Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Sampling Theory with R Software
Lecture - 23
Simple Random Sampling
Estimation of Mean, Variance and Confidence Interval in SRSWOR using R

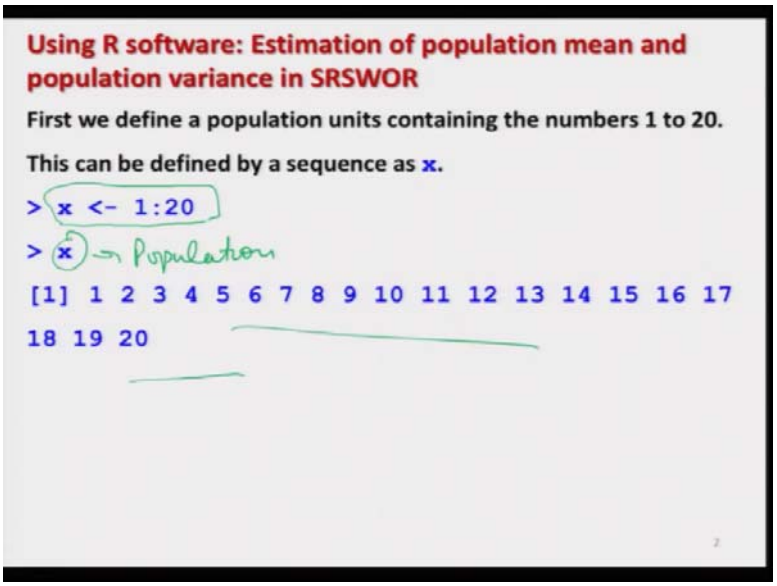
Hello, welcome to the course Essentials of Data Science with R Software 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. And, in this module, we are going to continue with the topics of Sampling Theory with R Software and we are going to consider the simple random sampling with R software.

So, up to now we have considered the topic of simple random sampling and we have concentrated on the estimation of population mean. We have found its point estimate; its variance, estimate of variance as well as the confidence interval. Now, the question is how to estimate them on the in the R software.

Well, in R at least to the best of my knowledge the packages to draw samples are available, but directly computing the estimate of variance confidence interval is are not available. So, we need to write down a small function or a small command using the basics of R that we have learnt earlier.

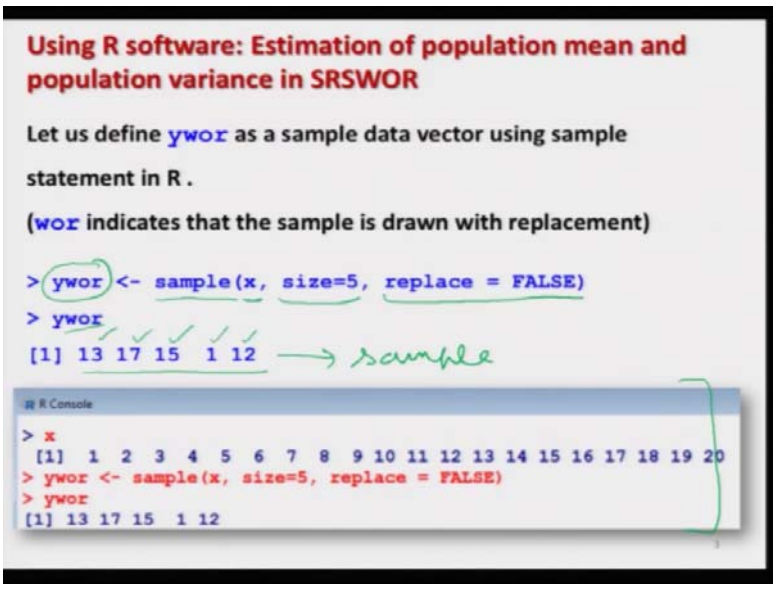
In this lecture, I am going to concentrate on the simple random sampling without replacement and in the next lecture I will try to repeat the same things with the simple random sampling with replacement ok. So, let us try to start our lecture.

(Refer Slide Time: 01:55)



So, just for the sake of simplicity what I am doing here that I am trying to create here a population. So, I am trying to create here a population of numbers 1 to 20; 1, 2, 3, 4 up to 20. So, this can be done by this command, x is equal to 1 colon 20. So, now, this x is going to denote my population. So, you can see here these are the values.

(Refer Slide Time: 02:27)



Now, you had learnt earlier that in case if you want to draw the simple random sample of a given size from this population, then we can use any one of the packages sample or

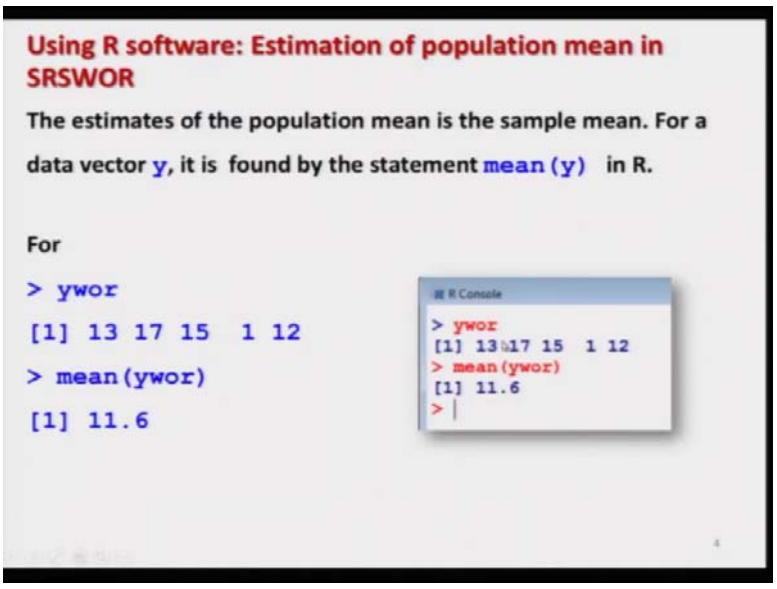
sampling, but here I am using sample and similarly, you can use sampling also that is the same thing because the job of sample or sampling package is limited to drawing of the samples only.

So, in case of wor, I am using the command here sample and this is my here population x and I am trying to draw here a sample of size 5. And, since this is simple random sampling without replacement so, I am using here replace is equal to FALSE. And, whatever is the outcome this I am storing here in a variable ywor. So, that will indicate the sample on sample values with without replacement, ok.

I am not using here wor because just to make sure that population and sample they are denoted by different symbols without any confusion. So, you can see here the sample drawn here is like this; 13^{th} unit, 17^{th} unit, 15^{th} unit, 1^{st} unit and 12^{th} unit and you can see here the screenshot also.

So, this is my here sample. Yeah, obviously, as I discussed couple of time earlier also, when I will try to do the same thing on the R console then you may not be able to get the same sample. But, whatever expressions we are going to discuss, whatever methodologies we are going to discuss they will remain the same, right.

(Refer Slide Time: 04:17)



So, now first I discuss computation of mean. Now, at this stage I can explain you very clearly that why theory is needed. For many people statistics is like just clicking on the

software which is not. Now, I am asking you one thing- you have drawn the sample and your objective is to infer about the population mean.

What are you going to find; the arithmetic mean of the values, geometric mean of the values, median of the values, mode of the values or what? Now, here comes the foundations and theory of statistics which connects the classical statistics or the theoretical statistics with the data science.

Now, to a data scientist, the knowledge of statistics is giving the information. Well, if you try to use the arithmetic mean of the sample values, then you will get a very good estimate. When I say very good estimate that is in the statistical sense. For example, in most of the cases, the sample mean turns out to be an unbiased estimator, an efficient estimator, a consistent estimator, sufficient estimator and in most of the cases, the complete and sufficient statistics also at least under the normal distribution.

So, but that is the job of those people who are working in the theory of statistics. They try to look into these statistical properties which are from the theoretical point of view and they try to establish that whether this estimator is going to give us a good outcome or not.

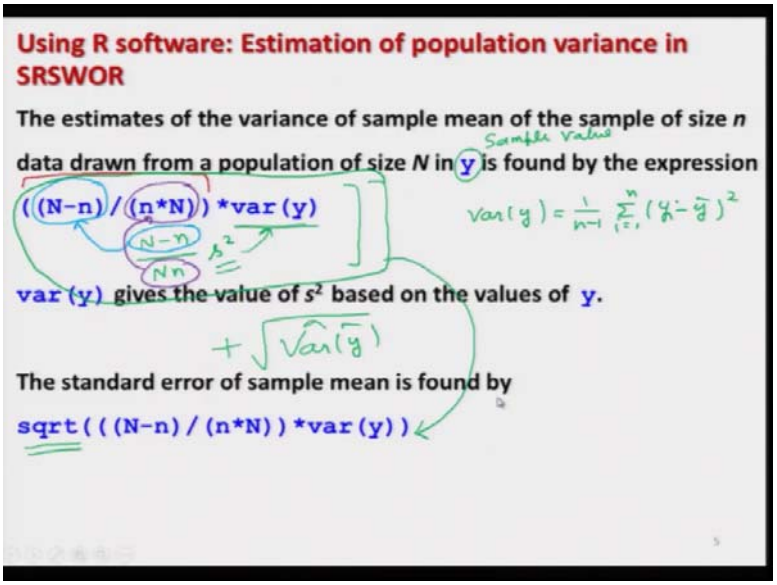
So, now the statistics has played it role and it has told us that you please try to find out the sample mean by arithmetic mean. So, now here there are two things in R. There is a built-in command to find out the mean values which is m e a n – mean and there is a built-in command to find out the variance also v a r.

But, now here the question comes, this syntax var is giving us what, σ^2 or S^2 - type of sample variant; that means, if I try to compute the variance of the values are they going to be $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ or $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$?

So, this may change from software to software. So, at least in R, I am telling you that this will give you the value with the divisor (n – 1). So, that is essentially giving you the value of s^2. So, this is what you have to keep in mind. But, definitely if you are trying to work with any other software you have to go to the help and try to see what the variance command is giving in the out outcome.

Well, I agree if your sample size is pretty large then whether you are taking the divisor n or say (n – 1) that will practically not make much difference, but at least you must know that what is happening ok. So, now I am going to use two commands here mean; and variance that is mean and var to find out the estimated values of population mean and its standard error ok. So, let us begin.

(Refer Slide Time: 08:32)



So, we assume that the sample values are stored in the data vector y and you can recall that the expression for the variance of $\bar{y}$ under SRSWOR was (N-n)/Nn and see here s^2. That was the estimate of variance under this.

So, now what I am doing here? I am using here a simple R command to write down this thing. So, this factor here N - n this is written here and then the quantity in the divisor, quantity in the denominator N into n; this is mentioned here under this thing.

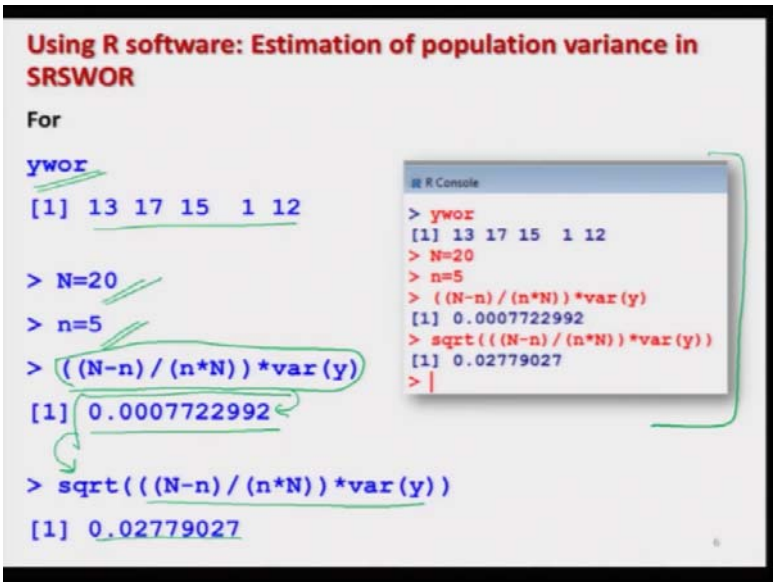
So, this whole quantity from here to here, this is trying to give us the quantity (N-n)/Nn and now, we want to compute s^2. This s^2 can be directly computed by the command var y because in R ,variance of y this is var of y this computes the quantity $\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

So, I can use here the direct command and from here I can compute the value of s^2, right. So, this will give us an estimate of the variance of $\bar{y}$ under SRSWOR. In case if you

want to find out the standard error so, we know that the standard error is defined as the $+\sqrt{Var(\bar{y})}$ estimator.

You estimate the variance of $\bar{y}$ and take its positive square root. So, what I have to do? Whatever expression I have written here, I simply have to take its square root. So, that was exactly I am writing here. In order to find out the square root we have a built in command in R, s q r t, right. So, you just try to take the square root of the value which we have obtained earlier and this will give you the standard error, ok.

(Refer Slide Time: 11:26)



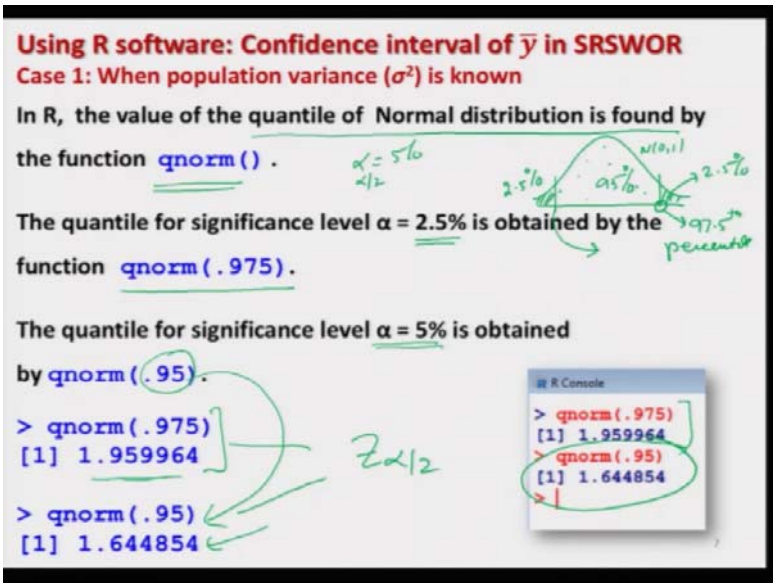
So, now, first we try to execute all the things. So, our data vector y is denoted here as say ywor which has these values 5 value; 13, 17, 15, 1 and 12. The N which is the population size that we have already fixed to be 20, the sample size we already have fixed to be 5.

So, now I try to use the same command or variance of $\bar{y}$ here and this will give me the value here like this ok. And, now in case if you try to find out the square root of this quantity this will give you the standard error. So, square root of this quantity here this is 0.02.

So, you can see that it is pretty straight forward and simple to find out the mean and variances or standard errors in the case of simple random sampling in R and here is the

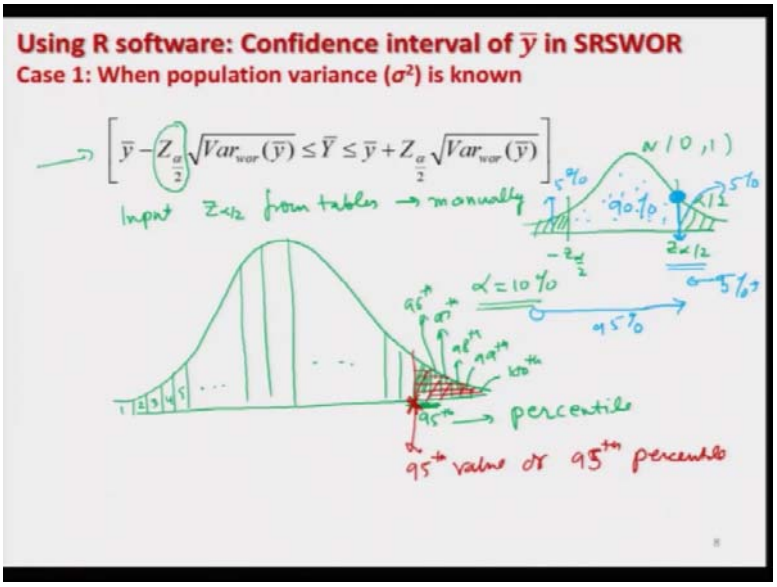
screenshot which you can verify that these are the same thing. But, as I said, these are the results when I conducted this computation while during the preparation of my slides, but now when I will do it on the screen it will give me a different result.

(Refer Slide Time: 12:42)



Now, when I. So, now, we consider the confidence interval estimation.

(Refer Slide Time: 13:03)



So, if you remember the confidence interval that we computed earlier in case of SRSWOR was given by this expression and you can see here there was a quantity here $Z_{\alpha/2}$. What was $Z_{\alpha/2}$? It was the upper $\alpha/2$ percent points on the distribution of N(0, 1).

Now, when you are trying to construct the confidence interval in R, then you have two options. First option is this that you try to input the value of $Z_{\alpha/2}$ from tables, but you have to do it manually and you have to write down this value. Second option is this that this value can also be computed directly on the R software, how?

If you try to understand what is this value $Z_{\alpha/2}$ on this table. For example, if I say that I am trying to take here α is equal to suppose 10 percent. So, I am trying to say here that in the case of confidence interval I have here two values - $Z_{\alpha/2}$ and + $Z_{\alpha/2}$ and the α is 10 percent. So, 5 percent points are on the left hand side and 5 percent points are on the right hand side.

And, the area in between here this is, 100 - 5 - 5 which is 90 percent and you are trying to find out the value of here Z, which is essentially the value on the normal curve corresponding to which whatever is the value of, value is on the x-axis. So, $Z_{\alpha/2}$ is the value on the x-axis which is dividing the total area under curve into 2 parts such that this part is 5 percent and this part here is 95 percent.

Now, if you try to look at this concept from the conceptual point of view means, I can say here that suppose I try to divide the area under the curve under the N(0, 1) into 100 equal parts. 1 say here 1, 2, 3, 4, 5 and here you can see here like this and then somewhere here like this and suppose this is the last part 100th part.

So, this is 100^{th} part, this is 99^{th} part, this is 98^{th} part, this is 97^{th} part, this is 96^{th} part and this is here is the 95^{th} part. So, here this is the value here which is trying to give us the 95th part. What is this 95^{th} part? In statistics, we call it as 95^{th} percentile. When you try to divide the total area into four equal parts it is called as quartiles.

When you try to divide the total area under the curve into 10 equal parts this is called as deciles and similarly, if you try to divide the total area into 100 equal parts then every part is called as percentile. And, what are you doing? You are trying to say here that the area on the right hand side of this point that is here it is 5 percent.

So, you would like to find out the value of on the x-axis corresponding to which this area is 5 percent. So, this is essentially the 95^{th} value or this is the 95^{th} percentile. This is the 95^{th} value if you try to divide the total area into 100 parts. So, now in R we have a facility that we can compute the percentile.

And that is also an advantage that if you try to see in the classical books usually people are trying to construct only 99 percent confidence interval, 95 percent confidence interval because it was not possible earlier in the earlier days to compute all those values manually.

So, people have created the table for given choices of α, but now with the R software you can do anything. You can choose any value of α and you can compute the quantiles or the percentiles. So, in R if you want to compute any quantile of the normal distribution then the command here is qnorm and inside the bracket you have to give the value of this α or say $\alpha/2$ whatever you want.

So, for example, if I try to take here α is equal to 5 percent, then we are looking at the $\alpha/2^{th}$ percent. So, in general I can say that we are trying to take here α to be here 2.5 percent. So, that means, this area under the N(0, 1) is being partitioned into 3 parts.

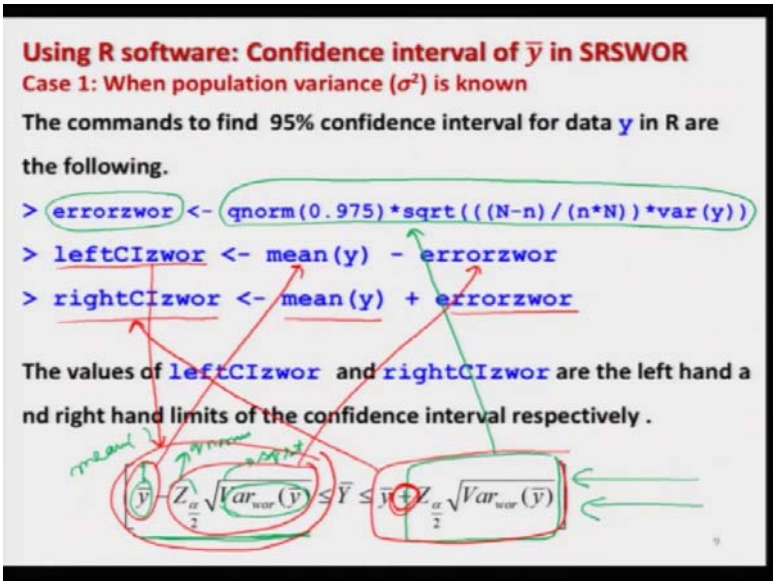
On the left hand side this is 2.5 percent; on the right hand side it is again 2.5 percent, these 2 shaded area and the dotted area in between is 95 percent. So, essentially this value here this is the 97.5th percentile and now, if you want to know this value on the left hand side you have two options either you try to take the minus sign of the 97.5 percentile or you try to simply find out the percentile at 0.025 right, ok.

So, now in case if you try to see here I can find out here the this function by qnorm 0.975 and if you try to use here this thing you can see here this value is coming out to be 1.9599 and you can see that in most of the books it is usually approximated at 1.96, right and this is here is the screenshot.

And, similarly if you try to take here α to be 5 percent or so, then in that case you have to give here the value of $1 - \alpha$ which is 0.95 and corresponding to this the percentile the 95th percentile will come as 1.644 and this is here the outcome and this is here the screenshot.

Now, you have got essentially these values are the values of $Z_{\alpha/2}$. So, now, I would suggest you that why should you go back to table, why do not you use this command and compute the lower and upper limits of the confidence interval.

(Refer Slide Time: 20:58)



So, you can recall that your confidence interval is given by this expression in the bottom of the slides, try to concentrate right. So, this is the expression. So, now what I have to do? That I need to compute the lower limit and upper limit manually. So, if you look at this expression of the lower limit I can compute the first quantity by here the function mean.

This quantity here Z, this I can denote by qnorm and this square root I can use the command sqrt and this command here variance of $\bar{y}$ this we already have actually computed earlier as ((N-n)/Nn) s^2 or the variance of $\bar{y}$, right. So, I have to simply write down here a small function for these commands.

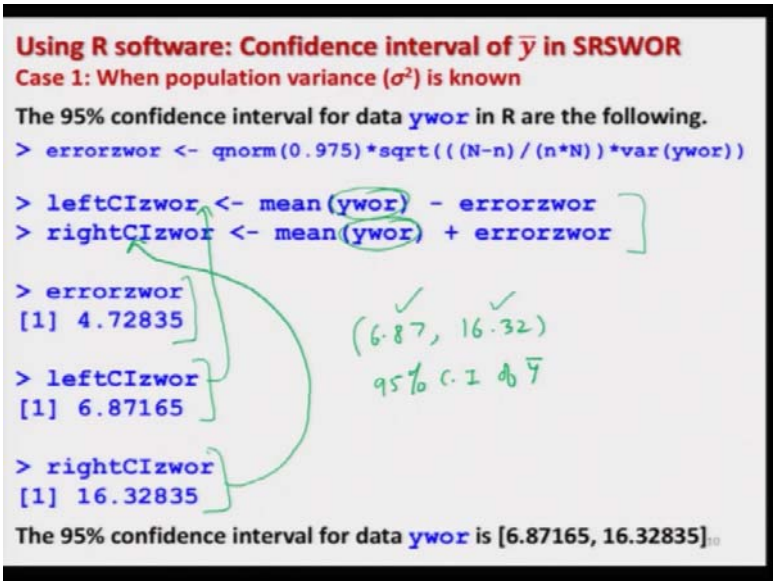
So, first I try to compute this value; $Z_{\alpha/2}$ $\sqrt{Var(\bar{y})}$ $\bar{y}$ under wor. So, because this is common in the left limit and right limit. So, it will help us in a simple representation. So, I try to compute here this quantity say qnorm and variance of $\bar{y}$ under SRSWOR that we already have actually computed, right and let me call this quantity as errorzwor, right and

once I have computed this quantity, then I can compute the left limit which is here like this.

So, you can compute here. This $\bar{y}$ will come here mean of y and this quantity come it comes here errorzwor. So, this is mean of y - errorzwor. So, this will give you the value of lower confidence limit. So, this entire limit will be computed by this function and once you have computed the lower limit, the upper limit goes exactly in the same way because the difference is only of the plus sign here.

So, this upper limit this is being computed here in the function rightCIzwor. So, which is nothing, but, here the same command as above but with mean of y plus errorzwor. And, these are the lower and upper confidence limit or these are the upper or lower confidence limit for 95 percent confidence interval.

(Refer Slide Time: 24:00)

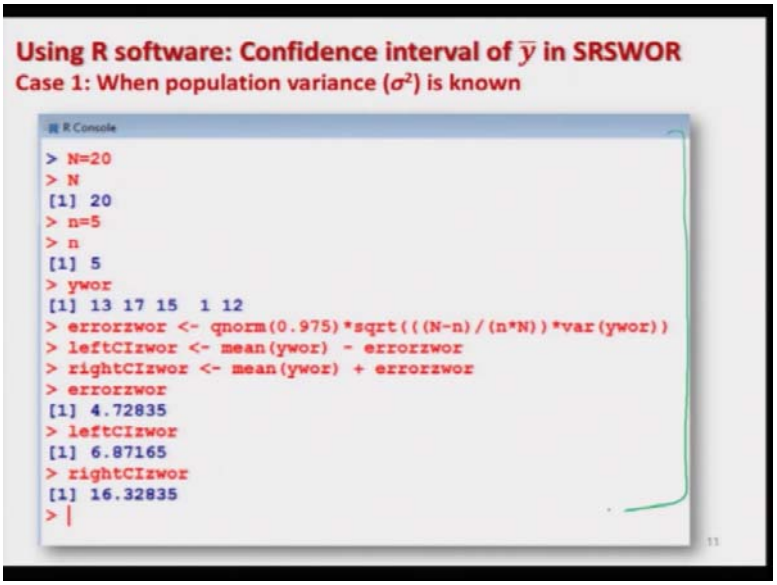


So, now let us try to compute it for the given sample of data. So, we already have this data and we try to use this command over here. So, right. So, if you try to see I have written here the same command, but now I have given here the sample values which we are earlier stored in the data vector ywor and I tried to compute just to show you the part which was $Z_{\alpha/2}\sqrt{Var(\bar{y})}$.

And, then the left confidence interval can be computed using this command and right confidence limit can be estimated can be computed using this command. So, now if you try to see this interval becomes 6.87 to say 16.32 approximately. So, what is this? This is your actually here 95 percent confidence interval of $\overline{Y}$.

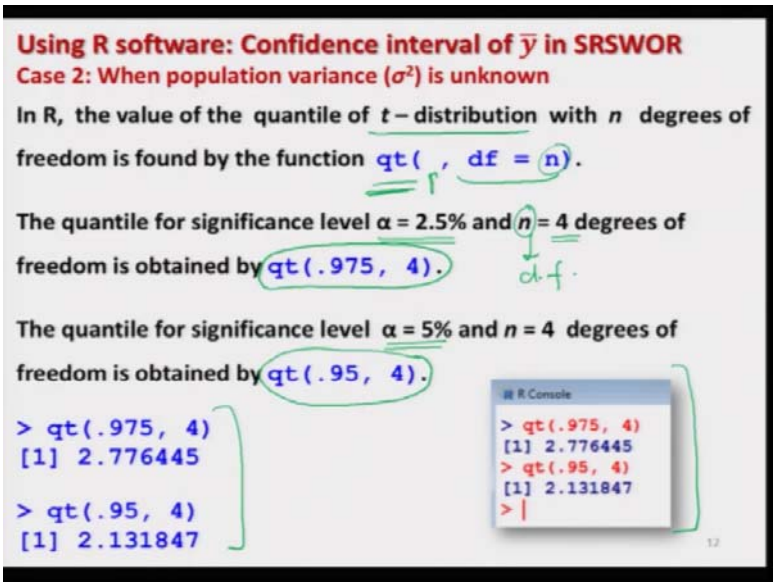
So, what are you trying to say? That if you try to find out the confidence interval for the unknown values of $\overline{Y}$ using the sample mean, then there are 95 percent chances that the estimated value will give you the values of $\overline{Y}$ lying between 6.87 and 16.32. So, you can say in general in a common language that the population mean will lie between 6.87 and 16.32, right.

(Refer Slide Time: 25:37)



And, this can be clubbed together with your standard error mean whatever you want. And this is here the screenshot of the same thing what I have done here right.

(Refer Slide Time: 25:47)



And, when σ^2 is unknown to us then as we have discussed during the construction of the confidence interval we use t – distribution. So, similar to how we have computed the Z value using the quantile function, similarly we can also compute the quantiles of t-distribution and the commands to compute the quantiles of t distribution is qt. And, we have to give here the value of square is α and we have to define the degrees of freedom.

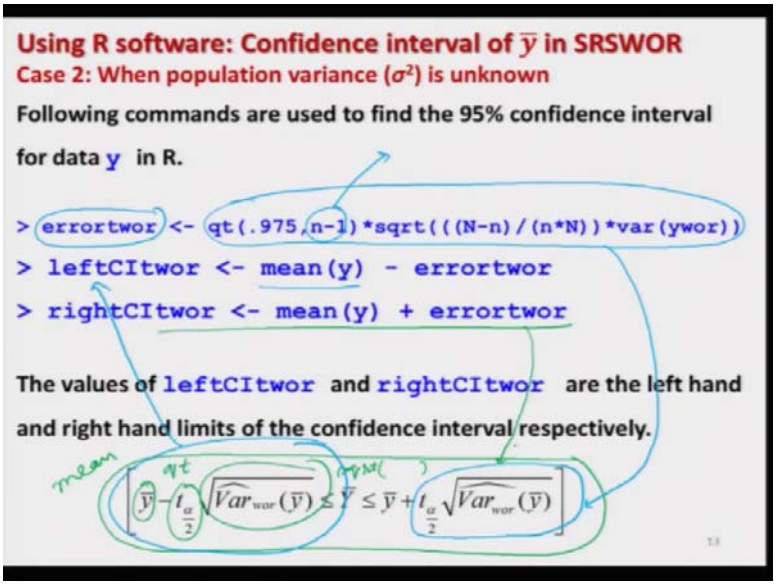
So, remember in during the confidence interval estimation means our degrees of freedom were n - 1. So, here in this command we have to give the degree of freedom as here n. So, you have to be careful that you give here the value of n - 1 while computing the confidence interval.

So, suppose if I take α to be 2.5 percent and our sample size was 5. So, these degrees of freedom this here n is actually degrees of freedom, this becomes 5 - 1 equal to 4. So, the 97.5 percentile of t-distribution with 4 degrees of freedom can be found by this command. And similarly, if you try to change your α to 5 percent, then the 95th percentile with 4 degrees of freedom on the t-distribution will be found by this command.

So, actually this is central tree. There are two types of t distribution central t and non-central t, but here this gives you a you the central t. So, the quantiles are computed on the central t-distribution and the values of the quantiles for this α to be 2.5 percent of 5

percent they are obtained here like this 2.77 approximately and 2.13 approximately. And, this is here the screenshot of what you will obtain, right.

(Refer Slide Time: 28:02)



And, now in case if you try to recall if you try to look in the bottom of the slide here this was your here the confidence interval when σ^2 was unknown. So, now here one can see here that there is here one quantity which can be computed using the function mean.

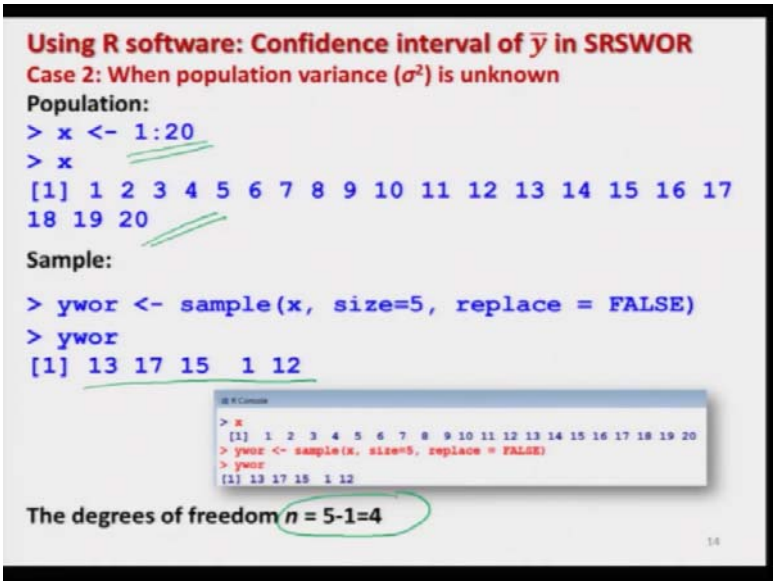
This t this can be computed using the function qt and this function can be using the function square root of t. and we already had computed the variance of $\bar{y}$ under wor when σ^2 is unknown, right. So, I have followed exactly the same concept which I just use for the normal distribution.

So, I am writing out here the this part $t_{\alpha/2}$ $\sqrt{Var(\bar{y})}$ or wor $\bar{y}$ is estimator here by this errortwor. So, this quantity here is giving you here like this and yeah, I mean just to be on the safe side. Because I am using the sample size so, just to avoid any confusion I already have written here n - 1.

But, here you have to be careful that how are you trying to write down this function. So, this function will compute this part and if you would try to see the lower limit is given by this part which is computed here just say mean of y - error. So, now you can see here we

have computed the lower limit as well as here right confidence limit also here. So, it is not difficult you can see here.

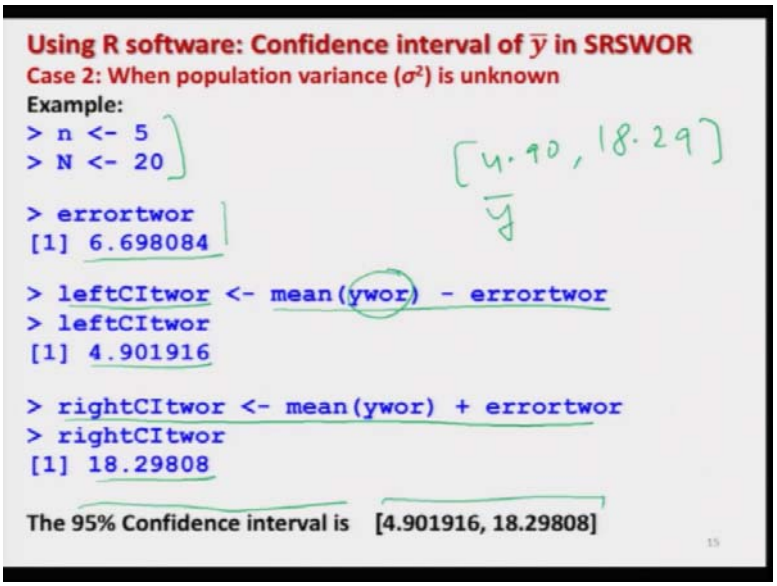
(Refer Slide Time: 29:51)



And, now I try to execute it on the R console and for that this is just for your quick review that we had considered the population to be a values from 1 to 20 and our sample was 13, 17, 15, 1, 12. And here now, the degrees of freedom will be 5 - 1 which is here 4.

You have to be careful with the symbol n and I have used to denote the sample size, but in case of t that is also the standard symbol. And so, I have not a used any other simple otherwise that will create more confusion, but that is going to happen only in case when we are trying to use the t-distribution that is actually the degrees of freedom.

(Refer Slide Time: 30:35)

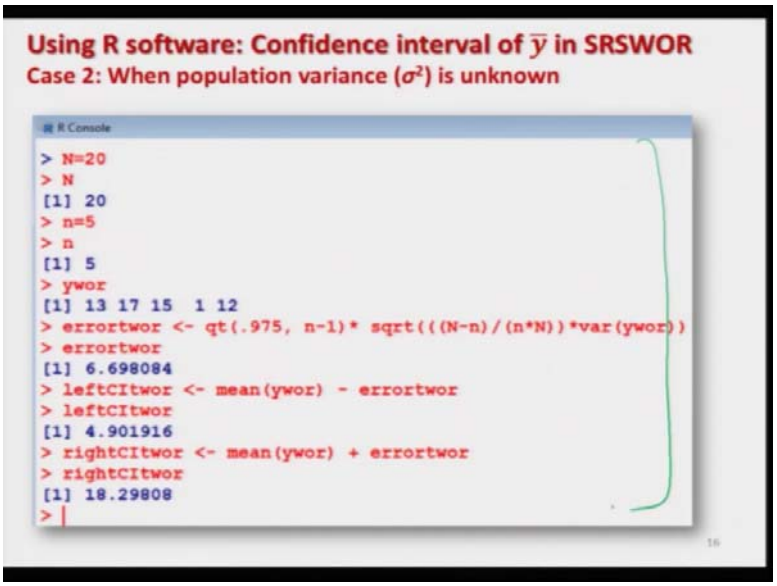


So, now when I try to execute these commands on the R console so, my n is 5, N is 20 and then I try to compute the expression for errortwor which comes out to be 6.69 approximately. And, then using the expression of left confidence interval which was here like this I try to put here the values of the data vector ywor and then the left confidence interval comes out to be like this.

And, similarly I try to write down the expression for the right confidence interval and this interval comes out to be say approximately 18.29. So, in this case the 95 percent confidence interval for Y bar is obtained as say approximately say 4.90 to 18.29. So, what is the meaning?

The interpretation and the meaning is the same as in the earlier case that there are 95 percent chances that on the basis of this sample when you try to estimate the population mean using the sample mean, then there are 95 percent chances that the estimated value will indicate that the where the population mean is between 4.9 and 18.29. So, this is the 95 percent probability of this event.

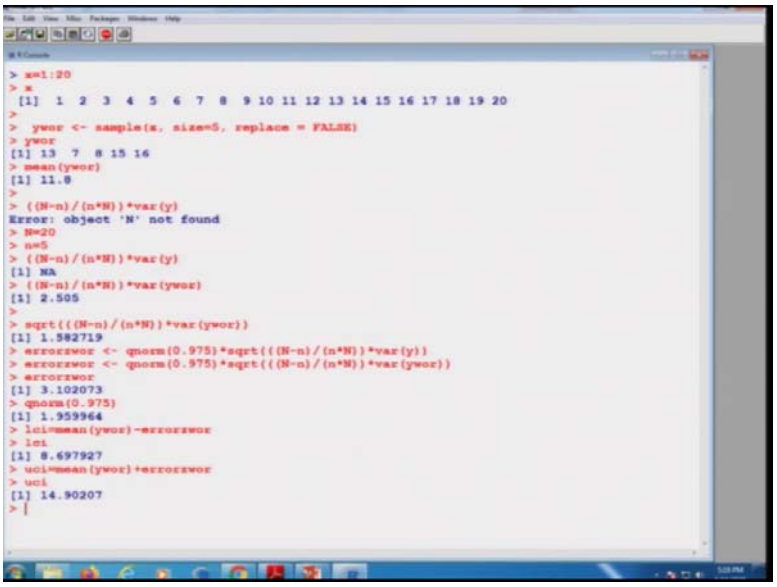
(Refer Slide Time: 32:06)



So, you can see here now you can give all this interpretation and this is here the screen shot of the same thing which I shown you. Now, I will try to do the same calculation on the R console also, so that you get confident that whatever I am doing here this is correct. So, first I come to beginning. I will define my population and then sample.

So, first I try to define my here sample. Well, this sample which I will get here that is going to be different from the sample which is reported here 13, 17, 15, 1 and 12.

(Refer Slide Time: 32:39)



You can see here that I am trying to give here x as population 1 to 20 values. So, this is here x. Now, when I try to compute this sample use the sample is here you can see here it is different 13, 7, 8, 15 and 16, but that does not make any difference for me. So, in case if you try to find out it is mean and see here it is a standard deviation standard error.

So, I will say here first I try to find out here the mean of here ywor, this comes out to be 11.8 and if you try to find out it is variance this comes out to be here something error. Why? Yes, you have not given the value of the population and the sample. So, now I am giving it here and the values of the variance of $\bar{y}$ under SRSWOR comes out to be here like this, but why it is giving you NA?

Is it not available? Yes, these things will happen if you simply try to copy and paste the command because you have not given here the right data vector. So, your data vector here is ywor. Well, I am trying to show you that; that if you try to do it yourself on the R console what type of possible mistakes you can make. And you can see here that this is 2.505 and if you want to find out the standard error.

So, I simply have to take the square root of the same function and this will come out to be 1.58 and so on. And, similarly if you go for the confidence interval estimation, so, I will try to find out here the first I try to compute the error part. So, you can see here, but I have not I made a mistake because it has to be ywor.

So, I try to give it here and you can see here error zwor comes out to be here like this. And yeah, means here I can also show you that how this qnorm will look like. So, you can verify that. This is going to be the same because this quantile will not change wherever you do it, right. So, you have obtained here this thing.

And, your the lower limit the I can see give it here a new name lci lower confidence interval that will be here mean of mean of ywor - error zwor. So, you can see here, this is here lower confidence interval. So, this is the lower confidence interval of the sample that you have drawn.

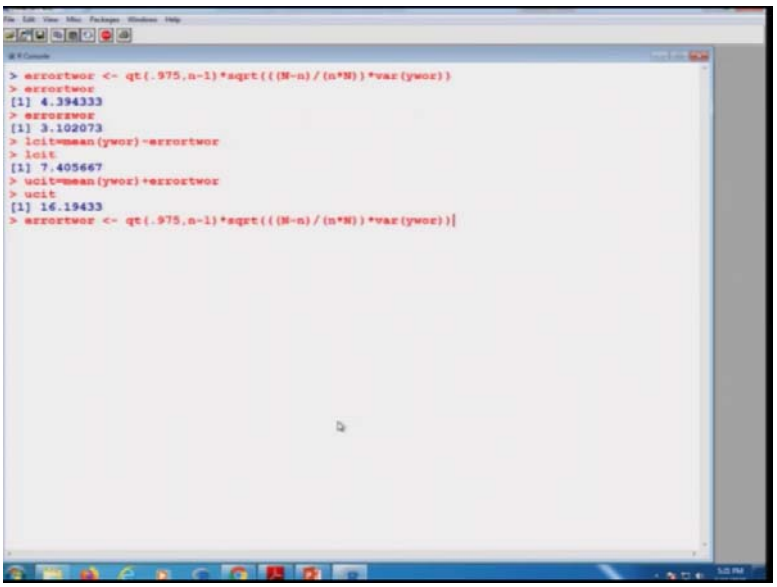
And, if you go for the upper confidence interval I simply have to make here the - sign to be the plus sign. So, upper confidence interval will come out to be like this. So, you can

see here that 8.69 to 14.97. So, you can now see actually if you try to compare this result with the result what I shown you here.

You can see here these are the values of left hand confidence interval and right confidence interval. You can see that that which you do that even if you try to change the sample these values are not differing much. Why? Because, you have said that that there is a 95 percent probability that these values will lie between this confidence interval, right.

Now, let me try to show you here the confidence interval based on t. So, I try to first compute that key part here. So, you can see here. Well, I can clear the screen.

(Refer Slide Time: 37:08)

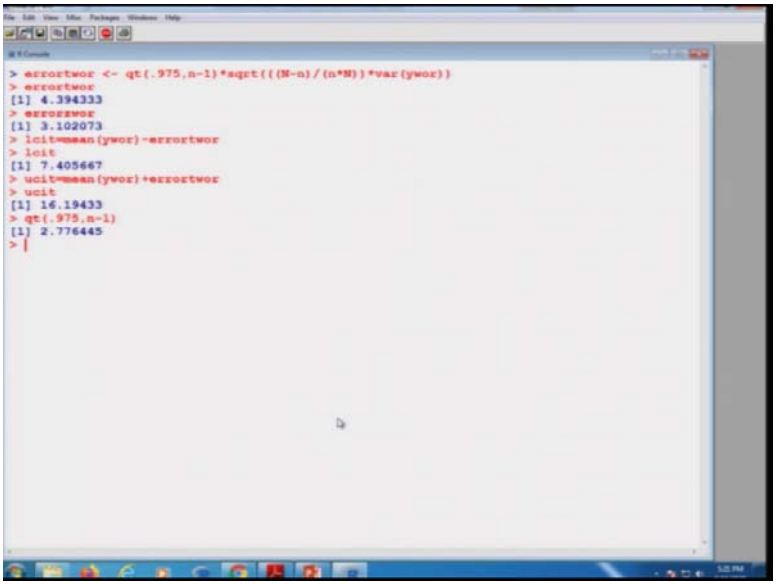


So, you can see here the precursor is you are here. Error part: so, you can see here, this is errortwor and this is different than what you had obtained. Earlier in the case of Z you can if you want to compare it with the with here Z you can see here this is like this. So, this is this is changed. The mean $\bar{y}$ will remain the same in both the cases.

So, in case of this if you try to find out the lower confidence interval of t, so, I can define it here as a mean of ywor - error twor. So, you can see here this is your here lcit. And, if you want to define here the upper confidence interval using the t distribution when σ^2 is

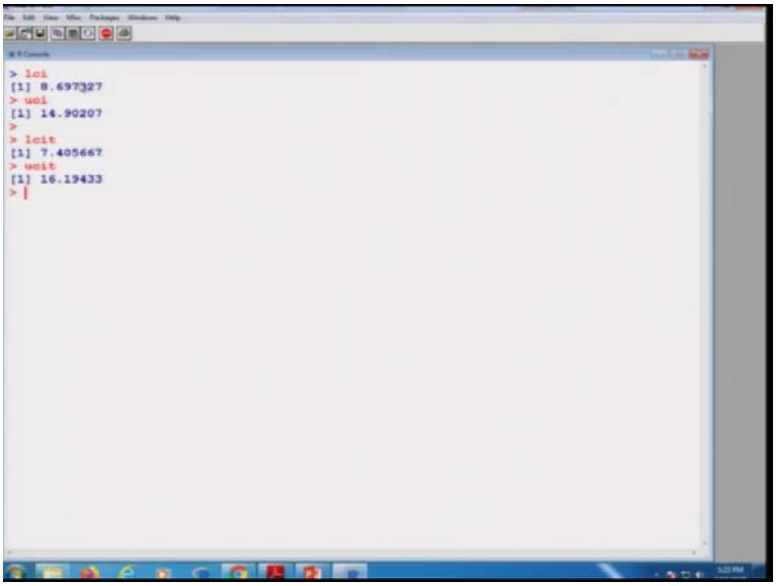
unknown, so, I can just change the minus sign to plus sign and you see here, this is here ucit, right.

(Refer Slide Time: 38:23)



And, now, if you want to see that how the this quantile of t will look like so, you can see here this is will give me the quantile of t 2.77.

(Refer Slide Time: 38:28)



Now, I try to show you something more. If you try to see here when you assumed the σ^2 to be known, then your lower confidence limit and upper confidence limit came out to be

like this 8.69 to 14.97. But, when you assume the σ^2 to be unknown and you computed it on the basis of your sample, then your confidence interval comes came out to be like this.

So, now you have here 2 confidence interval; one is here like this which is between 8.69 to 14.9 and second confidence interval when σ^2 is unknown is 7.4 to 16.1. Well, if you see this confidence interval is shorter; that means, the width of the confidence interval when σ^2 is known is shorter, then the confidence interval when σ^2 is unknown to us.

Do you think is it normal? Well, that is normal here you are assuming that σ^2 is unknown in the first case, but now you are saying that this is unknown. So, once a parameter is unknown you are trying to estimate it on the basis of the random sample. So, definitely it will incorporate some more uncertainty and that is why the confidence interval based on t will be wider.

The width of the confidence interval based on the t distribution will usually be wider than the corresponding confidence interval under normal case, right. So, now I stop in this lecture. So, I have given you the idea that how are you going to find out the mean; how are you going to estimate the variance; how are you going to find out the standard errors and how are you going to find out the confidence interval when σ^2 is known and unknown in the R software.

So, now I would request you that you try to take some data set maybe a bigger data set or a smaller data set and then try to practice these commands. My advice will be try to take a small data set first. The advantage is that you then you know means you can also compute the population mean at least in the in this situation.

In practice, it will be unknown to us and then you try to see that once you are trying to find out different samples, then how the sample mean is differing, how their standard errors are differing, how their confidence intervals are changing. And you yourself can make certain experiments that what will happen if the values in the samples are varying a lot.

You can take a sample of size 5, say you can take it artificially also just choose 5 values 1, say 5, 20, 50, 100 and try to choose another set of values which are very close to each

other, try to find out their mean, variance, standard errors, confidence interval and try to see what happens.

So, practice with this smaller example will give you an insight and that will develop a sort of feeling that when you are trying to expose these results on a bigger data set, then what is really going to happen. This is the same story that if you try to put your fingers on the forehead you can know that if the body temperature is normal or higher or the person has got a fever or not. You are not using any thermometer for this.

But, you are trying to use your knowledge because you have developed a sort of insight that if the body temperature is feels like this one then the person is person does not have fever or has fever. Similar is the story with data science. These tools are used on some unknown population on the bigger data sets where we cannot use our hands to test the forehead, but we have to depend on the thermometer.

So, what I am asking you? Put your hand on the forehead and try to decide whether the person has fever or not and then try to put the thermometer inside the mouth and see whether thermometer and your information are they matching. So, if you try to practice with the smaller data set you will gain more confidence and in the longer run you will have an idea what is really happening when you are trying to deal with bigger data sets in data sciences.

So, this is the interconnection. So, that is the first time I am trying to show you or I am trying to explain you that how the classical statistics is playing a very important role in the data sciences which is the newly born baby which born couple of years back, where the statistic born many many years back. So, you try to practice and I will see you in the next lecture till then good bye.