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

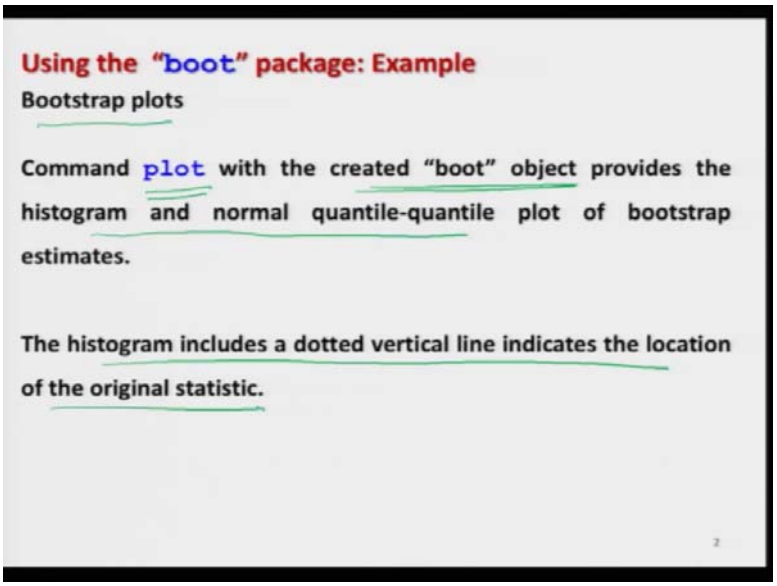
Sampling Theory with R Software
Lecture - 37
Bootstrap Methodology
Bootstrap Confidence Interval Using boot and bootstrap Packages in R

Hello friends, 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. In this module on Sampling Theory with R software, we are going to talk about the Bootstrap Methodology with R Software.

So, you can recall that in the earlier lecture, we had considered the construction of Confidence Interval Using the bootstrap methodology and we had considered five different types of confidence interval. Now, in this lecture, I am going to use the R software and I will demonstrate that how you can obtain those confidence interval and for that, I will be using here two packages; one is boot and say another is bootstrap.

The reason is the following, that out of those five confidence interval, four confidence interval can be obtained by the package boot that we have used earlier also and for the fifth bootstrap confidence interval, we have to use a different package. So, that confidence interval, I will try to show. That is actually based on t-distribution. Bootstrap t confidence interval, I will try to show using the bootstrap package, ok. So, let us begin our slides.

(Refer Slide Time: 01:49)



So, as I told you that when you want to construct the confidence interval using bootstrap methodology, first you have to check whether the distribution of the bootstrap values of parameter which you have obtained, do they have a smooth histogram, smooth density curve, smooth distribution, there are no spikes, there are no extreme values etc.

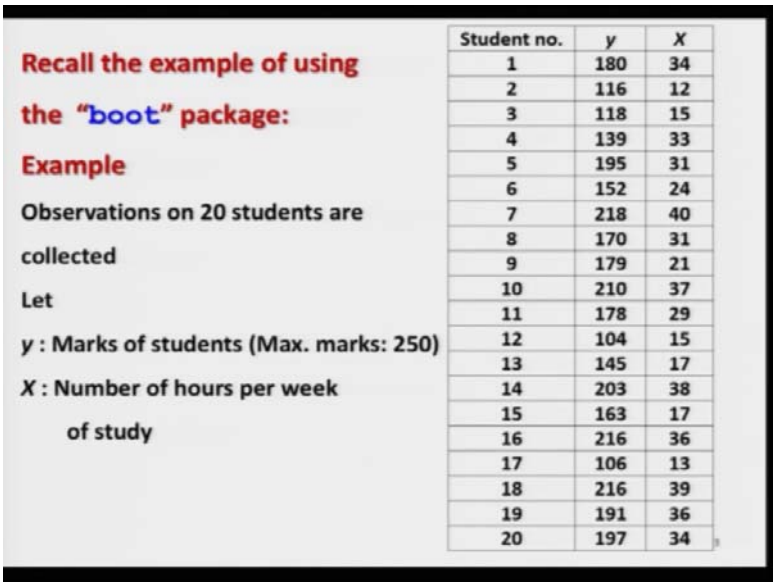
So, first step in creating the confidence interval is that you must plot those bootstrap values and as I told you that it is sometime advantageous to compare the distribution of bootstrapped values with the normal distribution. And if and as we know that in R software, there is a command Q-Q plot which provides us the quantile-quantile plots.

But here in this boot package itself, there is a command plot which creates both the things on the same graphic. So, first I will try to demonstrate the use of plot command and then, I will try to take the same example which I considered earlier for the estimation of bias and standard error and on the same data set, I will try to construct the confidence interval ok.

So, now, the first thing I am going to consider that how to construct the bootstrap plots. So, we are going to use here the command plot and this plot command has to be used on the boot object that was created earlier. For example, if you remember the outcome of the boot packages was package was stored in an object called boot, right.

So, using that information, we can create a histogram and a normal quantile-quantile plot of the bootstrap estimates. And this histogram includes a dotted vertical line that indicates the location of the original statistics that means the value of the statistics in the original sample.

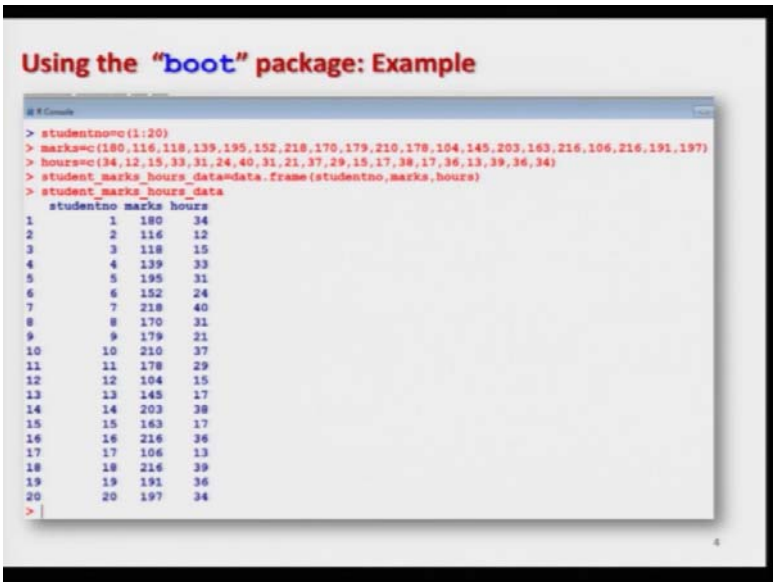
(Refer Slide Time: 04:07)



Student no.	y	X
1	180	34
2	116	12
3	118	15
4	139	33
5	195	31
6	152	24
7	218	40
8	170	31
9	179	21
10	210	37
11	178	29
12	104	15
13	145	17
14	203	38
15	163	17
16	216	36
17	106	13
18	216	39
19	191	36
20	197	34

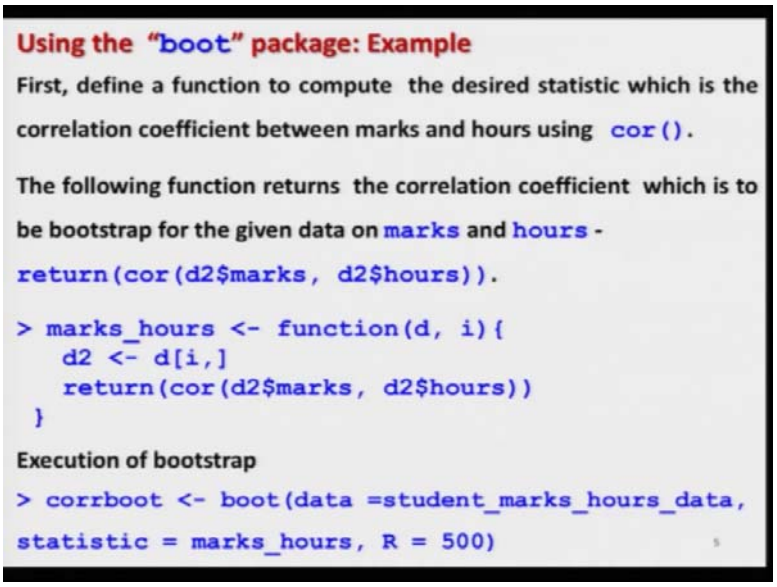
And so, just for a quick review, you can recall that we have use this example, where we had obtained the data on 20 students about the marks obtained out of 250 marks which are denoted by y and the corresponding values of X, which is indicating the number of hours per week of study and in this case earlier, we had found the correlation coefficient between X and y, right.

(Refer Slide Time: 04:36)



So, for that, we had created this data frame. So, I will not repeat these steps over here.

(Refer Slide Time: 04:47)



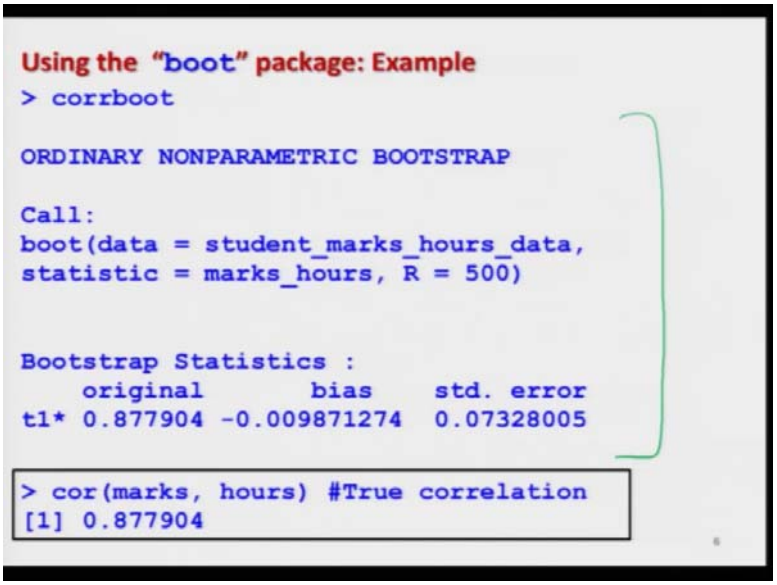
And in order to compute the bias standard error and to create an object using the boot package, we had used this slide. If you remember, I have simply copied the same slide over here so that it is not difficult for you to remember.

We had written there a function, which is going to compute the correlation coefficient between marks and hours and then, we had used the boot command to find out the

estimated value bias and standard error of the correlation coefficient between marks and hours.

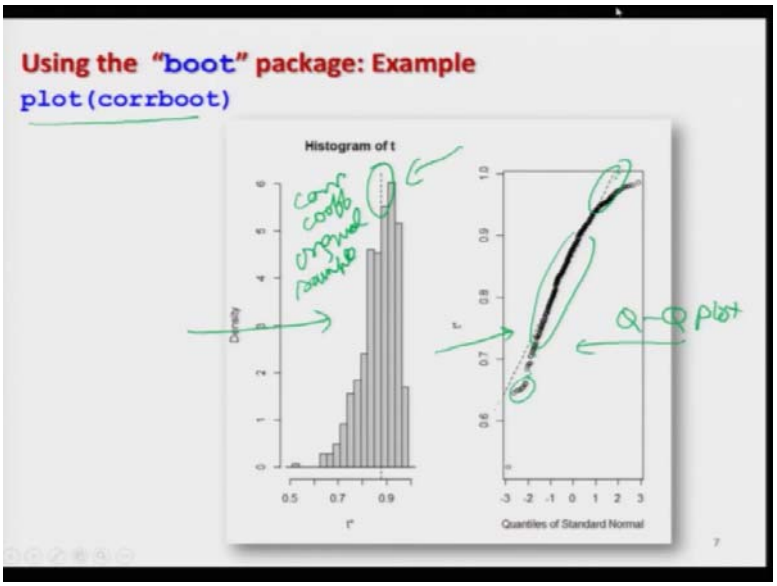
So, the correlation coefficient was computed based on the bootstrap sample and we had used 500 boot samples. So, for every sample that was drawn by SRSWR, the value of the correlation coefficient was computed and based on that, the bias standard error and the value of the estimator were obtained and those values were stored in the object corrboot; c o r r b double o t that is the name which we had given, right.

(Refer Slide Time: 05:53)



So, and this was the outcome which we had obtained. This is just for your recollection so that you do not have to go back to the earlier lecture; but, you have to remember all those things.

(Refer Slide Time: 06:06)

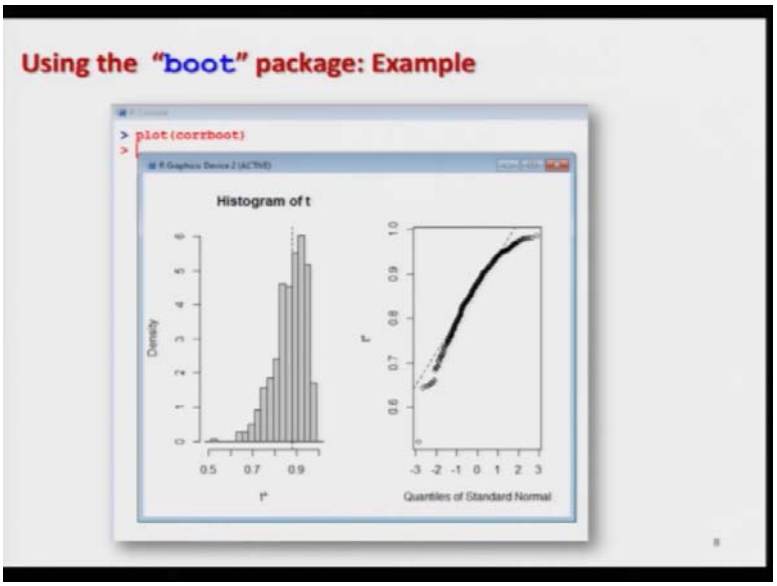


So, now, I try to use the command here plot on the object corrboot. So, I have to write plot and within the parenthesis c o r r b o o t. You can see here this is the type of graphics which you will obtain here. So, this is here the histogram and this is here the Q-Q plot, right.

So, you can see here, there is here is a dotted line which is trying to give us the value of the correlation coefficient in the original sample, right. And then, whatever are the values that we have obtained for the correlation coefficient in the 500 estimated values; based on those values, the normal probability plot is obtained here.

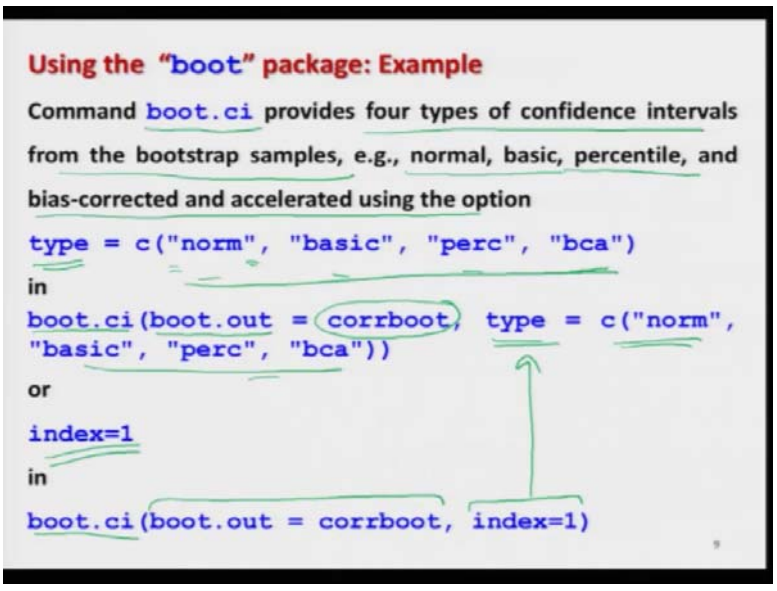
So, you can see here, I will remove this line here. So, you can see here all these points, points over here, they are lying very close to this dotted line, right. So, we can safely assume yes, they are coming from a normal distribution. In practice, you need some experience to interpret the quantile-quantile plots, whether you can safely assume the normality or not.

(Refer Slide Time: 07:35)



Anyway, so this is the screenshot. If you try to do it on the R console, you will get a similar thing, because when I will be doing possibly, the results are going to be little bit different although I have saved the results of the earlier simulation but they are going to be little bit different.

(Refer Slide Time: 07:53)



Now, I come to the construction of confidence interval. In order to find out the confidence interval, we have a command here boot dot ci; boot mean bootstrapping and

ci means confidence interval. Actually, this command provides four types of confidence interval that we have considered; except the t distribution base confidence interval, it provides us all other remaining confidence interval.

So, we get here the confidence interval based on normal distribution, basic confidence interval, percentile method as well as “bca” method, right and this thing can be controlled by using the option here; type is equal to I mean inside the parenthesis within the double quotes, you can state here the options which you want to apply.

Suppose, I want to apply; suppose, I want to have all four types of confidence intervals, so I am just giving all the names over here. So, “norm” is for normal confidence interval, “basic” is for the basic confidence interval, “perc” is for the percentile confidence interval and “bca” is the bca or the bias corrected and accelerated confidence interval and if you want to use only one option that is also possible.

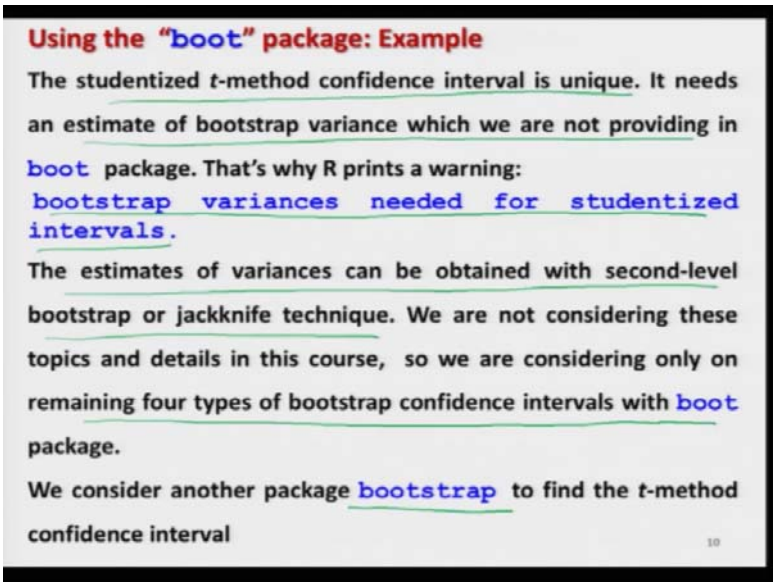
Instead of giving here all the four values, you can choose any value; but remember they have to be given as a string. So, they have to be given within the double quotes, ok. So, now, I use here the first I try to show you how are you going to use here.

The command here is boot dot ci and then, you have to give the value boot dot out, which means you have to give the outcome of the command boot or the boot object in which you have stored the outcome and you want to construct the confidence interval. So, we have stored our outcome in the boot object say corrboot; corrboot c o r r b o o t, right.

So, I am giving it here. Now, I am giving here, what type of confidence interval I want. So, I am just giving here all the four names, right. The other option is that instead of using because here this type equal to c non basic percentile bca etc., I can also use the option that index is equal to 1.

So, in that case, you have to simply write down here boot dot ci and boot dot out remain the same, the boot object. But once you write here index equal to 1 in place of a type, then you will get the same outcome; all the four confidence interval.

(Refer Slide Time: 10:41)



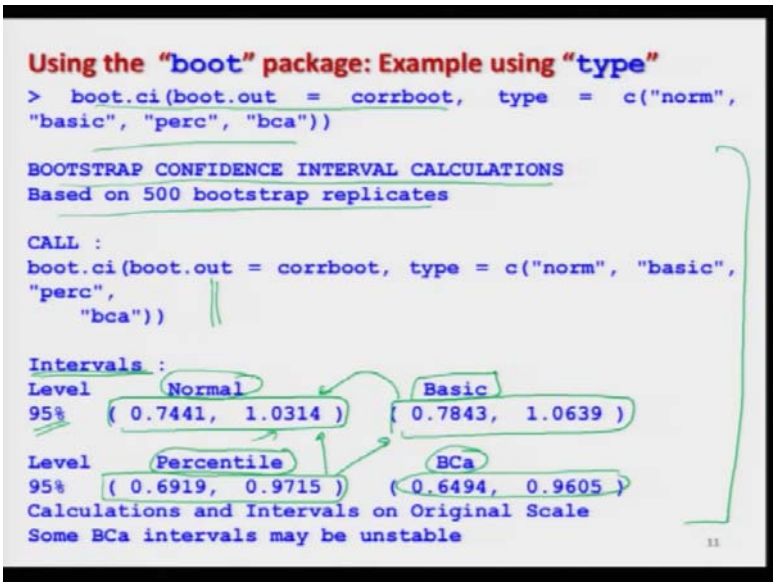
So, it depends on you, which option you want to use; but I will try to show you all the four. But before I go further, let me try to give you an idea that why this t-method is not computed under this thing.

So, actually the students, t-method confidence interval is unique and it needs an estimate of the bootstrap variance which we are not providing in the boot package and that is why when you try to do it, then this command will give you an outcome which will read like bootstrap variances needed for studentized interval.

And actually, in order to obtain this thing, we need jackknife technique also or a second-level bootstrap, which we have not considered here. So, since these estimates of variance can be obtained with the second-level bootstrap or jackknife techniques which we have not considered here, so this so only these four types of bootstrap confidence interval can be obtained from this boot package.

In order to obtain the confidence interval based on the t-method, we will use another package which gives us directly bootstrap. Although, you can write your own program also; but at this moment, I am not considering that option because that option will always be there with you in all sorts of situation.

(Refer Slide Time: 12:05)



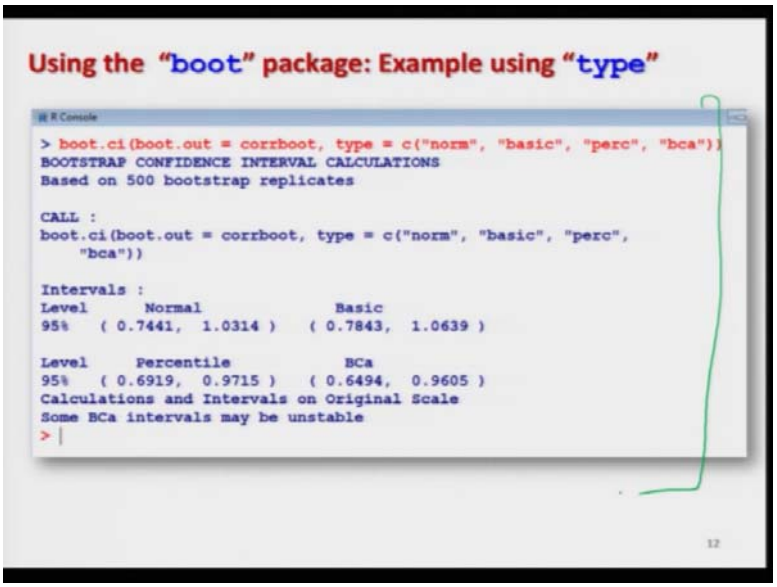
So, now, I try to execute this command on the R console and here is the outcome and later on I will try to show it on the R console. So, I try to give you give here the same command on the R console and here is the outcome which you can see. So, you can see here the heading is BOOTSTRAP CONFIDENCE INTERVAL CALCULATION which is Based on 500 bootstrap replicate that is correct. And here, it is trying to give the outcome and here, it is the details of the intervals.

So, you can see here the Level which has been considered is 95 percent that is also in your control that you can decide it, right. And, you can go to the help menu and can find out more details. The confidence interval which is based on the Normal distribution is given here it is 0.7441 to 1.0314, right. Similarly, the confidence interval computed using the Basic confidence interval methodology, it is obtained here, you can see these are and you can compare it with the normal.

So, you can say they are pretty close, but they are different, right. Similarly, on the same data set, if you try to compute the confidence interval using the percentile method, you can obtain the confidence interval here like this; 0.6919 to 0.9715. So, you can see here this confidence interval is again different from the normal and basic confidence interval. And the fourth one is bca type confidence interval.

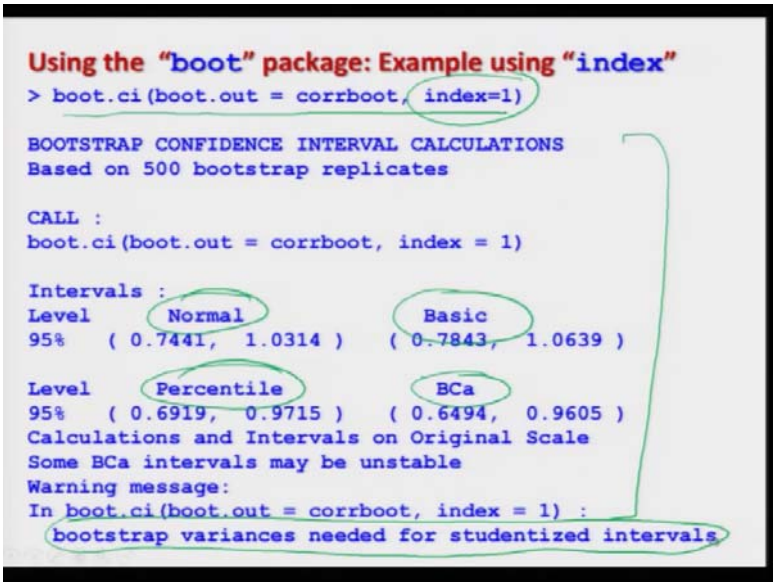
So, which is obtained here like this, which is again different than all other confidence interval, right. So, this is how you can obtain the four types of confidence intervals using the command boot dot ci.

(Refer Slide Time: 14:01)



And this is here, the screenshot of the same outcome which I have shown you here.

(Refer Slide Time: 14:08)

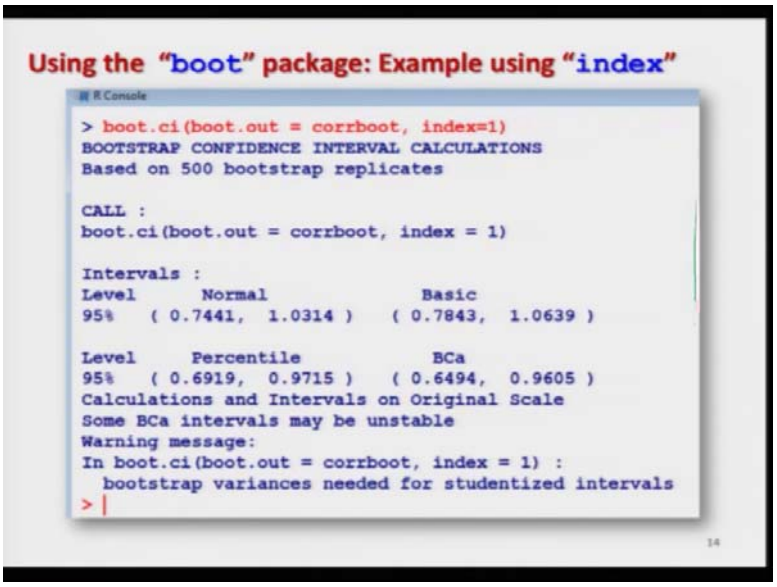


And now, in case if you want to use the index option, then the command can be modified here boot dot ci and we write here index equal to 1 and you get here exactly the same

outcome; you can see here, right. And if you try to use your index equal to 1, you can see here it is giving you a message here that bootstrap variance is needed for studentized interval.

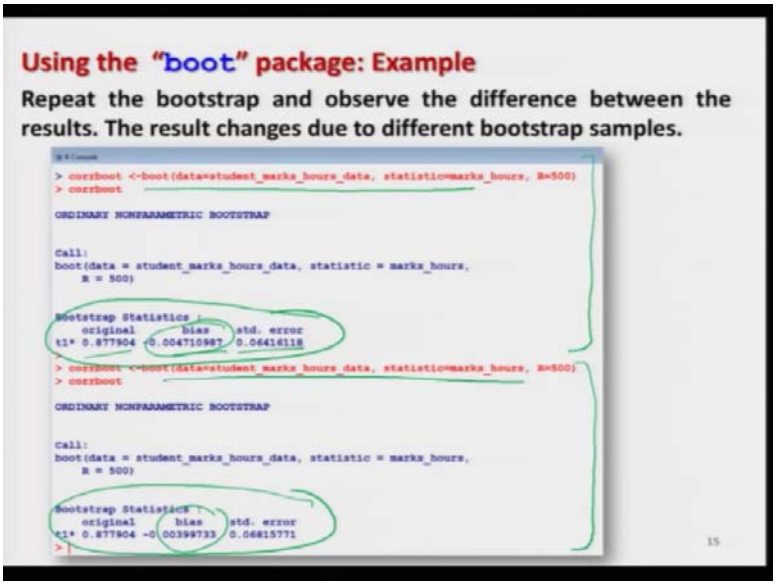
So, that is why it is not computing the bootstrap t-intervals, but it is giving you here the normal, basic, percentile method and bca confidence interval, right. So, you can use any one of them, right.

(Refer Slide Time: 14:48)



Now, well, this is the screenshot of the same outcome which you will obtain when you try to do it on the R console.

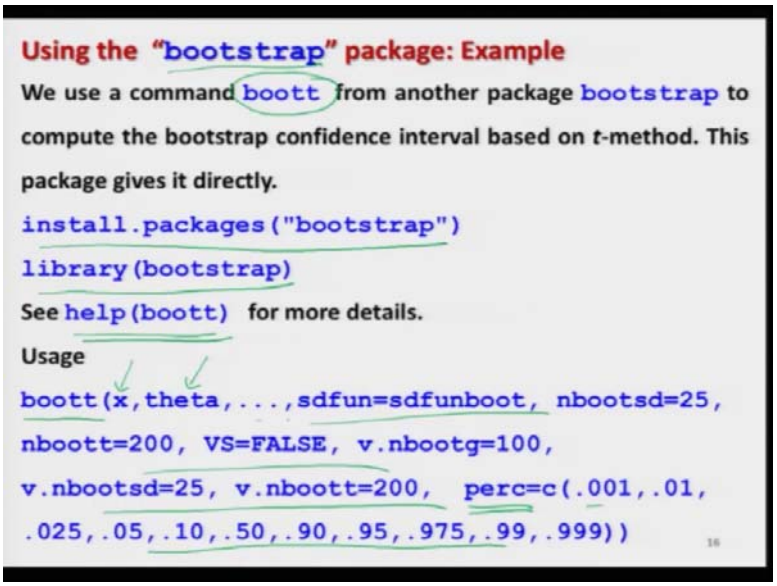
(Refer Slide Time: 14:53)



And before I go further, I just want to show you that if you try to repeat the bootstrapping process; then in every process, you will get different outcome. For example, here I have put I am trying to show you here two outcomes of the same command. From the same data set which I have used here on the students marks and hours, I try to repeat this command two times; one here and two times here and these are here the outcome.

So, you can see here, here is the bootstrap statistics which is based on the correlation coefficient, bias and standard error and here, if you try to see this is again different. So, here it is 0.0047. It is here -0.003. So, what I am trying to say that every time, if you try to repeat the bootstrap, you will get a different outcome and if you try to obtain the confidence interval, then they will also be different, right ok.

(Refer Slide Time: 16:03)

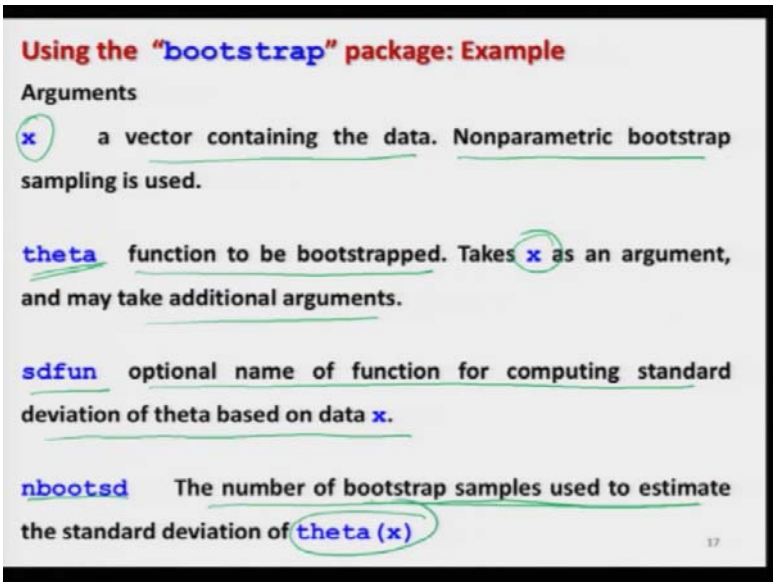


Now, I come to the other aspect that how to obtain the bootstrap confidence interval using the t-methodology. So, I am going to use here the package bootstrap and for that, the command is here different which is b o o t t, right. So, be careful about the spelling which are very very similar, right.

There is only a difference of t, right. So, first I need to install this package using the command install dot packages, then I load it using the library and inside parenthesis bootstrap and I will say that if you want to see the more details about this one, you would simply go to the help of this function.

And the structure of this function is something like you have to give here b o o t t, boott; then, you have to give here x theta etc. etc. So, I try to give here you a brief introduction or brief description of these values which we want and you can see here there are many many options which are given over here and here also, you can define your percentiles also what you want 1 percent, 5 percent, 10 percent and so on or if you want to choose any particular one also, that is also possible, right.

(Refer Slide Time: 17:34)

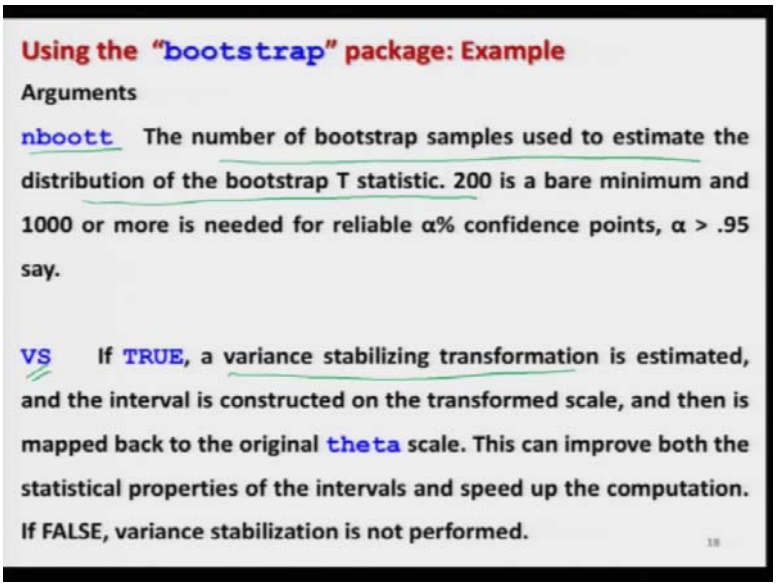


So, let me try to just give you the quick information on these options. So, x here is indicating the vector containing the data. So, that is what we have to give on that what is the data on which we want to compute the confidence interval and here in this case, the non-parametric bootstrap sampling is used. Then, we have another option which we have to give and this is indicated as theta; t h e t a.

Actually, this is a function which has to be bootstrapped and it and in case if you take this x as an argument, then it may require some additional arguments also. This, I will try to show you with an example, right. Similarly, there is another there are several function like here is here sdfun, which is the optional name of the function for computing the standard deviation of the theta, based on the data in x. Then, there is here nbootsd.

And then, we have nbootsd. So, this will indicate the number of bootstrap sample used to estimate the standard deviation of theta based on the data vector x.

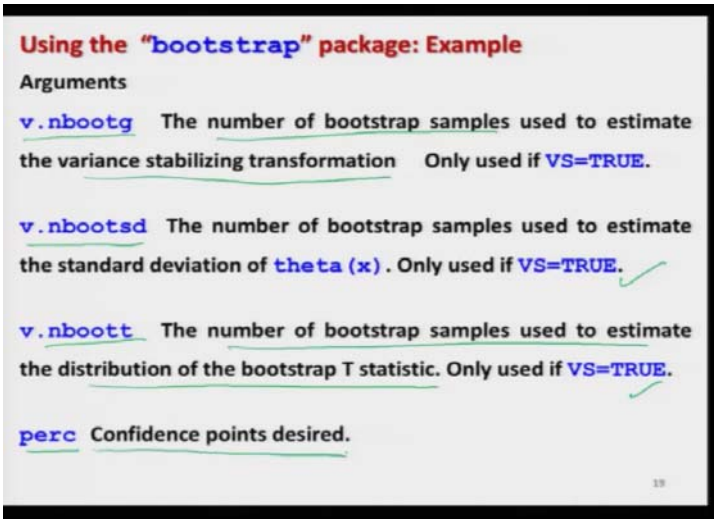
(Refer Slide Time: 18:45)



Simply, we have here nboott which is the number of bootstrap sample which are used to estimate the distribution of the bootstrap T statistics and in such a case the minimum number what is needed is say 200, but I suggest that it should be more than 200.

And, it if you try to increase this number, you will get more reliable confidence points. Similarly, there is here another option here VS which is a logical value which takes two option either TRUE or FALSE and it indicates the requirement of variance stabilizing transformation and so on.

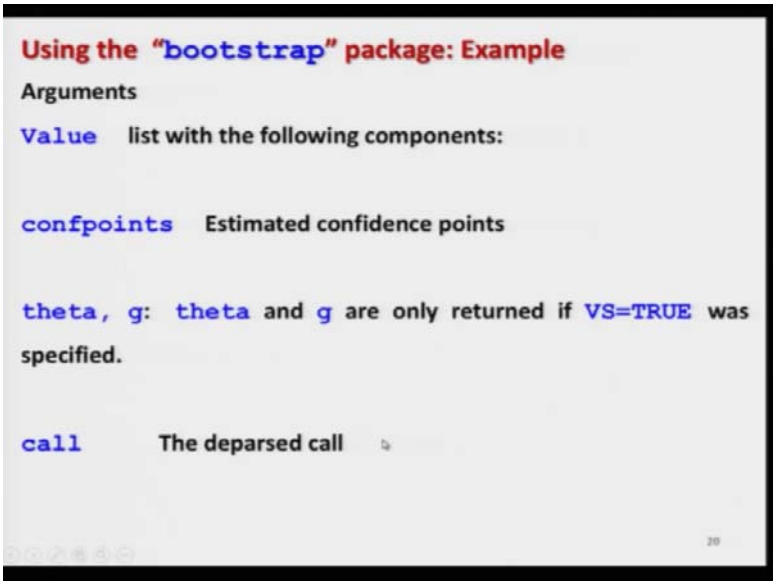
(Refer Slide Time: 19:26)



I am not going into that detail because I have not covered these topics and similarly, we have here another option here v dot n boot g which indicates the number of bootstrap samples to estimate the variance stabilization transformation. And then, we have simply have here another option v dot nbootsd, which are again related to the number of bootstrap samples used to estimate the standard deviation of theta x.

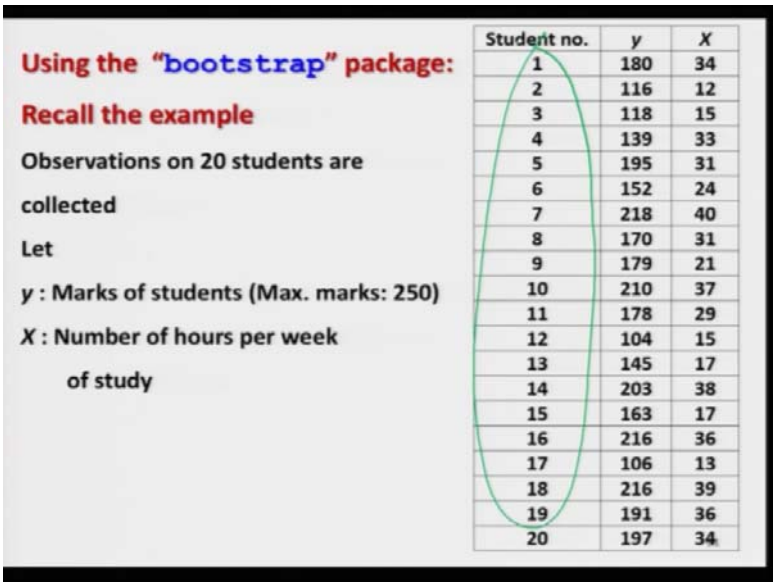
And, I mean there are here v dot n boott and which is the number of bootstrap sample used to estimate the distribution of the bootstrap T statistics and this and actually, these two options are going to be used only when you are trying to use the variance stabilizing transformation, right. And then, there is here perc which is the percentile, so that will give you and give you the option to specify the confidence point which are needed, right.

(Refer Slide Time: 20:22)



And then, inside the arguments, you have to give the value which will give you the list of different following confidence like for different components, like confidence point which will give you the estimated confidence point; then, we have here theta, g. So, theta and g value are returned if you say the variance stabilizing transformation to VS and so on.

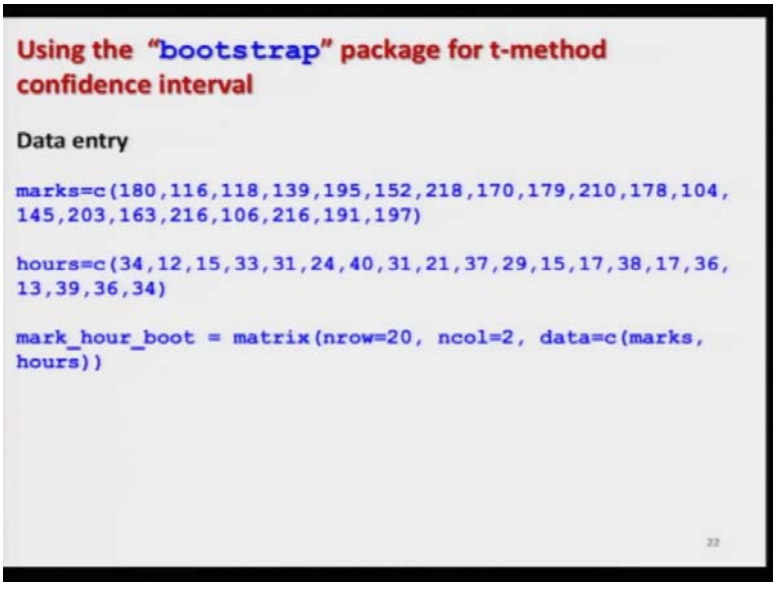
(Refer Slide Time: 20:54)



Student no.	y	X
1	180	34
2	116	12
3	118	15
4	139	33
5	195	31
6	152	24
7	218	40
8	170	31
9	179	21
10	210	37
11	178	29
12	104	15
13	145	17
14	203	38
15	163	17
16	216	36
17	106	13
18	216	39
19	191	36
20	197	34

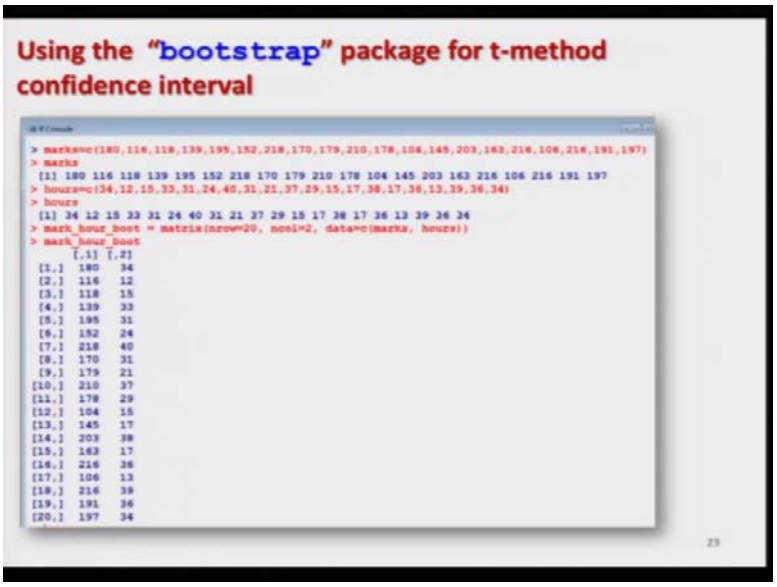
And so, you can see here, there are many options. But now, let me come back to my example and try to show you how easily you can compute this confidence interval using this package. So, this is here the same data set on the 20 students on which we have obtained the confidence interval earlier.

(Refer Slide Time: 21:11)



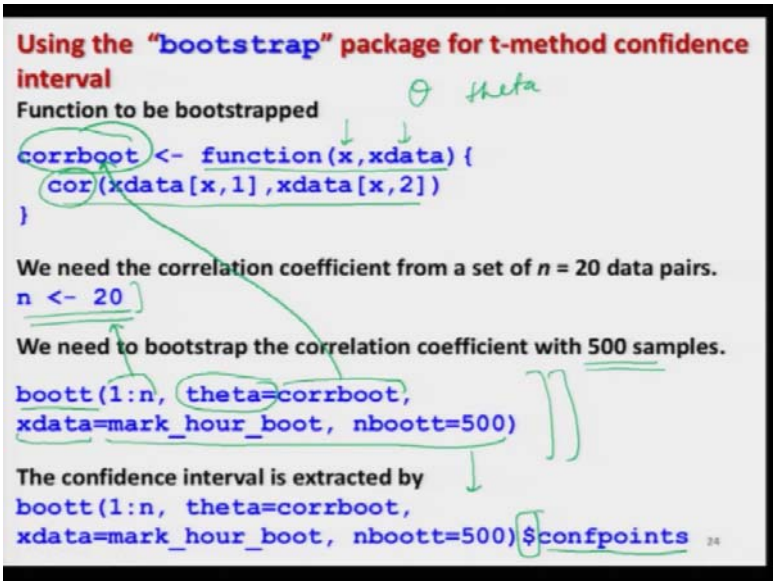
And this is the data entry part. Once again, just for your quick revision.

(Refer Slide Time: 21:16)



And this is the data frame that we had obtained earlier

(Refer Slide Time: 21:19)



And this is the and now here, we start with the construction of confidence interval. So, you remember, we had an option theta; t h e t a which was denoted like this. So, theta is the function which has to be supplied for the computation of bootstrap statistics.

So, this is something like what we had used earlier also. What I try to do here that I try to create here a function which is creating which is computing the correlation between the data on marks and number of hours. So, you can see here, I have simply use here correlation function c o r and then, I am trying to give here in the first argument as x and second argument as x data and here, we need 20 observation because our the data size is 20. So, I am writing here n is small n is equal to 20.

So, now, and I want to compute the correlation coefficient based on 500 samples. So, since we are using here a bootstrap package. So, this package will also compute different types of thing related to the bootstrap. So, we will let us try to do how it can be done. So, now, I try to write down here the command.

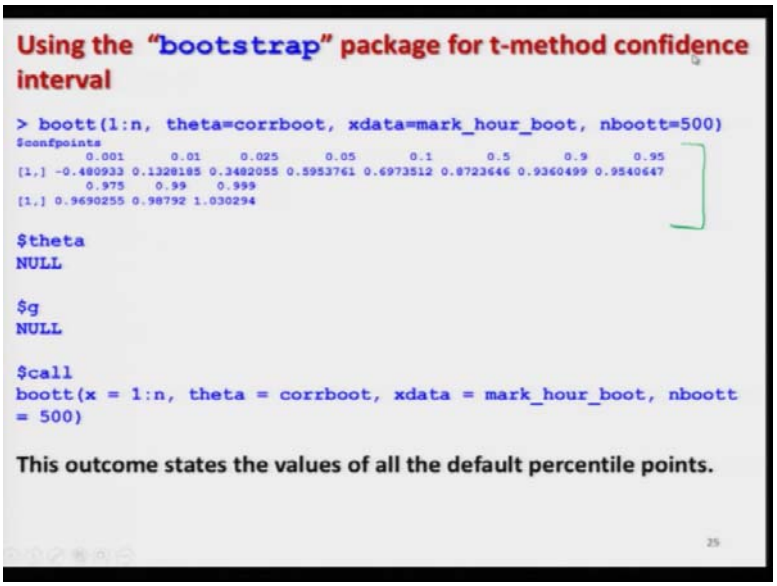
So, command here is b o o t t which is the function, then I am writing here 1 to n, this is trying to give the data that what data set I have to use to sample the bootstrap samples. So, this is here and there are 20 observations. So, I am asking in the first place that please read the data from 1 to n; that means 1 to 20. Then, I am giving you giving here the function theta.

So, theta is the function which has to be computed; that means this is the statistics for which we want to compute the bias, standard error, confidence interval. So, this function here, I am denoting here by corrboot which is obtained here. You can see here, this is corrboot function.

So, this theta is the function which we want to estimate and I simply have to write the name of the function over here and then, this computation has to be based on which the data set? This information is being given here by x data and then, I have to write down here the name of the data frame.

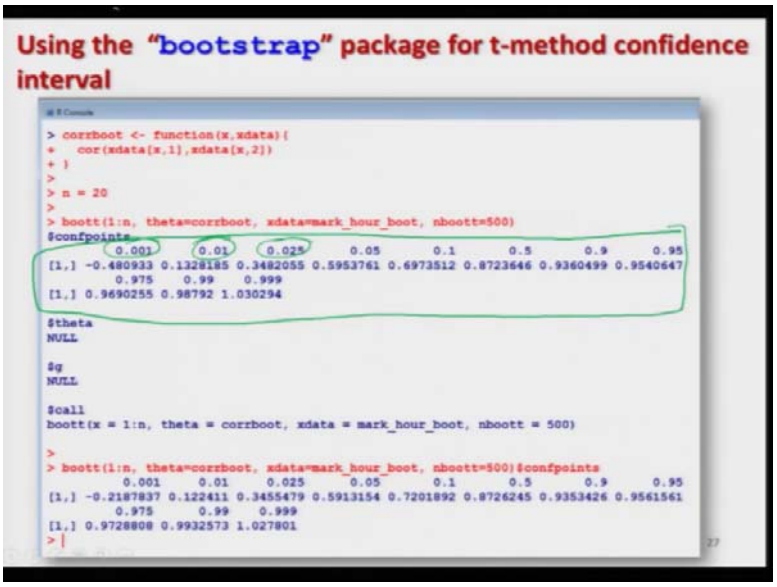
So, it this will give you several outcome. But in case if you just want to extract the confidence interval, then what you have to do? You have to simply write the entire program and then, you have to write that dollar sign and then, write confpoints. So, you can see here, I had explained you that how you can extract different types of quantities from the outcome. So, this is how this will give you the confidence interval.

(Refer Slide Time: 24:32)



So, let us try to do it and see what type of things we will get here. So, I used the same command on the R console and I get here this type of outcome. Well, I have made the font size to be smaller so that I can exactly show you how it looks like.

(Refer Slide Time: 24:52)

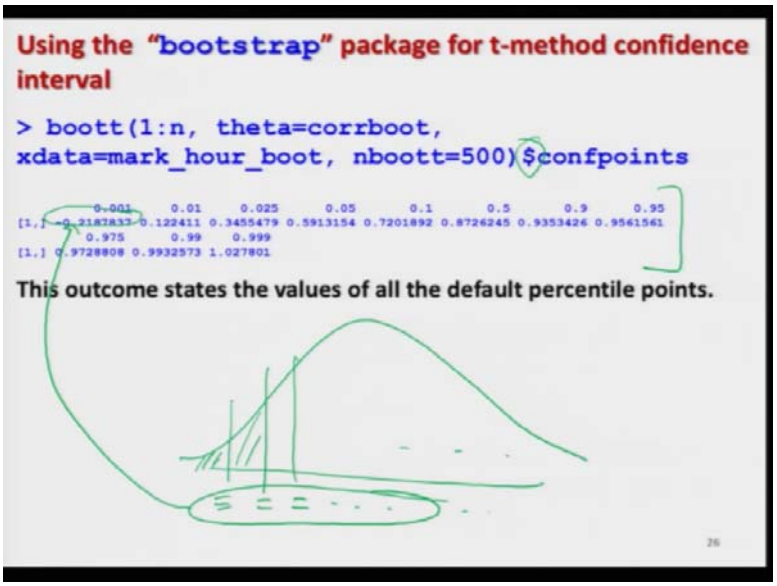


But if you want to be more clear I can show you here, this is the outcome which will look on the screenshot, right. So, you can see here this is the outcome which I have copied and pasted over there, right. So, these are the something like values of α equal to

0.001, 0.01, 0.025, these are your percentile points actually, which have been computed on the basis of bootstrap sample.

So, you can see here that all sorts of confidence points are obtained over here and I will show you that if you want to extract only particular values, then how you can do it, right.

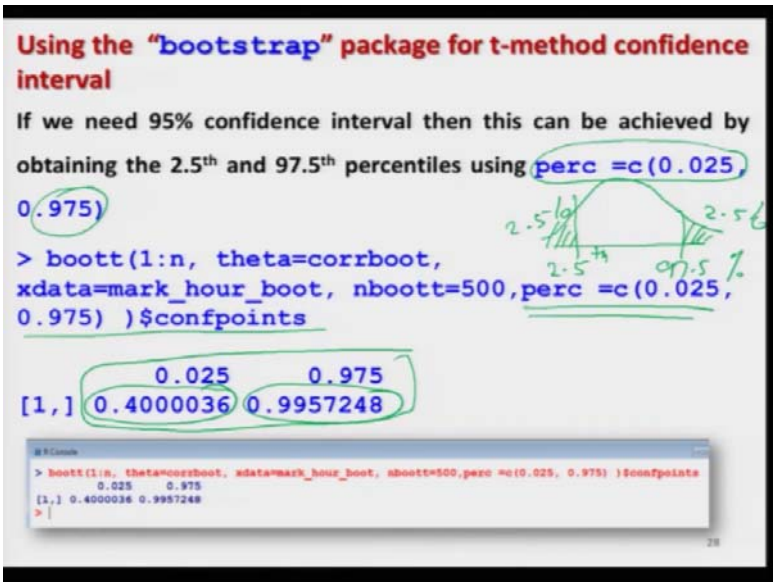
(Refer Slide Time: 25:27)



So, well, in case if you try to extract only the confidence points, then I use the command here the same command and we which is followed by dollar and confpoints, confidence points and I can extract this outcome from this big outcome. It is trying to give us many information, but we do not want all this.

We are interested only in the confidence intervals, so it is trying to give us the confidence point. So, these are the values of the confidence point. For example, if I try to take here this type of distribution. So, if I try to take here say this alpha is equal to 1 percent, 5 percent, 10 percent and so on. So, it is trying to give us all these values which you will obtain here on the x axis, here they are given here in the second row, right.

(Refer Slide Time: 26:30)



For example, I can show you that well if you want to find out only the 95 percent confidence interval. So, 95 percent confidence interval means, you want here this to be 25 percent. So, this is going to be your here 2.5th percentile and here, this is going to be 2.5 percent. So, this is going to be 97.5; 97.5th percentile. So, I can use here the command percentile is equal to c equal to 0.025 to 0.975.

So, and I use the same command with the change perc. So, you will get here this value. So, this is the lower limit of the confidence interval and this is the upper limit of the confidence interval and you can see here, these are the same value which you have obtained here. You can see here 0.025 and 0.975, right.

So, this option is giving you all possible confidence interval and this is here the screenshot, here you can see this is my function here corrboot and then, I try to execute it and I get here this outcome and from there, I try to extract the confidence interval. So, right. So, now, I already have explained you that how you can extract a particular confidence interval. So, this comes to an end of this lecture.

So, in this lecture, I have shown you how you can compute all types of confidence interval and you can compare that the values of the lower and upper limits are differing from each other. Well, in this case, the distribution was not bad; so that is why the difference in the upper and lower limits in different types of procedures, they are more or

less same. There is not much difference. But if the distribution is disturbing like as having extreme values, spikes etcetera then these values will change a lot.

So, you can see here that the time taken in explaining how to use it or how to compute the confidence interval is nearly half of the time which I took in explaining the theory and philosophy behind the confidence intervals. So, you can see that it is not difficult at all to find out the confidence interval and definitely, when you are trying to go for data sciences without confidence interval, I doubt if you can really survive.

You always want to know the unknown value. The unknown value can be estimated in terms of as a point estimate as well as an interval estimate and both are helpful. What type of information are needed, why? Because, you are working in blank with the population which is not known to us.

So, it becomes rather than a rule, then that you have to find out point estimate, you have to find out standard error as well as you have to find out the confidence interval. Without this, you cannot proceed further. So, now, I will do one thing, just to make it make the things more clear on the software, in the next lecture, I will try to take a very simple example and I will try to find point estimate, standard errors as well as confidence interval in the same data set in one lecture.

So, till then, you practice and I will see you in the next lecture. Till then, good bye.