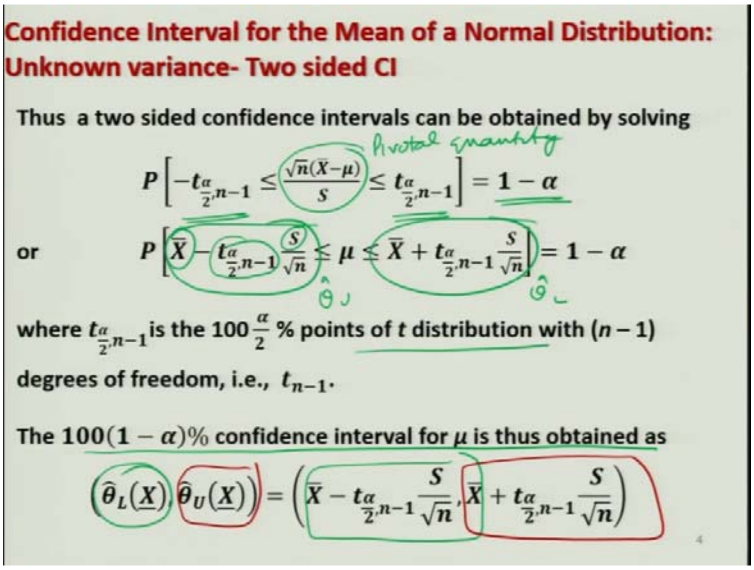
So, now can you recall the t distribution? It was like this that if $X_1, X_2, \ldots, X_n$ that is a random sample from $N(\mu, \sigma^2)$ then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and from the Chi-square distribution do you recall that $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ and we also have done one result while discussing the distributional properties of the sample mean in an earlier lecture that both $\bar{X}$ and $(n-1)\frac{S^2}{\sigma^2}$ are independently distributed.

So, now we can define the pivotal quantity as say $\frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\frac{S^2}{\sigma^2}}{n-1}}}$. So, if you try to simplify it, this will come out to be $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$. So, this is a statistics which is going to follow a t distribution with n - 1 degrees of freedom.

Now, I do not need to repeat it that I will simply refer you to t distribution where we have discussed all these things. So, now this is going to be a pivotal quantity why? Because

$g(\underline{X}; \theta = (\mu, \sigma^2))$ that is depending upon μ σ^2 $X_1, X_2, \ldots, X_n$ but this quantity $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ although this is a function of here μ but the probability distribution of this quantity is independent of any of the unknown parameter like μ and it is depending only on the degrees of freedom n - 1. So, now we can use this pivotal quantity to construct the confidence interval.

(Refer Slide Time: 06:32)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

Thus a two sided confidence intervals can be obtained by solving

Pivotal quantity

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

or

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

$\hat{\theta}_U$ $\hat{\theta}_L$

where $t_{\frac{\alpha}{2},n-1}$ is the $100\frac{\alpha}{2}$ % points of t distribution with $(n-1)$ degrees of freedom, i.e., t_{n-1}.

The $100(1-\alpha)\%$ confidence interval for μ is thus obtained as

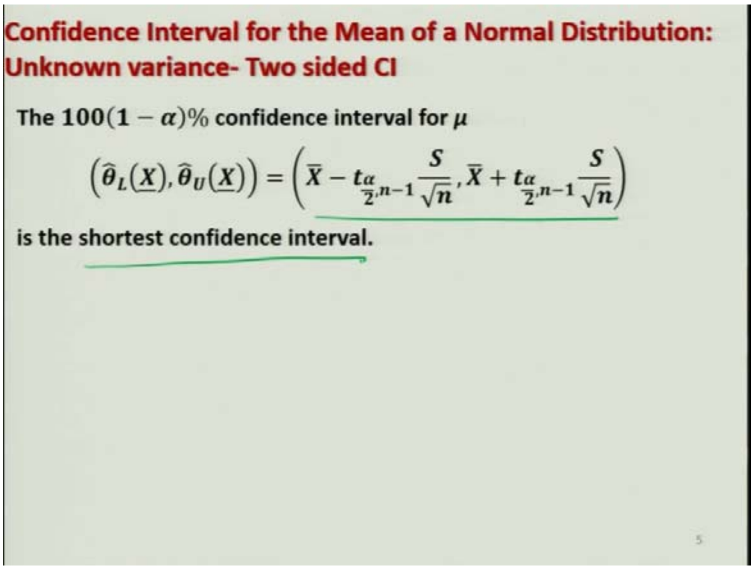
$$\left(\hat{\theta}_L(X), \hat{\theta}_U(X)\right) = \left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right)$$

Now, once you know this thing then after that the light becomes very systematic and simple. If you want to find out the two-sided confidence interval, you simply have to see here that this is the pivotal quantity which is lying between $-t_{\frac{\alpha}{2},n-1}$ and $+t_{\frac{\alpha}{2},n-1}$ and the probability of this event is equal to 1 – α according to the definition of confidence interval.

So, now if you simply try to solve it exactly in the same way as we did in the case of normal distribution, you can obtain the lower limit and upper limit here like this. So, where this value $t_{\frac{\alpha}{2},n-1}$ can be obtained from the probability tables of t distribution or you can also obtained directly from the software also nowadays, right this capital $\bar{X}$ can be obtained on the basis of sample, capital S can be obtained that is the standard error can be obtained by using the observation $X_1, X_2, \ldots, X_n$.

So, all these $\hat{\theta}_U$ and $\hat{\theta}_L$ which are obtained here they can be computed very easily. So, this is your $\hat{\theta}_U$ and this is here $\hat{\theta}_L$. So, I can write down here that 100(1 – α)% confidence interval for μ in such a case is obtained as $\hat{\theta}_L$, this is the lower limit of the interval which is here like this $\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$ and the upper limit is $\hat{\theta}_U$ which is dependent on $X_1, X_2, \ldots, X_n$, this is given by $\bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$. Exactly on the same way if you try to see the structure is very much similar to what you have obtained in the case of normal distribution. The structure is like that you are just using here $t_{\alpha/2}$ in place of $Z_{\alpha/2}$ and you are using here S in place of σ_0.

(Refer Slide Time: 08:33)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

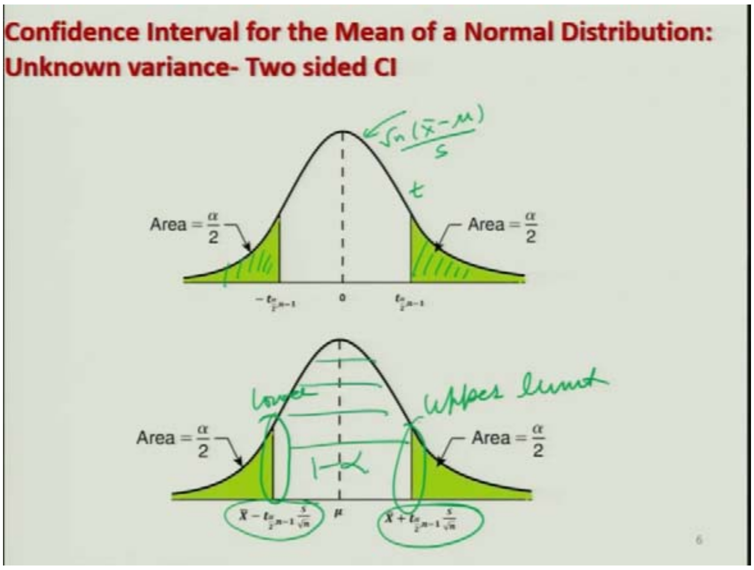
The $100(1-\alpha)\%$ confidence interval for μ

$$\left(\hat{\theta}_L(\underline{X}), \hat{\theta}_U(\underline{X})\right) = \left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \underline{\bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}}\right)$$

is the shortest confidence interval.

So, now I can say that this is the confidence interval in the case when σ^2 is unknown to us and in case if you try to judge this confidence interval on the basis of the length of the confidence interval, then one can very easily show that the interval what we have obtained this is the shortest confidence interval. So, that means this is a good confidence interval.

(Refer Slide Time: 09:02)



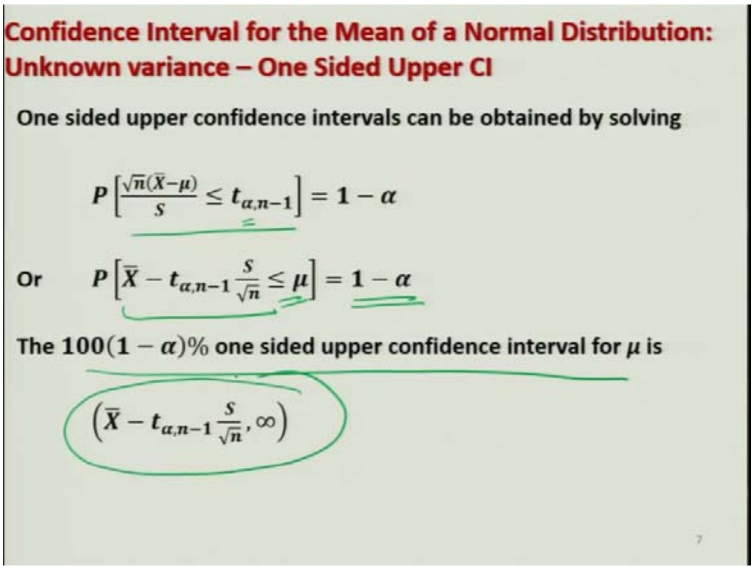
Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

$\frac{\sqrt{n}(\bar{x}-\mu)}{s}$
t
Area = $\frac{\alpha}{2}$
Area = $\frac{\alpha}{2}$
$-t_{\frac{\alpha}{2},n-1}$
0
$t_{\frac{\alpha}{2},n-1}$
lower
upper limit
Area = $\frac{\alpha}{2}$
Area = $\frac{\alpha}{2}$
$1-\alpha$
$\bar{X}-t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$
μ
$\bar{X}+t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$
6

And if you try to see here how it will look like right, so here you can see that we are trying to take the value of this area to be α/2 on the left hand side and right hand side of the t distribution. So, this is actually $_{\alpha/2}$. This is the distribution of this quantity, this is the curve of this quantity.

And now in case if you try to translate it the confidence interval will look like here, this is here the lower limit whose value is $\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$ and this is here the upper limit whose value here is $\bar{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$. So, this is how this two sided confidence interval will look like. Two-sided means because the confidence limits are on both the sides and the middle part this area is 1 – α.

(Refer Slide Time: 10:00)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – One Sided Upper CI

One sided upper confidence intervals can be obtained by solving

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

Or $$P\left[\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}} \le \mu\right] = 1-\alpha$$

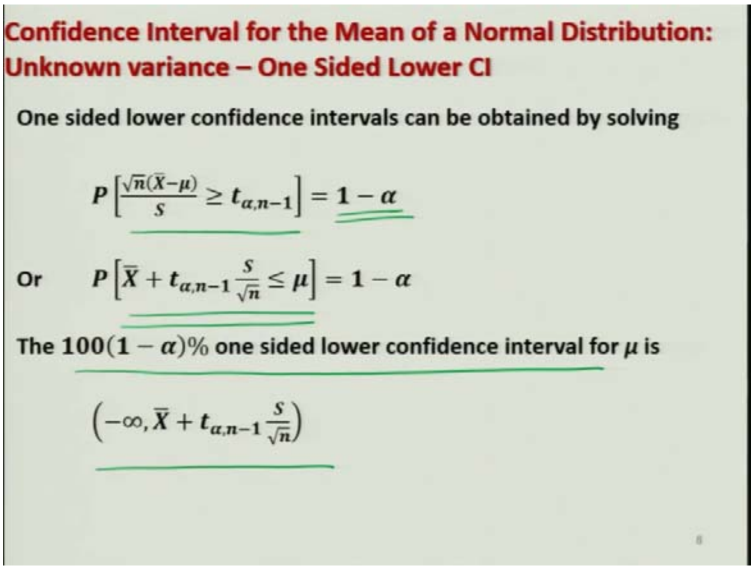
The $100(1-\alpha)\%$ one sided upper confidence interval for μ is

$$\left(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right)$$

And in case if you want to find out the one sided confidence interval, so the one sided upper confidence interval can be obtained just by solving exactly in the same way as we have done in the case of normal distribution when σ^2 is known to us. So, we can write down here $P\left[\frac{\sqrt{n}(\bar{X}-\mu)}{S} \le t_{\alpha,n-1}\right] = 1-\alpha$. Now, it will be see here t_α because the entire area is only α.

So, if you try to solve it simply this limit will come out to be $\left[\bar{X} - t_{\alpha,n\ 1}\frac{S}{\sqrt{n}} \le \mu\right] = 1-\alpha$. So, the 100(1 – α)% one sided upper confidence interval for μ is obtained here like this $\left(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right)$.

(Refer Slide Time: 10:58)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – One Sided Lower CI

One sided lower confidence intervals can be obtained by solving

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

Or $$P\left[\bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}} \leq \mu\right] = 1-\alpha$$

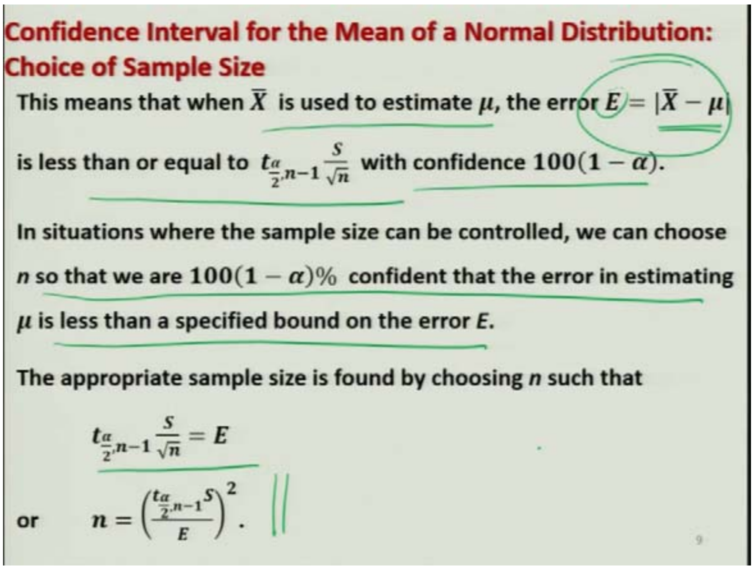
The $100(1-\alpha)\%$ one sided lower confidence interval for μ is

$$\left(-\infty, \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$$

And similarly you can also obtain the one sided lower confidence interval by solving this condition that $P\left[\frac{\sqrt{n}(\bar{X}-\mu)}{S} \geq t_{\alpha,n-1}\right] = 1-\alpha$ and if you simply try to solve it, you get here μ is greater than or equal to $\bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}$.

So, the 100(1 – α)% one sided lower confidence interval for μ is obtained as the region between $\left(-\infty, \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$. So, you can see here this is exactly on the same line as you did in the last lecture, that is what I had requested you in the beginning that if you have done that lecture then you will understand it very easily.

(Refer Slide Time: 11:51)



Confidence Interval for the Mean of a Normal Distribution: Choice of Sample Size

This means that when $\overline{X}$ is used to estimate μ, the error $E = |\overline{X} - \mu|$ is less than or equal to $t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$ with confidence $100(1-\alpha)$.

In situations where the sample size can be controlled, we can choose n so that we are $100(1-\alpha)\%$ confident that the error in estimating μ is less than a specified bound on the error E.

The appropriate sample size is found by choosing n such that

$$t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = E$$

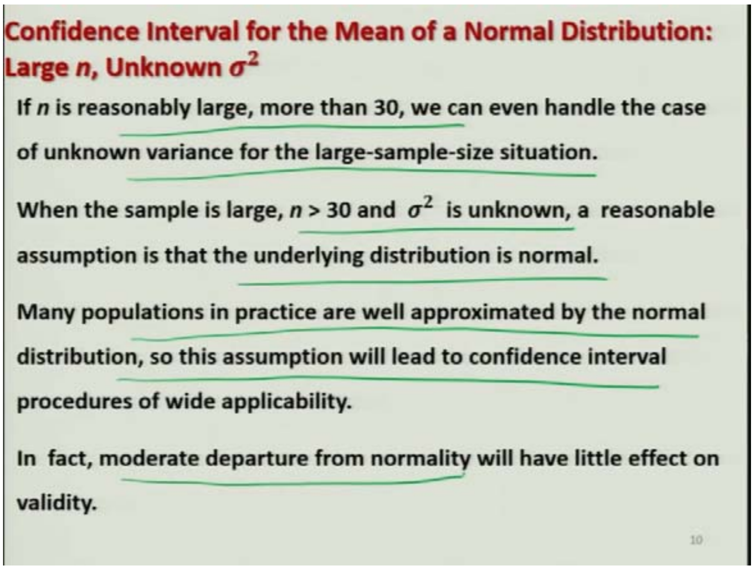
or $$n = \left(\frac{t_{\frac{\alpha}{2},n-1}S}{E}\right)^2.$$

9

So, now as we have obtained the sample size in the case of when σ^2 was known, so similarly in this case also we can obtain the sample size. So, here since $\overline{X}$ is being used to estimate the mu, so some of the values of $\overline{X}$ will be less than μ some values will be more than μ. So, the absolute error between $\overline{X}$ and μ can be expressed by here quantity E. So, we are now saying that we would like to find out the value of sample size and we would like to find out the sample size in such a way such that our error is less than or equal to $t_{\frac{\alpha}{2},n}\ 1\frac{S}{\sqrt{n}}$ with confidence coefficient 100(1 – α)%.

So, in such situation where the sample size can be controlled, we can choose the value of small and that is the sample size so that 100(1 – α)% confidence interval or at least we are 100(1 – α)% confident that the error is in the estimating the value of parameter μ is less than a specified bound on the error E which is given here. So, now the appropriate sample size can be found by solving the equation that $t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = E$ and this will give us a value of here $n = \left(\frac{t_{\frac{\alpha}{2},n-1}S}{E}\right)^2$.

(Refer Slide Time: 13:25)



Confidence Interval for the Mean of a Normal Distribution: Large n, Unknown σ^2

If n is reasonably large, more than 30, we can even handle the case of unknown variance for the large-sample-size situation.

When the sample is large, $n > 30$ and σ^2 is unknown, a reasonable assumption is that the underlying distribution is normal.

Many populations in practice are well approximated by the normal distribution, so this assumption will lead to confidence interval procedures of wide applicability.

In fact, moderate departure from normality will have little effect on validity.

So, from there you can obtain the optimal value of n but you have to just be careful in the case of this case when σ^2 is unknown and you are using the t distribution. In this case suppose if there is a condition that your sample size is more than 30 and σ^2 is unknown, so do you remember that we had a good discussion in understanding that when the degrees of freedom in the t distribution becomes more than 30, then the curve of t distribution and the curve of normal the distribution they becomes almost the same.

So, in such a cases either you are trying to find the critical values from t distribution or normal distribution they are going to be the same. So, when we are trying to consider the situation where your sample size is large say more than 30 and σ^2 is unknown then in that case you can also use the critical values $z_{\alpha/2}$ in place of the critical values $t_{\alpha/2}$.

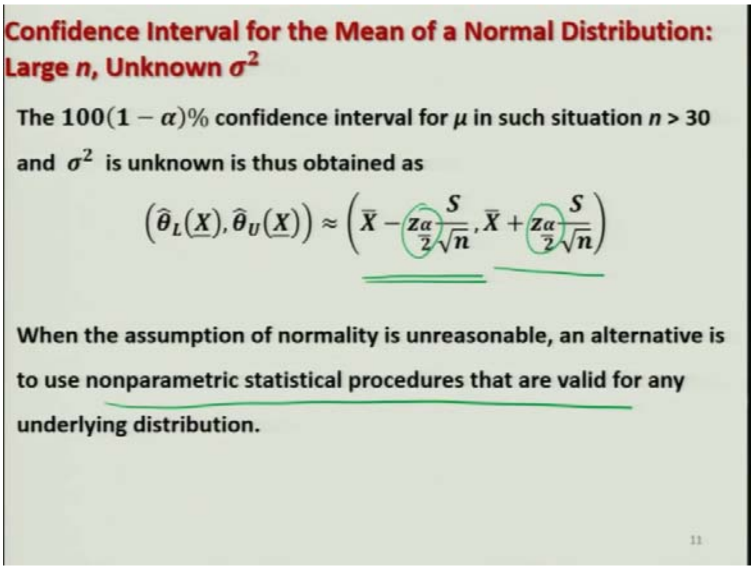
Why I am calling them critical values that will be clear to you when we try to start the discussion on the test of hypothesis in the next lecture but at this moment you can just assume that they are called as critical values. So, the moral of the story is whenever you have the sample size more than 30, then you can use either the normal distribution or the t distribution and both are going to give you the same result.

Although in the case when σ^2 is unknown, you are going to estimate it on the basis of the sample but still both the confidence interval are going to be almost the same. So, in case if n is reasonably large say more than 30 then even we can handle the case of unknown variance directly by the last sample size situation.

So, when n is greater than 30 and σ^2 is unknown a reasonable assumption is that the underlying distribution is normal instead of here t. So, and you see in particular in data sciences you can expect that usually the sample size is expected to be definitely more than 30. So, that is the reason that this case becomes there more popular and that is the reason you will see that in the software also people are more interested in developing a software for the case when σ^2 is unknown to us because that is more practical situation.

So, in many populations in practice there are well approximated by the normal distribution, so this assumption will lead to the confidence interval procedures which are of wide applicability. So, they are easy to use so they become more popular they are more convenient, people can use them easily, people can find them easily right and in fact if there is some moderate departure from the normality that may be possible but that is going to have very little effect on the validity of the confidence interval.

(Refer Slide Time: 16:49)



Confidence Interval for the Mean of a Normal Distribution: Large *n*, Unknown σ^2

The $100(1-\alpha)\%$ confidence interval for μ in such situation $n > 30$ and σ^2 is unknown is thus obtained as

$$\left(\hat{\theta}_L(\underline{X}), \hat{\theta}_U(\underline{X})\right) \approx \left(\bar{X} - z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

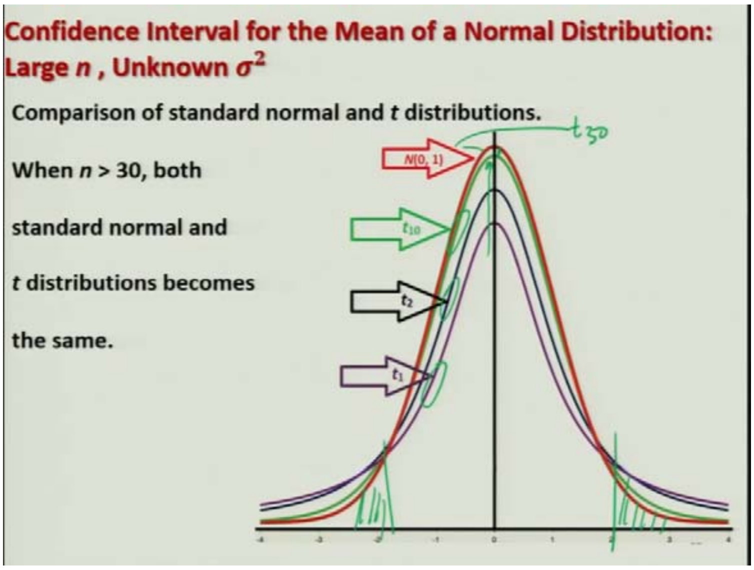
When the assumption of normality is unreasonable, an alternative is to use nonparametric statistical procedures that are valid for any underlying distribution.

11

So, you do not have to worry for those things. So, in such a case when your n is greater than 30 and your σ^2 is unknown, you can redefine the same confidence interval like this $\left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right)$. So, instead of using here the t values we are going to use here the normal values and in case, if you think that the assumption of normality is really going to be not satisfied.

And the data is not really coming from normal distribution, then if you find it or even if you are finding it difficult to verify that data is coming from any given distribution then under those situations my suggestion will be that you can use the non-parametric statistical procedures which do not require the information on the underlying distribution. Well, I am not going to discuss here but surely I can give you this idea once you know once you have done so many topics you can do now non parametric inference yourself without any problem and I assure you that is not going to be difficult.

(Refer Slide Time: 18:09)



Confidence Interval for the Mean of a Normal Distribution: Large n, Unknown σ^2

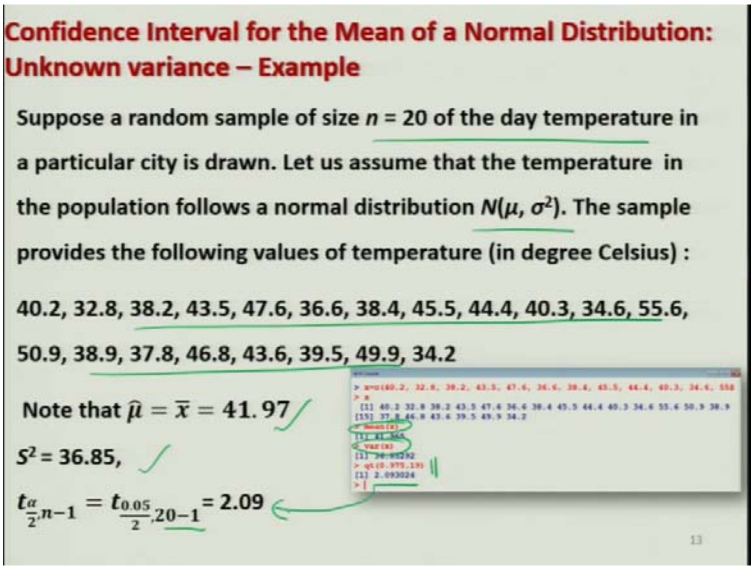
Comparison of standard normal and t distributions.

When $n > 30$, both standard normal and t distributions becomes the same.

So, now if you try to get convinced here I have just constructed some graphics so you can see here this is the curve with t distribution with 1 degrees of freedom, this black color is the curve for the t distribution with 2 degrees of freedom and then this green curve is the t distribution curve with 10 degrees of freedom and you can see here the degrees of freedom are increasing and then height of t distribution is increasing and here in this red color this is the normal 0,1.

So, you can see here at t equal to 10, they are very close but if you try to increase it to n equal to 30 that means 30 degrees of freedom then N(0, 1) and t 30 they will become identical. So, both distribution are the same, now if you want to know this value somewhere here or somewhere here on the left hand side of that curve then either you try to use the normal distribution or t distribution with the say 30 degrees of freedom or more than 30 degrees of freedom that will not make any difference.

(Refer Slide Time: 19:17)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – Example

Suppose a random sample of size n = 20 of the day temperature in a particular city is drawn. Let us assume that the temperature in the population follows a normal distribution $N(\mu, \sigma^2)$. The sample provides the following values of temperature (in degree Celsius) :

40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

Note that $\hat{\mu} = \bar{x} = 41.97$

$S^2 = 36.85,$

$t_{\frac{\alpha}{2},n-1} = t_{\frac{0.05}{2},20-1} = 2.09$

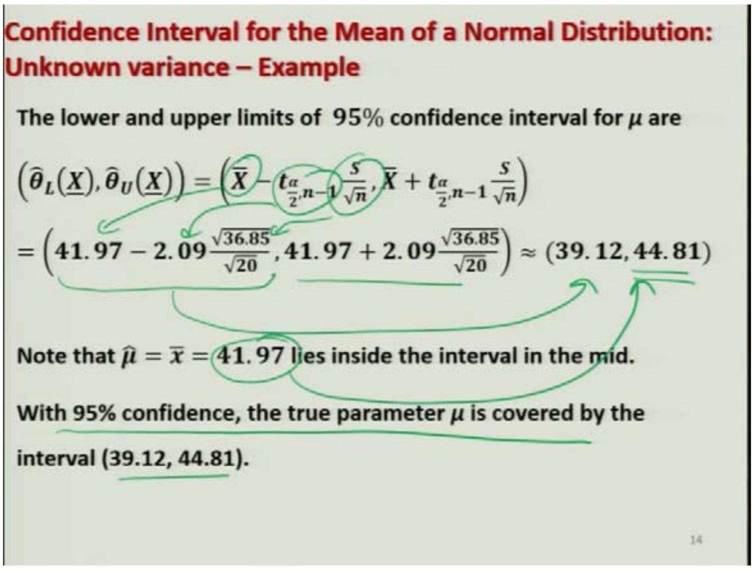
```
> x=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55
> x
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6 50.9 38.9
[15] 37.8 46.8 43.6 39.5 49.9 34.2
> mean(x)
[1] 41.965
> var(x)
[1] 36.85292
> qt(0.975,19)
[1] 2.093024
> |
```

13

So, now let me try to take an example and try to compute this confidence interval manually and on the R software also. So, I am going to take the same example which I considered in the last lecture. The only difference is that there I assumed the σ^2 to be 36 but in this case we are I am assuming that σ^2 is unknown and we are going to estimate it on the basis of given sample of data.

So, we had 20 values on the day temperature in a particular city and they were and the temperature were assumed to follow a $N(\mu, \sigma^2)$ distribution and these are the values of the sample in degrees Celsius that we had obtained and means if you want to compute the confidence interval manually, I have computed here the mean of x variance of x which are coming out to be here 41.97 and 36.85 respectively and the value of this t α/2 n - 1 which is here at say t 0.05 divided by 2 and with 20 - 1 that is 19 degrees of freedom, you can see here I have computed here and this is coming out to be 2.09. So, I have computed all these values separately in the R software and this is screenshot is given here that you can actually verify yourself.

(Refer Slide Time: 20:32)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – Example

The lower and upper limits of 95% confidence interval for μ are

$$\left(\hat{\theta}_L(\underline{X}), \hat{\theta}_U(\underline{X})\right) = \left(\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right)$$

$$= \left(41.97 - 2.09\frac{\sqrt{36.85}}{\sqrt{20}}, 41.97 + 2.09\frac{\sqrt{36.85}}{\sqrt{20}}\right) \approx (39.12, 44.81)$$

Note that $\hat{\mu} = \bar{x} = 41.97$ lies inside the interval in the mid.

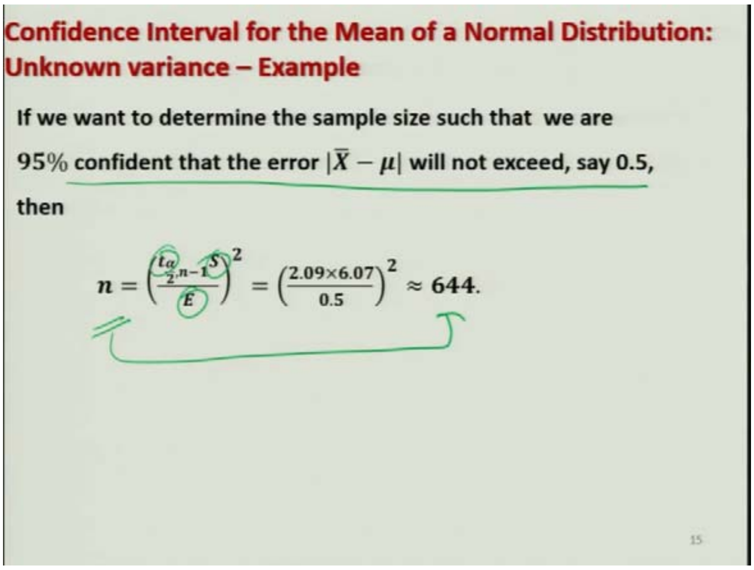
With 95% confidence, the true parameter μ is covered by the interval (39.12, 44.81).

14

But my objective here is to put all these values in the expression and try to obtain these values manually and then I will try to obtain the same values in the R software also. So, I try to substitute here the value of capital $\overline{X}$ value of t and value of S by root n here like this and I obtain the value of this expression as 39.12.

So, the lower confidence limit of μ in this case is coming out to be 39.12 and similarly you can compute the upper limit also just by replacing - by plus sign and this value is coming out to be 44.81 and you can see here this is sample mean whose value is 41.97, this is lying inside this interval in fact in the mid of the interval. So, now I can see that with 95 percent confidence, the true parameter μ is covered by this interval 39.1 122, 44.81. So, you can see it is not difficult at all to compute such confidence interval even when σ^2 is unknown to us.

(Refer Slide Time: 21:41)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – Example

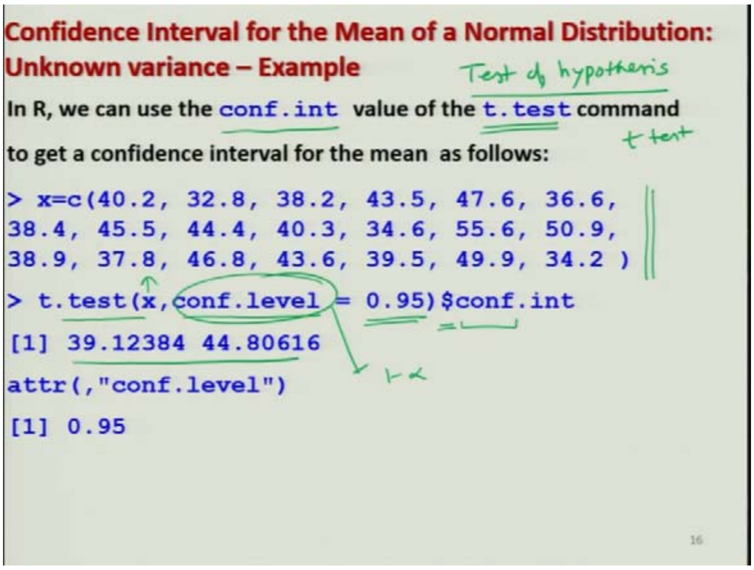
If we want to determine the sample size such that we are 95% confident that the error $|\bar{X} - \mu|$ will not exceed, say 0.5, then

$$n = \left(\frac{t_{\frac{\alpha}{2},n-1} S}{E}\right)^2 = \left(\frac{2.09 \times 6.07}{0.5}\right)^2 \approx 644.$$

And if you want to find out the value of the sample size such that we are 95 percent confident that the absolute error $\bar{X}$ - μ will not exceed say some value say 0.5, then we already have obtained the expression. You simply have to substitute the value of t, S and here E and you can compute the value of here n which is coming out to be here approximately 644.

So, you can see here that this value has increased in comparison to the sample size that you had obtained in the earlier lecture when σ^2 was assumed to be known. That is expected means when you are trying to estimate something, then definitely you are going to increase the variability and once you know something that means there is no error into it there is no randomness.

(Refer Slide Time: 22:30)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance – Example

Test of hypothesis

In R, we can use the conf.int value of the t.test command to get a confidence interval for the mean as follows:

t test

```
> x=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6,
38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9,
38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2 )
> t.test(x,conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
attr(,"conf.level")
[1] 0.95
```

$1-\alpha$

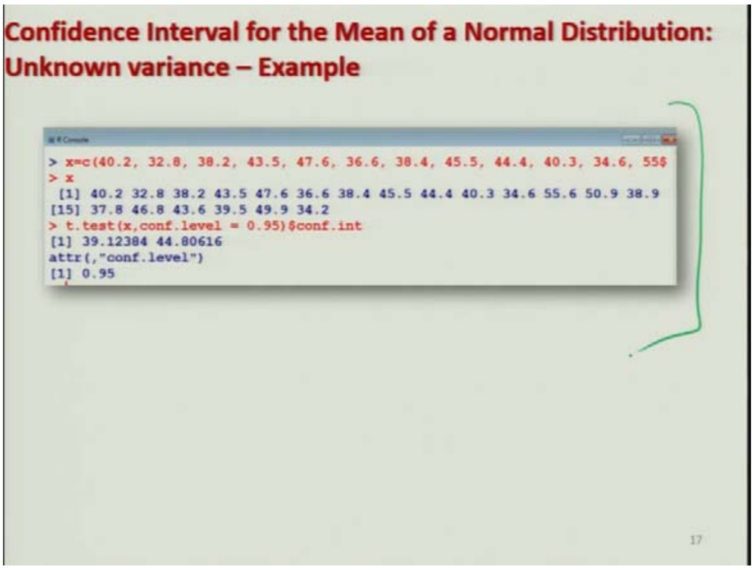
16

Now, I try to do the same example in the R software but let me be honest, here I am going to use a command which is related to the topic which I am going to cover in the forthcoming lecture. It is about test of hypothesis, actually this test of hypothesis and confidence interval estimation they are inter-related. So, that is why I am just using here the command, I am giving you the command here but when I will be considering the topic of key test, then I will try to give you the complete detail of this R command.

So, we are using here the command conf dot int which is confidence interval from the value of the t test. And suppose we have given the data here as say here like this suppose here x and then you are simply try to write down here t dot test x which is the data vector here and confidence level that is c-o-n-f dot l-e-v-e-l is equal to 0.95 and then you have to write down here dollar and then c-o-n-f dot i-n-t.

So, this is here the value of 1 – α which is given by conf dot level. So, at the moment I would request you that you try to accept it as on the as a face value and later on I will try to give you the entire results. And if you try to get this value, this will come out to be a 39.1224.80 and this is the interval that you can obtain here. So, you can see here that it is not difficult at all to obtain such confidence interval.

(Refer Slide Time: 24:18)



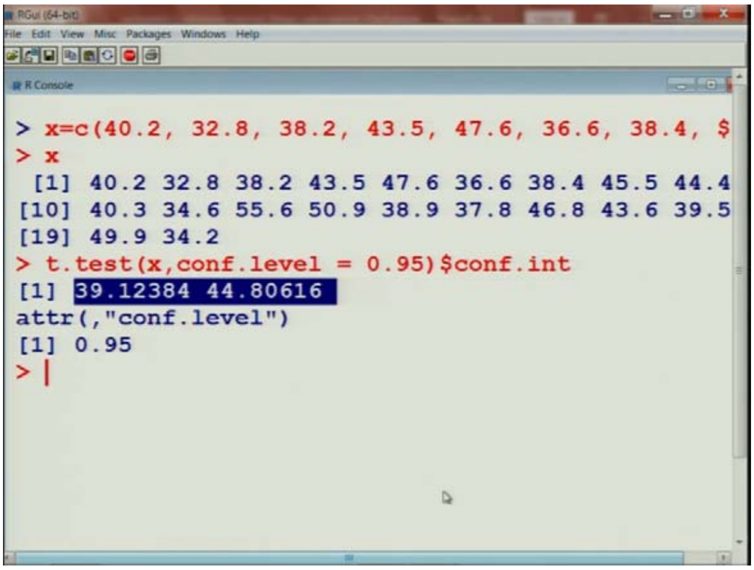
Confidence Interval for the Mean of a Normal Distribution: Unknown variance – Example

```
> x=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55$
> x
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6 50.9 38.9
[15] 37.8 46.8 43.6 39.5 49.9 34.2
> t.test(x,conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
attr(,"conf.level")
[1] 0.95
```

17

And this is here the screenshot but now let me try to first show you these things on the R console.

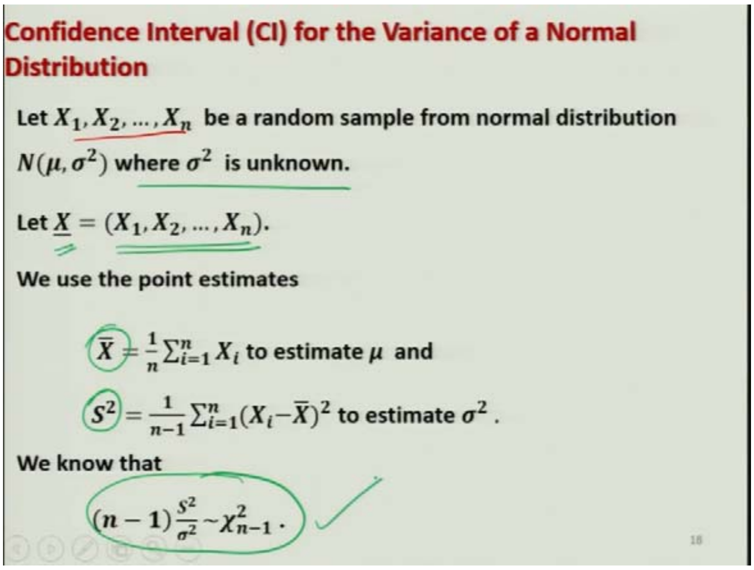
(Refer Slide Time: 24:25)



```
> x=c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, $
> x
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4
[10] 40.3 34.6 55.6 50.9 38.9 37.8 46.8 43.6 39.5
[19] 49.9 34.2
> t.test(x,conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
attr(,"conf.level")
[1] 0.95
> |
```

So, let me try to copy this data to save some time you can see here this is here the data say I am calling it here I say x just for the sake of simplicity and then I try to use this command for obtaining the confidence interval and you can see here this is coming out to be like this 39.12384 to 44.80616. So, this is the 95 percent confidence and these are the lower and upper limits of the confidence interval.

(Refer Slide Time: 25:02)



Confidence Interval (CI) for the Variance of a Normal Distribution

Let $X_1, X_2, \dots, X_n$ be a random sample from normal distribution $N(\mu, \sigma^2)$ where σ^2 is unknown.

Let $\underline{X} = (X_1, X_2, \dots, X_n)$.

We use the point estimates

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate μ and

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to estimate σ^2.

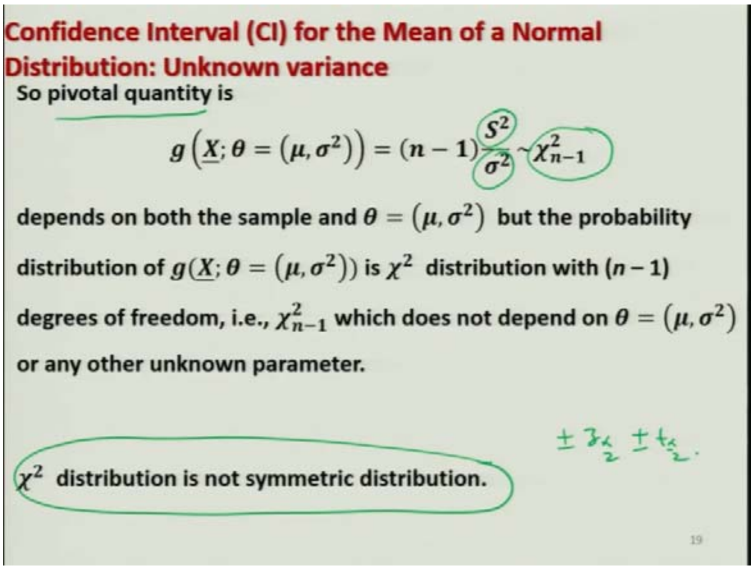
We know that

$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$.

So, now after this we come back to our slide and we try to discuss one more topic. This is about the confidence interval estimation for the variance of a normal distribution. You have two parameters in the normal distribution μ and σ^2 . So, up to now you have only assumed that σ^2 is known or unknown and then you have tried to provide the confidence interval for the mean but now we are assuming that σ^2 is unknown, so we can estimate the σ^2 through point estimate as well as through the confidence interval estimation.

So, how to get it done that is what I am going to now discuss here. So, we have a random sample $X_1, X_2, \dots, X_n$ from $N(\mu, \sigma^2)$ and where σ^2 is unknown and I am indicating the data vector here as the x underscore just for the sake of simplicity and we know that in this case the mean and variance can be estimated by say sample mean and this quantity capital S square as we have done earlier. And now we have to create a pivotal quantity. So, for that we are going to take the help of Chi-square distribution and the topic that we already have understood actually. So, we are going to use them, so we know that $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$.

(Refer Slide Time: 26:21)



Confidence Interval (CI) for the Mean of a Normal Distribution: Unknown variance

So pivotal quantity is

$$g\left(\underline{X};\theta=(\mu,\sigma^2)\right)=(n-1)\frac{S^2}{\sigma^2}\sim\chi^2_{n-1}$$

depends on both the sample and $\theta=(\mu,\sigma^2)$ but the probability distribution of $g(\underline{X};\theta=(\mu,\sigma^2))$ is χ^2 distribution with $(n-1)$ degrees of freedom, i.e., χ^2_{n-1} which does not depend on $\theta=(\mu,\sigma^2)$ or any other unknown parameter.

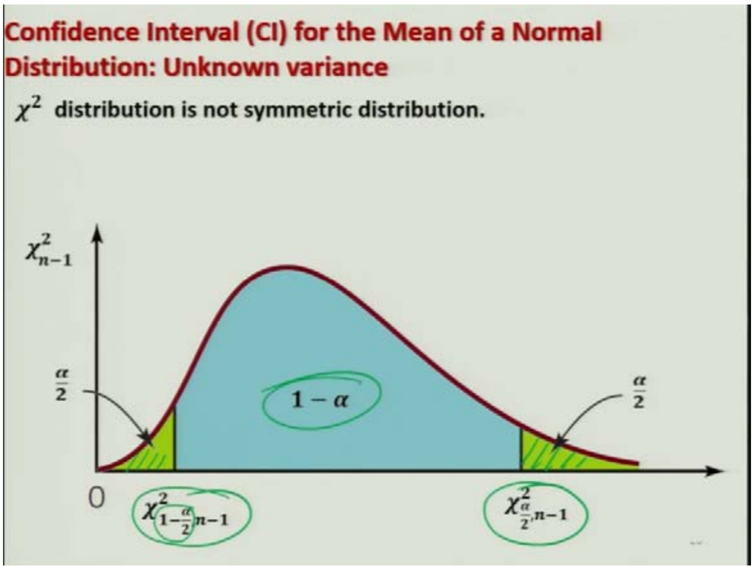
$\pm z_{\frac{\alpha}{2}}$ $\pm t_{\frac{\alpha}{2}}$.

χ^2 distribution is not symmetric distribution.

19

So, now I can consider this quantity as pivotal quantity, why? Because this is the quantity which is depending on the unknown parameter σ^2 and X_1, X_2,…, X_n through S square but its distribution is independent of the unknown parameter which is Chi-square with only n - 1 degrees of freedom. So, now I can use this statistic to construct the confidence interval. Remember one thing that Chi-square distribution is not a symmetric distribution, so you cannot use here like plus - $z_{\alpha/2}$ or plus - $t_{\alpha/2}$. You have to be careful.

(Refer Slide Time: 27:01)

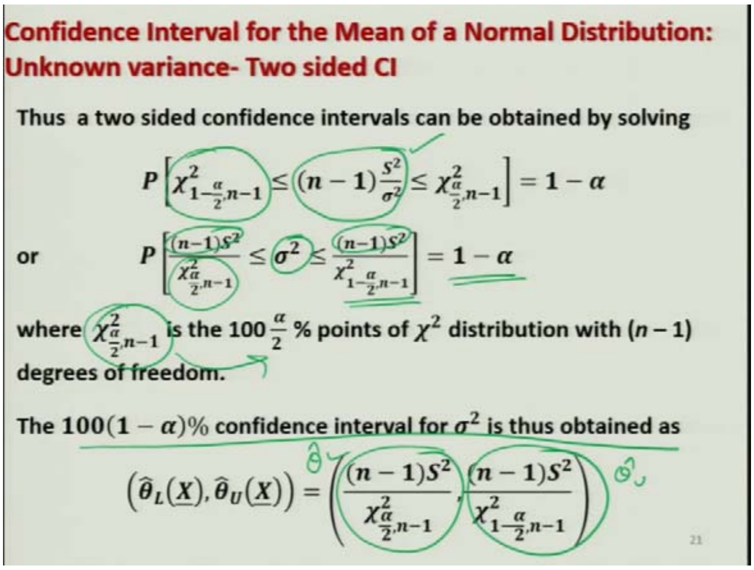


Confidence Interval (CI) for the Mean of a Normal Distribution: Unknown variance

χ^2 distribution is not symmetric distribution.

So, actually this will look like this. So, in case if you are trying to take here this area to be here α/2, then this critical point is going to be $\chi^2_{\frac{\alpha}{2},n-1}$ and then this area will become here like α/2 but this point is going to be indicated by $\chi^2_{1-\frac{\alpha}{2},n-1}$. And this middle blue part will be 1 – α. So, always remember that this Chi-square and event f they are not the symmetric distribution.

(Refer Slide Time: 27:34)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

Thus a two sided confidence intervals can be obtained by solving

$$P\left[\chi^2_{1-\frac{\alpha}{2},n-1} \le (n-1)\frac{S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right] = 1-\alpha$$

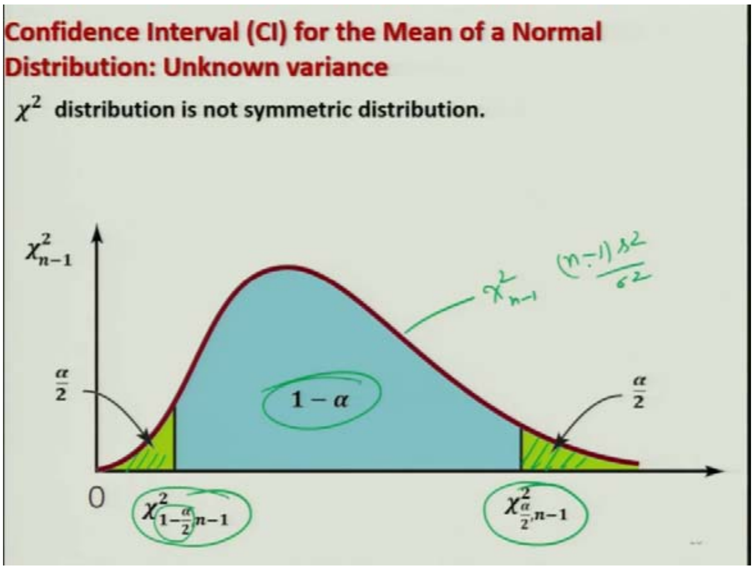
or $$P\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = 1-\alpha$$

where $\chi^2_{\frac{\alpha}{2},n-1}$ is the $100\frac{\alpha}{2}$ % points of χ^2 distribution with $(n-1)$ degrees of freedom.

The $100(1-\alpha)\%$ confidence interval for σ^2 is thus obtained as

$$\left(\hat{\theta}_L(X), \hat{\theta}_U(X)\right) = \left(\overset{\theta_L}{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}}, \overset{\theta_U}{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}\right)$$

21



Confidence Interval (CI) for the Mean of a Normal Distribution: Unknown variance

χ^2 distribution is not symmetric distribution.

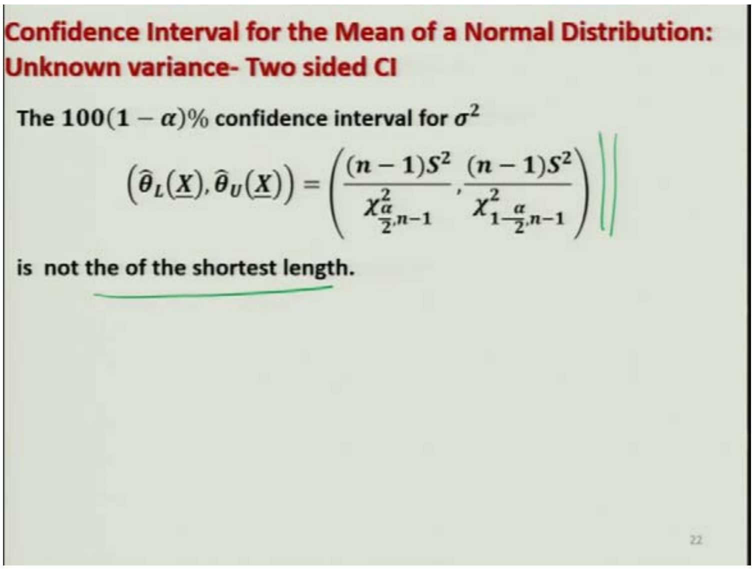
So, now I use this statistics $(n-1)\frac{S^2}{\sigma^2}$ to find out the confidence interval. So, I can say that this $(n-1)\frac{S^2}{\sigma^2}$, so this is the curve of Chi-square with n - 1 degrees of freedom, this is actually

$(n-1)\frac{S^2}{\sigma^2}$. So, this is the curve of this quantity. So, now this quantity is going to lie between these two values you can see here this and here this.

So, that is what precisely I am going to write down here that this quantity is lying between these two limits Chi-square value with 1 - α/2 and α/2 and with n - 1 degrees of freedom, as I shown you in the figure also. So, now you can simply solve it and then this $\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$and the probability of this event is going to be exactly 1 – α. So, this $\chi^2_{\frac{\alpha}{2},n-1}$ is the 100 α/2 percent point on the Chi-square distribution with n - 1 degrees of freedom that we know.

So, now I can write down here that the 100(1 – α)% confidence interval for σ^2 is obtained here like this. This is here the $\hat{\theta}_L$ which is based on $X_1, X_2, \ldots, X_n$ and this is here $\hat{\theta}_U$which is again based on $X_1, X_2, \ldots, X_n$ and we have $X_1, X_2, \ldots, X_n$ so we can compute the value of capital S^2 and then we can obtain the value of Chi-square from the table or from the R software and we can compute this interval.

(Refer Slide Time: 29:22)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

The $100(1-\alpha)\%$ confidence interval for σ^2

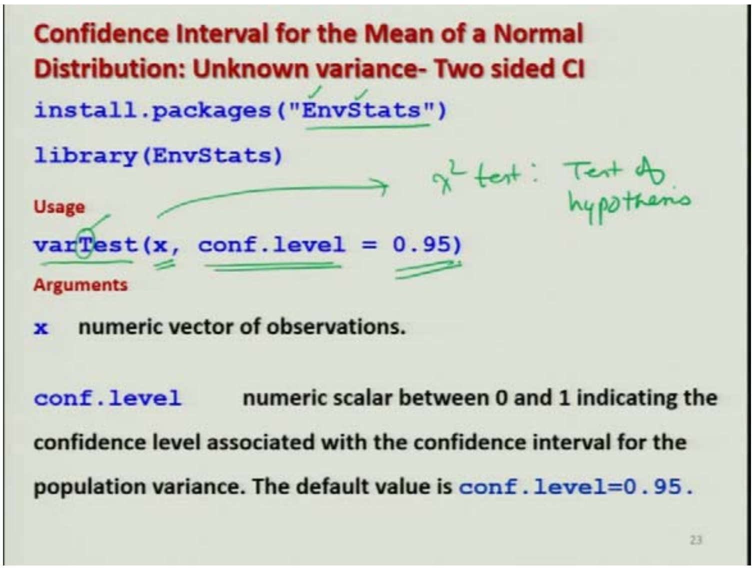
$$\left(\hat{\theta}_L(\underline{X}), \hat{\theta}_U(\underline{X})\right) = \left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$$

is not the of the shortest length.

22

But you have to keep in mind that this 100(1 – α)% confidence interval for σ^2 is not the shortest length confidence interval just like it was in the case of normal distribution where we tried to find the confidence interval of mean when σ^2 is known and unknown. So, this is what you have to keep in mind.

(Refer Slide Time: 29:43)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI

```
install.packages("EnvStats")
library(EnvStats)
```

Usage

```
varTest(x, conf.level = 0.95)
```

χ^2 test : Test of hypothesis

Arguments

`x` numeric vector of observations.

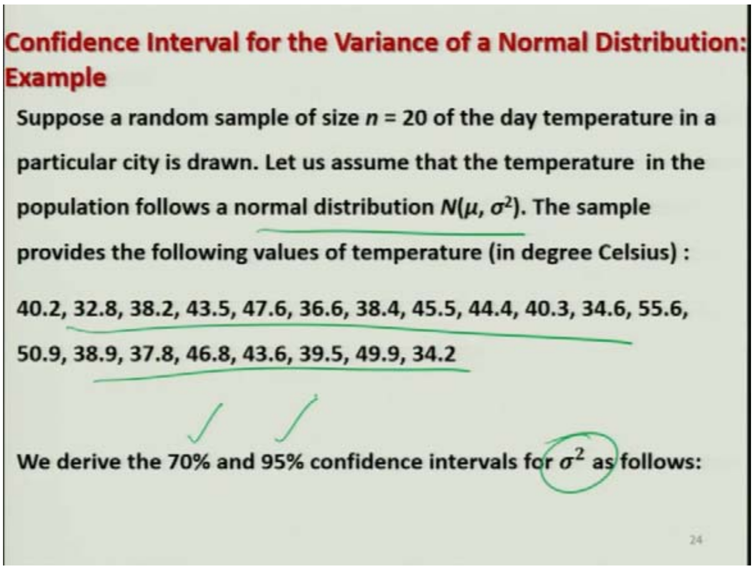
`conf.level` numeric scalar between 0 and 1 indicating the confidence level associated with the confidence interval for the population variance. The default value is `conf.level=0.95`.

23

Now, the question here is how we can obtain this confidence interval in our software? Well for that you have to use here a special package this is e-n-v-s-t-a-t-s and here this E and this S they are in capital letters. You have to be careful, so this is a package in which some data sets of a book of environmental statistics is there. So, they have given this program to compute the confidence interval for σ^2 and let me be honest here that once again this command which we are going to use here for the construction of confidence interval it is v-a-r-T-e-s-t where T is going to be in the capital letter.

Actually this is for the Chi-square test which is a topic of test of hypothesis which I am going to take up after this lecture and it will come very soon before us. So, that is why I am unable to give you the complete details of this command but here I would request you that you please accept it on the face value and use this command to construct the confidence interval for σ^2 in the case of normal distribution. So, the command here is simple v-a-r-T-e-s-t where this capital T is here and then you have to give the data vector here x and then you have to specify the confidence interval whatever you want, means the default here is 0.95.

(Refer Slide Time: 31:27)



Confidence Interval for the Variance of a Normal Distribution: Example

Suppose a random sample of size n = 20 of the day temperature in a particular city is drawn. Let us assume that the temperature in the population follows a normal distribution $N(\mu, \sigma^2)$. The sample provides the following values of temperature (in degree Celsius) :

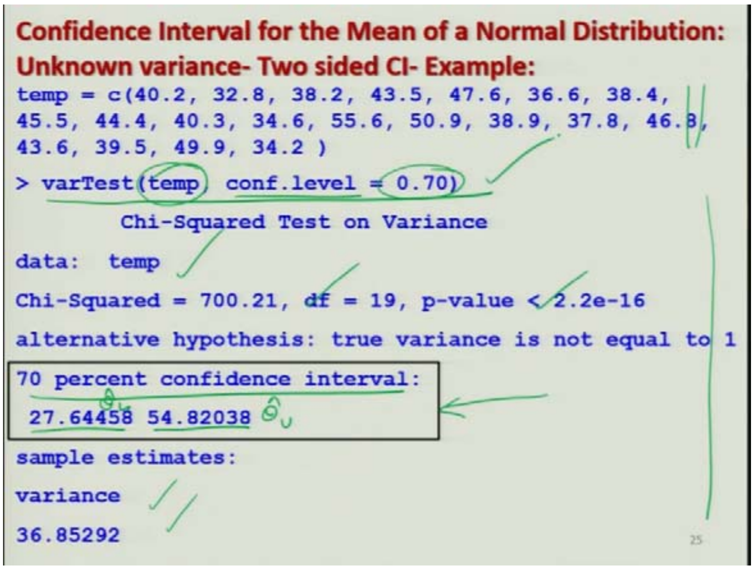
40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

We derive the 70% and 95% confidence intervals for σ^2 as follows:

24

So, now in case if you try to see here that I try to take the here the same example that I just consider that we have the 20 observation on the day temperature of our day in degree Celsius and we are assuming the population to be $N(\mu, \sigma^2)$ and this is the data here and we want to construct here a confidence interval for σ^2 and suppose just for the sake of illustration, I am taking here the confidence interval to be 70 percent and 95 percent you can see using the software, it is very easy.

(Refer Slide Time: 31:58)



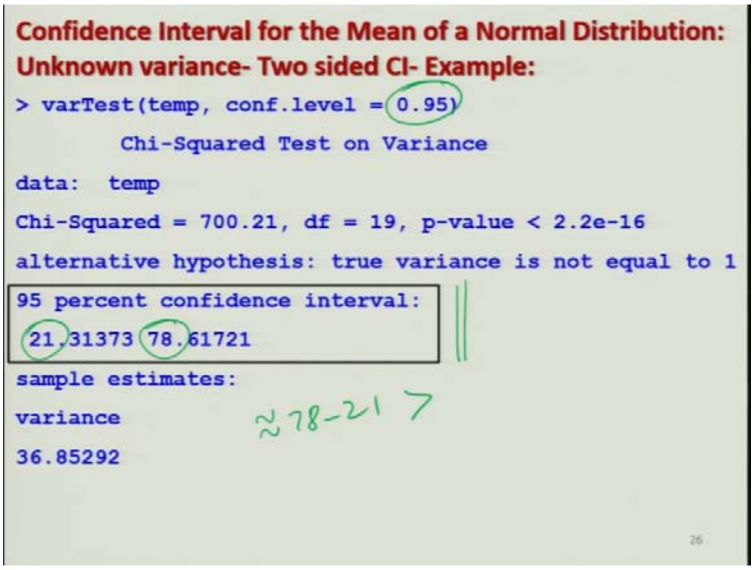
Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI- Example:

```
temp = c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4,
45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8,
43.6, 39.5, 49.9, 34.2 )
> varTest(temp, conf.level = 0.70)
            Chi-Squared Test on Variance
data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
70 percent confidence interval:
 27.64458 54.82038
sample estimates:
variance
36.85292
```

25

Now, if I ask you to do it manually you know that it is very simple now. So, this is here the data and now I try to give this command over here, v-a-r-T-e-s-t on the data here, temperature and then the confidence level here is 0.70 that you can change according to your wish. Your outcome will look actually like this but you have to concentrate only on this part which I have put inside a box. This is 70 percent confidence interval and there are two values here like this, so this is here the value of $\hat{\theta}_L$ and this is the value of a $\hat{\theta}_U$ that is the lower end upper bounds of the confidence interval of σ^2 . Rest all these values whatever they are here that I will try to explain you when I try to consider the topic of a test of hypothesis that is my promise to you we are going to do it.

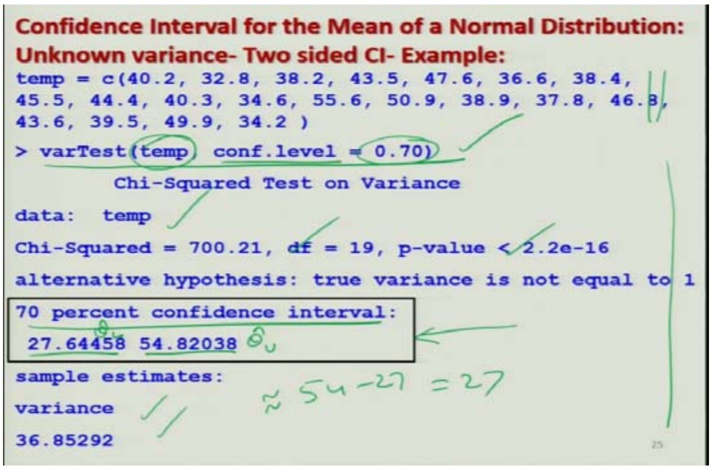
(Refer Slide Time: 33:06)



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI- Example:

```
> varTest(temp, conf.level = 0.95)
        Chi-Squared Test on Variance
data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
21.31373 78.61721
sample estimates:
variance
36.85292
```

$\approx 78-21 \; 7$



Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI- Example:

```
temp = c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4,
45.5, 44.4, 40.3, 34.6, 55.6, 50.9, 38.9, 37.8, 46.8,
43.6, 39.5, 49.9, 34.2 )
> varTest(temp, conf.level = 0.70)
        Chi-Squared Test on Variance
data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
70 percent confidence interval:
 27.64458 54.82038
sample estimates:
variance
36.85292
```

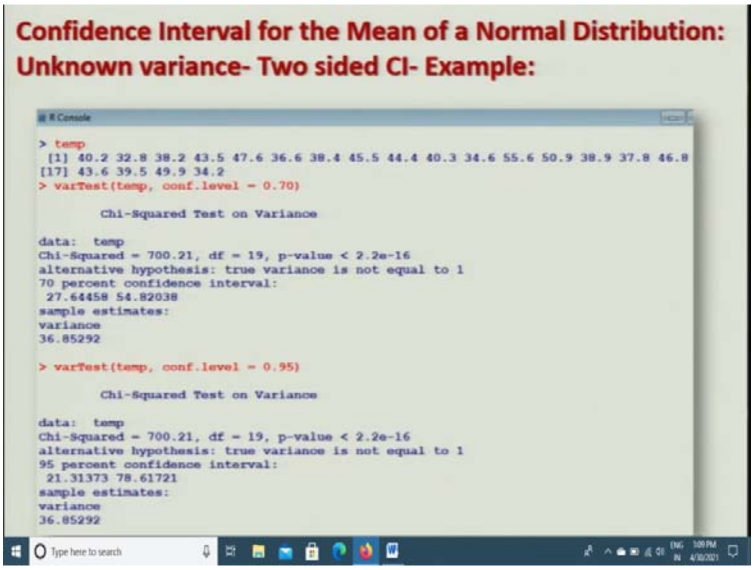
$\hat{\theta}_L$ $\hat{\theta}_U$

$\approx 54-27 = 27$

24

Now, if you try to change this confidence level from 0.70 to 0.95, you can you just use the same command over here and then you will get here this confidence interval but now can you see what is really the difference between the two. Here the confidence interval is from 27 to 54 so the length of the confidence interval is approximately 54 - 27 which is equal to 27 and here in this case the confidence interval limits are 21 nearly and say 78 nearly approximately. So, the length of the confidence interval is approximately 78 - 21. So, now the you can see here this is greater than the length what you have obtained when your confidence level is 70 percent. So, now if you try to recall that in the last lecture when we introduced the concept of confidence interval, we had talked about it.

(Refer Slide Time: 33:56)

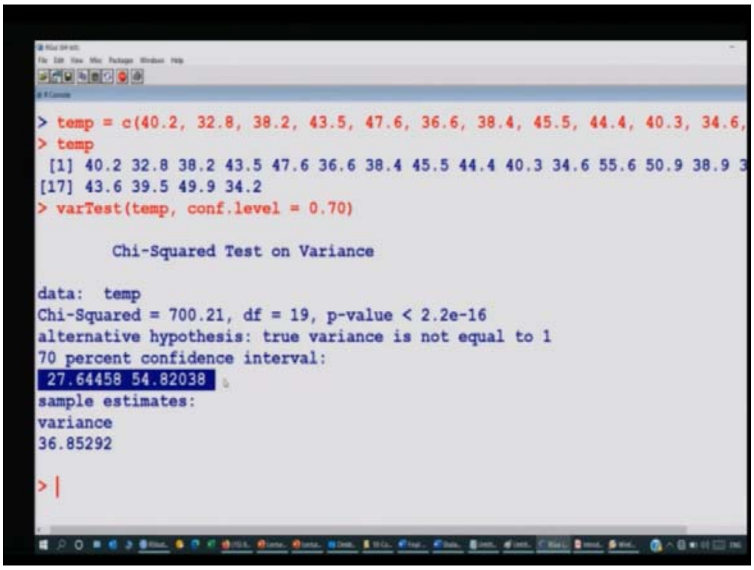


Confidence Interval for the Mean of a Normal Distribution: Unknown variance- Two sided CI- Example:

```
> temp
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6 50.9 38.9 37.8 46.8
[17] 43.6 39.5 49.9 34.2
> varTest(temp, conf.level = 0.70)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
70 percent confidence interval:
 27.64458 54.82038
sample estimates:
variance
36.85292

> varTest(temp, conf.level = 0.95)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
 21.31373 78.61721
sample estimates:
variance
36.85292
```

So, these are outcomes of the same program but now my choice will be that I try to show you these things on the R console. So, let us try to come to R console and try to do the same thing over here.

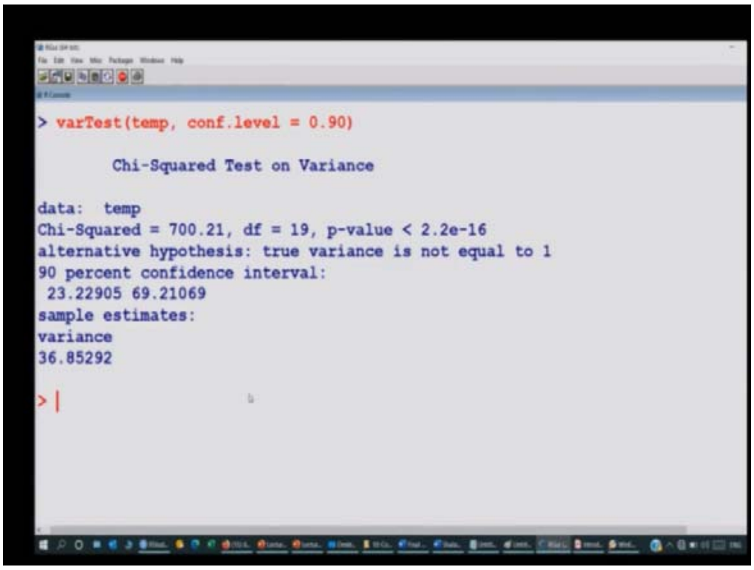
(Refer Slide Time: 34:11)



```
> temp = c(40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3, 34.6,
> temp
 [1] 40.2 32.8 38.2 43.5 47.6 36.6 38.4 45.5 44.4 40.3 34.6 55.6 50.9 38.9 3
[17] 43.6 39.5 49.9 34.2
> varTest(temp, conf.level = 0.70)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
70 percent confidence interval:
 27.64458 54.82038
sample estimates:
variance
36.85292

> |
```

So, let me try to first clear the screen and let me try to enter the data and try to execute it. So, now I have entered here the data on the temperature you can see here this is here like this and now I try to compute the confidence interval. You can see here using this command at confidence level equal to 0.70 this is the outcome. So, you can see here these are the confidence limits.

(Refer Slide Time: 34:44)



```
> varTest(temp, conf.level = 0.90)

        Chi-Squared Test on Variance

data:  temp
Chi-Squared = 700.21, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
90 percent confidence interval:
 23.22905 69.21069
sample estimates:
variance
36.85292

> |
```

And similarly if you try to change the confidence level to be suppose here 90 you can see here these limits are going to change. So, you can see here that it is not difficult to compute this confidence limit on the basis of given sample of data.

So, now with this illustration I come to an end to this lecture and also to the topic of confidence interval estimation. Well one thing I would like to clarify here that we also have confidence interval for different parametric function like if you have two populations and you want to construct the confidence interval for $\mu_1 - \mu_2$ where μ_1 and μ_2 are the means of the 2 population and so on.

But in order to understand these things I first need to explain you this type of requirement and this type of requirement I am going to explain you when I try to take the say test of hypothesis that is my next topic and possibly the last topic of this course. So, I am not covering those topics over here but if you wish if you know about this confidence interval, then you can go through with the book they are very simple and I will show you that they can be obtained exactly in the same way as you have done this confidence interval.

Now, after this, I have taken the example only for the normal distribution but you can construct a such confidence interval for any probability density function or probability mass function. But you have to be just be careful that when you are trying to deal with the discrete random variable like as if you want to find out the confidence interval for a population proportion using the binomial distribution then you have to take care of the continuity correction.

And there is another thing, that in this case we were able to find out the virtual quantity whose distribution is exact but many times finding out the exact distribution may be difficult then we try to use the standard normal curve after transforming the statistics that $\bar{X} - \mu$ divided by σ by root n type of transformation and we try to compute the confidence interval based on that approximated normal 0, 1 distribution. So, whatever concepts we have used, whatever things we have learned for such approximation they have to be implemented here also because you are simply trying to compute the probability.

If you try to see many times you have computed the probability that X is lying between say this X_1 and X_2 and you transformed it to a standard normal variate and then you computed the probability using the N(0,1) distribution. So, the type of calculation what you have done in the confidence interval estimation they are the same. So, whatever knowledge you have learned which you have gained while doing those exercises that is going to be used here also. I do not need to repeat it again and again. So, now I will request you that you try to take some examples from the book, try to solve them, try to practice them and try to see how you can employ them in the R software.

Now, you can see here that the use of the software will come only when you know what has to be done and what is the correct expression, what is the correct tool that has to be employed to get the correct answer of a problem. So, that is why this theory is very important that is gives you the basic fundamental by which you can decide what exactly you want to do.

Now, once you decide that this is what I want, now there will be challenges in the data science that if the data is too big how are you going to handle it and those types of complications for that we need the help of computer scientist also. So, I would request you, you try to revise this lecture, try to compute confidence interval for some other type of distributions also and I will see you in the next lecture with a new topic. Till then, good bye.