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 – 18
Simple Random Sampling
SRS with R using "sampling" and "sample" Packages

Hello, welcome to the course Essentials of Data Science with R Software 2, where we are trying to understand the basic concept of Sampling Theory and Linear Regression Analysis. In this module we are going to continue with the topics of Simple Random Sampling with R Software. You may recall that in the earlier lecture we consider the package R sample to draw the samples using simple random sampling under without replacement and with replacement cases.

In R usually a job can also be done by some other package why? Because different people who are working independently they can contribute different packages; those packages may not be identical, but possibly they might be doing a similar job or even the same job, but their outcomes, the pattern of outcome, the syntax, commands that may differ.

So, thinking on the same lines, today I am going to use another package whose name is sampling to draw simple random sample. And you will see that there is a difference between the two and whatever is the difference I will show you that I am going to utilize it.

Whatever is the outcome that will be coming from here, from sample and sampling that I will try to compare and I will try to show you, that how the output differs and how they can be used in the best possible way. So, let us try to start our discussion with the package sampling.

(Refer Slide Time: 02:13)

So, in order to use the package sampling first you need to install it. So, in order to install it you have to simply use the command, install dot packages and inside the double quotes within the parenthesis just write sampling and then you have to load it library sampling. And this software it provides the simple random samples using with and without replacement, right ok.

(Refer Slide Time: 02:43)

So, let me first try to show you what are the usage. So, the syntax or the command to draw a simple random sample in case of without replacement is srswor, all in letters and

inside the parenthesis you have to write n and N. The n and N are the same what we have assumed a n is the sample size and N is the population size.

So, essentially this srswor command is asking that write the command srswor and inside the parenthesis first write the sample size which you want and then write the population size, ok. So, this command is going to give us a simple random sample which has equal probabilities of selection for every unit and this returns a vector.

This vector will have two quantities 0 and 1 right and there is an indication the element k of this vector indicates the status of unit k. That means, in case if k = 1 this means the unit is selected and if k = 0 then the unit is not selected. Meaning of this that is k = 0 and k = 1, I will try to illustrate with some examples.

(Refer Slide Time: 04:23)

Drawing of sample using R : SRSWR

Usage   srswr(n,N)   → Popn.
                ↓
Description   Sample

Draws a simple random sampling with replacement of size n (equal probabilities, fixed sample size, with replacement).

Arguments
n   sample size.
N   population size

Value
Returns a vector of size N, population size.
Each element k of this vector indicates the number of replicates for unit k in the sample

But before going to the example let us also see how one can draw the simple random sampling with replacement. So, in order to draw a sample by srswr the command is srswr and inside the parenthesis you have to give n and N. So, n is the sample size and N is the population size.

And once we try to execute it, it will return a vector of size N each element of this vector will have some number of values and that will indicate that each element k of this vector is the number of replicate for unit k in the sample, what does this mean? I will try to show you with some examples now.

(Refer Slide Time: 05:15)

First let me try to show you the examples on the slides itself you please try to understand and later on I will try to execute these things on the R console. I am not doing it right now because in case if I try to use this command in the R console, the outcome is going to be different what we are looking on the screen and in further slides we are going to use this outcome.

So, that is why I will try to explain you keeping this outcome fixed and later on I can explain you on the R console, that if the outcome changes means everything changes accordingly, right ok. So, now if you try to see here, I am trying to take here a population of size 10. So, N=10 and n = 3; that means, if I would like to have a sample of size 3.

So, I try to use the command here srswor and with sample size 3 and population size 10. And whatever is the outcome of this vector I will try to store in a vector swor. So, the name indicates that this is a sample by without replacement. So, in case if you try to execute, you get here a outcome like this one.

Now, this is what I was saying that these are the values which contains only 0 and 1s and 0 and 1 they have an interpretation that this is trying to give me the position of values in the population say Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8, Y_9 and Y_{10}. So, those places where you have here 1 corresponding to these 1s the corresponding units are selected in the sample.

4

So, here you can see corresponding to this 1 the unit is Y 5, corresponding to this 1 the unit is Y 8 and corresponding to this one the unit is Y 10. So, this Y 5, Y 8 and Y 10 they are selected in my sample and all those places where we have got 0. These 0s are indicating that the that these units are not selected in the sample, but that is obvious we wanted to have a sample of size 3 and now here we have a sample of size 3, 1 unit, 2 unit and 3 units right.

So, that is what I am writing here that the interpretation is that number corresponding to 1 are selected in the sample. So, this is what I was trying to communicate with you convey with you convey to you, that different packages may have different interpretations and different styles, but the outcome is similar.

So, now, if you remember in the earlier lecture what I had done? I had taken an example and then I was trying to choose the values of students or their heights based on those variable. So, I would try to do the same thing with the package sampling, right.

(Refer Slide Time: 08:42)

So, if you remember, but before that let me try to show you the outcome of srswr also. So, the command here is to draw a sample of size 3 from a population of size 10 by srswr is srswr and inside the parenthesis 3 comma 10. So, now, this will give you an outcome like this one here, right.

So, you can see here this is here the screen shot. So, here also the outcome is there; the interpretation is the same that the number corresponding to 1 are selected in the sample right ok. So, that means, Y_1, Y_3 and Y_{10} they are selected in the sample and I am trying to store these values as swr vector, ok.

(Refer Slide Time: 09:39)

And so, now, let me try to show you that this data what we saw in the example of srswor that has been chosen according to variable 1 to 10, but now we would like to execute it on the example that we consider earlier and then we would try to choose the sample according to the variable name.

(Refer Slide Time: 10:05)

6

Drawing of sample using R
Y: Height of students in a class
N = 10 : Number of students in the class (Population size)
n = 3 : Number of students in the sample (Sample size)

Name of Student	Y_i = Height of students (in Centimeters)	
A	Y_1 = 151	
B	Y_2 = 152	
C	Y_3 = 153	
D	Y_4 = 154	
E	Y_5 = 155	5ᵗ
F	Y_6 = 156	
G	Y_7 = 157	
H	Y_8 = 158	8ᵗ
I	Y_9 = 159	
J	Y_{10} = 160	10ᵗ

So, if you remember we had consider this example, where we had a data set in the form of a table; where there are 10 students whose names are A, B, C, D up to J and their heights are denoted by Y_1, Y_2, Y_{10} and their values are given here in the table.

(Refer Slide Time: 10:26)

Drawing of sample using R

Suppose we want to select the name of student or Height of the student.

The data in R will usually be given in a data frame, CSV file or any other format.

Suppose the data is stored in a data frame **heightdata** by using the following commands:

```
height=c(151,152,153,154,155,156,157,158,159,160)
name=c("A","B","C","D","E","F","G","H","I","J")
heightdata=data.frame(name,height)
```

And using this data set we had created a data frame. So, I have entered the variables here height and the variable name for the name of the students, exactly in the same way what is given in the data. And then using the command data dot frame I have constructed the data frame of name and height.

(Refer Slide Time: 10:50)

So, that is exactly the same thing that we did in the last lecture and this is my here the data frame and from this data frame if you want to extract the data on individual variables. For example, if you want to extract the data on the name, which is here then I had given it the name names n a m e s and if you want to extract the data on the height that name I had given earlier as a heights.

And the command to extract this data is that first you try to write down the name of the data frame followed by the variable name which is joined by a dollar sign and the same thing is here also. If you want to extract the data on height, write down that data frame name height data then variable name height and then join them by dollar sign. So, you can see here that this is giving you the same data here whatever is given in the data frame, ok.

(Refer Slide Time: 11:47)

So, that is the same thing what we had done earlier and this is the screen shot. So, you can believe that whatever I am showing you here that that will hold correct.

(Refer Slide Time: 11:53)

Now I try to compare, now if you try to recall earlier when we use the command sample, then in order to get the sample on the basis of the variable names. We had use the command sample the variable name names and size of the sample which was 5 and then because we wanted to use srswor. So, we use replace equal to FALSE and this had given us a simple random sample without a replacement.

9

Now, we would like to use this command srswor and if you try to see here I am trying to just use this data set what we have obtained earlier. If you try to see here that I means in this example also I had consider N equal to 10 and n = 3. So, by this swor values are something like this. And here if you try to see this is my 5t value this is my 8th value and this is my here 10th value.

Now, my objective is this, that I have got here a sample in terms of 0's and 1's using the command srswor and based on that I would like to see that what are these 1's Which are corresponding to a given variable name. For example, I have consider in that example 2 variables names and heights. So, now, I am choosing here 5th unit, 8t unit and 10th unit.

So, I would like to find out here that in the names vector, what are my 5th, 8th and 10th units or what are the 5th, 8th and 10th names. Similarly when I try to use the variable here heights, then I would like to know, what are the values corresponding to 5th, 8th and 10th value in the variable heights.

So, that is what I want to do. So, if you remember when we considered the introduction to R software means I had quickly given you some commands related to logical operators. At that time we had considered an example like something x x double equal to 1 or 0 something like this one. So, I had explain you that when I try to write x double equality sign equal to 1, then this double equality to sign is the logical equality sign.

And we are trying to see here that in the data vector x what are the values which are equal to 1. So, this will give us an outcome in terms of TRUE and FALSE and then once I use here this command here x with this bracket sign then this entire command is going to give us the values in the data vector x for which x = 1.

So, if I write it here neatly, when I am trying to write down here like this. So, this command is going to tell us that what are the values in the vector x corresponding to which x = 1 or x is equal 2, right. So, this logic can be used over here and now I have here a data vector swor, where I have 0's and 1's.

So, when I say swor double equality sign that is the logical equality = 1. That means, I am trying to find out all those values in the data vector swor corresponding to which the values are 1. So, what are the values in the data vector swor? Which are 1. And now I am

asking that please tell me the corresponding names. So, in case if I try to combine both the sentence I can say very simply, that mister R, please tell me the values in the name vector corresponding to the data vector swor, where the swor values are 1.

So, now what it will do here? That, it will give us; it will give us the values of names corresponding to the places where swor was 1, swor was 1 where? That you can see here Y_5, Y_8 and Y_{10} and you can see here in this table that Y_5 name is E, Y_8 name is H and Y_{10} name is J. So, that means, I am looking for these values E, H and J. And now if you try to see in this outcome here what you are getting here? This is here E, H and J.

One thing you have to just keep in mind, that the command which I have used here names swor double equality 1, this will give you the same outcome E, H and J just like it here this one, but still I am using here as dot vector; because I just want to save the outcome of this command in the form of a data vector. So, that it is compatible with mathematical manipulations at a later stage otherwise only this command is sufficient to give this solution and answer.
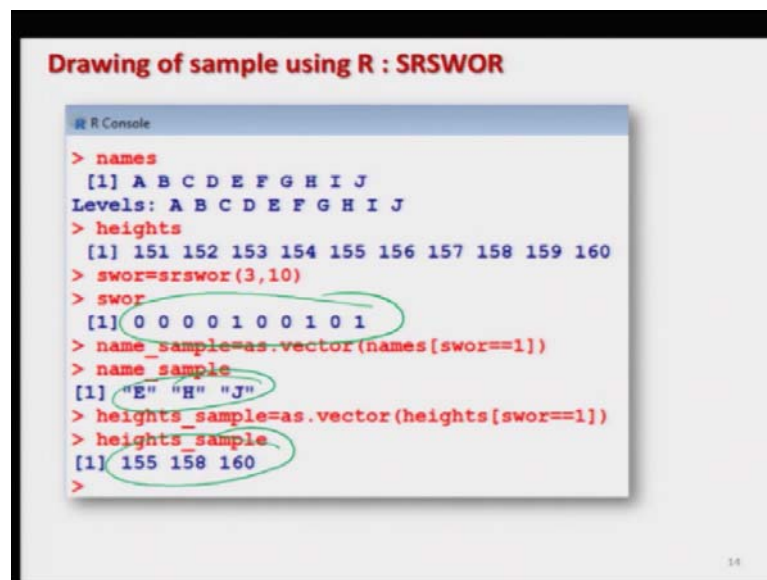
(Refer Slide Time: 18:04)

So, now in this example I have taken the variable names now I try to take another variable heights. And then corresponding to the same sample data that we have selected and that was stored under swor where we had 3 units Y_3, Y_5 and Y_8, they were selected.

11

So, now, I would try to find out the values of Y_3, Y_5, Y_8 under heights earlier we had found it under names. So, now, what I use the command? I try to use here the command here swor = double equality equal to 1. So, that means, please try to identify the values in the data vector swor which are equal to 1 and corresponding to them, please come to the variable heights and try to identify the corresponding sampling units.

So, in this case there are 3 values at 3rd, 5th and 8th position which are equal to 1 here. So, now, the control will come to heights and it will find the heights of the 3rd, 5th and 8th observation. So, this command will give us the same output, but as I said earlier we would like to store this outcome in the form of a vector.

So, I am just using here as dot vector to store the observation in a data vector and you can see here this is going to give us the outcome 155, 158 and 160. And if you try to look into your example the 5th unit here is 155, 8th unit here is 158 and 10th unit here is 160. So, now, we had chosen the 5th unit, 8th unit and 10th unit and corresponding to which we have found the names or their heights. So, this is how we can proceed further.

(Refer Slide Time: 20:07)

And you can see here this is the screenshot, this is the screenshot when I conducted this experiment on my computer, but now once I try to do it on the R console before you then surely this swor will change. And then the corresponding values of here E, H, J will also change and this will also be change the concept will remain the same.

12

(Refer Slide Time: 20:33)

Now, I try to take the example of a simple random sampling with replacement using the same example. So, for example, I can show you here, I am taking here as population of size 10 and a sample of size 8. So, I want to draw a sample of size 8 from a population of size 10, well I have taken it to be pretty large; because I wanted to show you something.

So, now, I use the command here srswr 8 and 10, remember earlier when I used the command here srswr here with 10 comma 3 you can see here the values were only here 1 and 0's, but now I would like to show you something more.

So, when I try to take here srswr 8, 10; that means, I want to choose 8 values out of 10 values by simple random sampling with replacement; then my outcome comes out to be like this you can see here that this time it is not actually only 0 and 1, but there are some other values 2 and 3 also ok.

So, what does this mean? The interpretation goes like this that let me try to write down the corresponding units for these numbers. So, now you can see here when I am getting here 3, this is indicating that there are three values of Y_1 are selected. So, the sample will look like Y_1, Y_1, Y_1, now it is saying that there is one value corresponding to Y_4. So, that means, another value selected is Y_4 and similarly here it is 2 corresponding to Y_7.

So, that mean there are two values corresponding to Y_7 and there are two values here corresponding to 8. So, Y_8 and Y_8 are occurring two times. So, you can see here now my

sample consist of 3 values of Y 1, 1 value of Y 4, 2 values of Y 7 and 2 values of Y 8 amounting to N = 8; that means, sample of size 8.

So, this is the interpretation when I say we are using the command srswr. So, the rule is the numbers which are not equal to 0 are selected in the sample and these numbers are representing the frequency of the numbers in the sample. Here you can see this is different than the case of srswor; because in the case of srswor these values can be can only be either 0 or 1; because either the value is selected or not selected a value cannot be selected more than once.

So, in this case suppose I have got the sample. Now my next objective is this I want to know that what are the units which are selected; remember I am asking only the unit which are selected I am not asking for the frequency. So, in case if I want to know the selected unit then I can use here the command 1 to 10; which was my this population size because here the population is given as integer.

And then I am trying to find out the values of swr which are not equal to 0, if you remember negation sign is given by a symmetry sign and equality sign; that means, not equal to 0 right. So, now from this swr vector it is trying to say that this value 3, 1, 2 and 2 they are not equal to 0 and that is what it is giving me here that the corresponding units are 1, 4th, 7th and 8th they are selected.

Now in case if I want to find out what is the frequency of these selected values, then I can use here the command ah, on the same lines that find out the values of swr which are not equal to 0 from the data vector swr using this command right. Now, you are familiar with this command and it will give me a value here 3 1 2 2. So, you can see here this 3 1 2 2 is the same thing which I have obtained here.

That means, the value here Y$_1$ this is repeated three times, the 4th unit this is occurring one time, 7th unit this is occurring two times and 8th unit which is occurring two times. So, you can see here I have tried to illustrate it in a more simple way, but just by looking at the outcome you can see what is really happening, ok.

(Refer Slide Time: 25:15)

Drawing of sample using R : SRSWR

Suppose we want this sample in terms of names of persons.

```
> name_sample_wr=as.vector(names[swr!=0])
> name_sample_wr
[1] "A" "D" "G" "H"
```

Suppose we want this sample in terms of heights of persons.
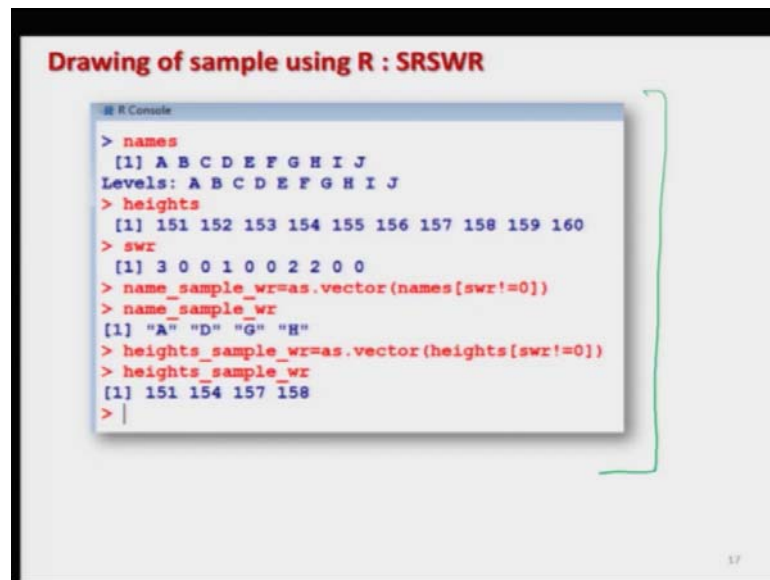
Command `heights[swr!=0]` will do the job but we want to store it as a vector.

```
> heights_sample_wr=as.vector(heights[swr!=0])
> heights_sample_wr
[1] 151 154 157 158
```

Now, the same thing I try to do that I would like to consider the same example and I would try to find out the values corresponding to the variable name. So, is so this is simply a repetition of whatever I have done. The only thing is this I simply have to need I simply have to change here the data vector.

So, I am trying to find out here the values of the name corresponding to which the data in the swr vector is not equal to 0 from here and these values are A, D, G and H right. And similarly if you want to find out the corresponding value of height, so, I can take here the values of height, corresponding to which the swr the vector has nonzero values right and you can see here this will come out to be 151, 154, 157 and 158.
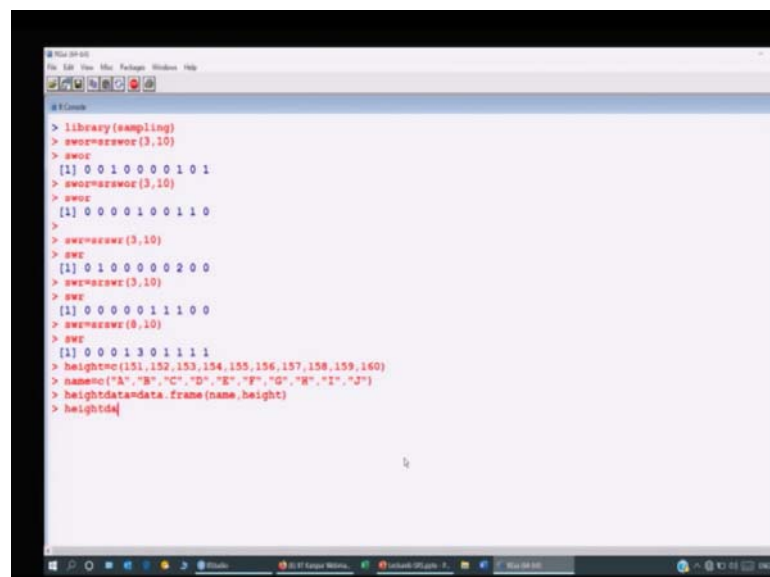
(Refer Slide Time: 26:19)

So, this is how you can do different type of manipulations and you can see here this is the screenshot which was coming when I was trying to conduct this example on my computer. Now I will try to show you this entire analysis on the arc console and based on that you can see that these things are working ok. So, let us try to come to R console right.
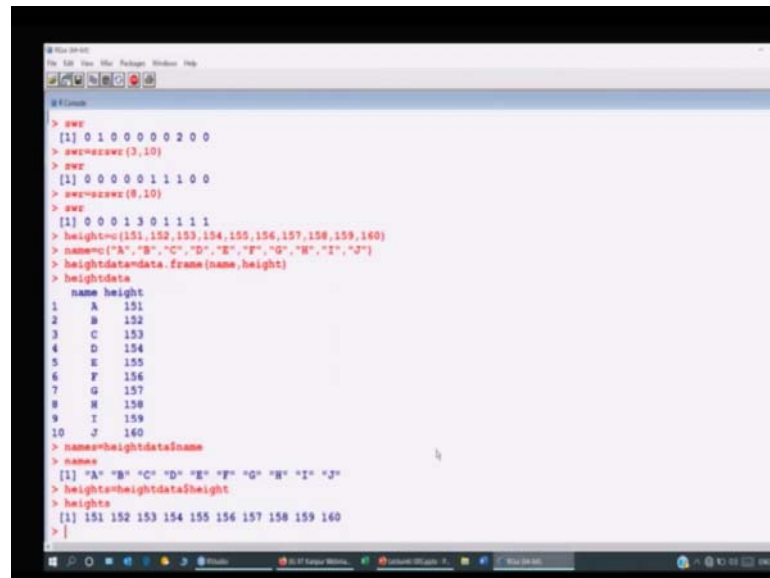
(Refer Slide Time: 26:48)

Let me first load the library, library sampling and yeah actually I already have installed this library in my computer. So, let me increase the font also, so, that you can see more clearly. Now, I try to come to the beginning. So, you see now I have installed the packages sampling and I have uploaded it; now the first example which I took in which I just drawn a simple random sampling with and without replacement of size 3 from the population of size 10.

So, you can see here if I try to do the same command over here, you can see here this is here like this, but definitely as I said now this sample will be differing from the sample which I reported in my slides; because you can see here this will as soon as I repeat it, it will miss every time it will give me a different sample.

So, and similarly after this I had obtain the simple random sample with replacement. So, if I try to take here the same command here. So, you can see here swr this is coming out to be say a sample of size 3 you can see here. Now this time it is coming out to be here 1 and here 2 right. And if you try to repeat it here you can see here now this time it is only 0 and 1 so yeah.

So, this is going to increase for example, if I try to take here a sample of size 8, then you will see here that the things are changing right. So now, we should be confident that all these things are working ok. Now I try to create this data set on my R console. So, I try to copy and paste and you can see here this is my data frame, ok.

(Refer Slide Time: 28:58)

And from this data frame if I want to abstract the information on this here names and height I have to use the command here, name sequel to the data frame and then the dollar sign followed by the name of the vector you can see here this is here names; and similarly if you want to extract the information on your heights.

So, you have to use the same command here, I am just copying and pasting before you. So, that you can also do the same exercise without any problem you can see here now this is giving me the data which is on heights. Now, after this I will try to use the same command here, let me try to show you because I said that this command is sufficient to give us the information's on the names, but still I have used here in this case you can see I have used it here as a vector.

(Refer Slide Time: 30:11)

So, first you can see here what is my let me clear my screen and you can see here the data which I had for the simple random sampling without replacement it was like this; that means, here in this case the 5th unit, 8th unit and 9th unit they are selected in my sample.

So, you can see here now I try to obtain here the value of the name corresponding to this 5th, 8th and 9th units. So, you can say see here that the 5th letter is A, B, C, D, E that is and then H, I, right. And similarly if you want to find out here heights; so, you can see here these values are 155, 58, 158 and 159. And these are very easy to verify that was the reason I had taken it was my here 5th value; that means, 155 this is 8th value; that means, 158 and this is my here 9th value which is 159, right.

So, this gives you this thing and now the only thing is this if you want to execute this command over here. So, that will only do one thing, now this will save the same outcome in the form of a data vector. So, that will be convenient for you to use some mathematical operation R functions over this thing that is the only difference, right.

So, similarly is the case of here height which I have shown you and similarly if I try to do it for the srswr unit that I have selected. So, if you want to see here what are the units which are selected here. So, I can use here the command here for example, if you remember the swr data here is given like this.

So, if you try to find out here 1 to 10 where swr is not equal to 0. So, you can see here what are the values where swr is not equal to 0 these are 1, 2, 3, 4th unit, then 5th unit, then 6 has 0, then 7th, 8, 9, 10th, this is exactly what is given here. That these are the places where the units are not equal to 0 and similarly if you want to find out here that what are the values corresponding to under do this swr which are not 0 and their frequency.

So, I can use here this thing you can see here because this is telling you that unit number 4 is occurring 1 time, unit number 5 is occurring 3 times and this 3 is the same as this 3 and unit number 7, 8, 9, 10 each of them is occurring one time, right. So, I hope I have made it clear, right and if you want to find out the names corresponding to the units which you have selected in the srswr.

So, you can see here the names corresponding to the values where we have nonzero values in the data vector swr they are D, E, G, H, I, J. And in case if you want to find out the values of the height variable corresponding to these nonzero values this is here like this.

Now, so, you can see here I have completed all the commands which I have shown you in this lecture ok. So, now, I think this is enough for you to learn and understand that how the samples can be drawn using srswr or wor in an automated way inside the R software.

So, we have two commands, we have two packages which we have utilized here. There may be more and in the forthcoming days who knows there can be some more development, but I have shown you that how you can learn a package also. Simply try to take the help from the built in software and try to spend some time in studying reading that material and I am sure that you will be successful in using any type of package.

So, I stop here and in the next turn I will try to now concentrate on some more fundamental properties; like as estimation of mean variance and confidence interval and then I will try to show you how you can do it on the R console. So, you practice and I will see you in the next lecture again, see you, good bye.