

**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 – 17**  
**Simple Random Sampling**  
**Example of SRS with R using “sample” Package**

Hello friends, welcome to the course Essentials of Data Science with R Software 2, where we are trying to understand the basic concepts of Sampling Theory and Linear Regression Analysis. And in this module, we are going to continue with the topic of Simple Random Sampling with R Software. So, you may recall that in the earlier lecture, we had considered the package `sample` and I had illustrated that how one can draw different types of samples using simple random sampling with and without replacement.

But there what I did, I had defined a data vector as a population and from there, I was trying to sample the observation of the required numbers. Now, in case if I try to extend it more and try to view this package from the point of view of data science; then you can consider that usually whenever a data is collected, it is not usually on the single variable. But that is in the form of some table, where there can be more than one variables and the data on all such variable is considered.

For example, if you try to consider the same example of online shopping; whenever a person creates the login for the online shopping on that website, the person has to give different types of information that can be name, address, telephone number, age, and some other aspects.

So, usually the data will be available in some tabular format; the format can be in the form of table or data frame or anything else. And the objective is to choose one of the column of that data set and to select a sample from that column only. For example, suppose I want to select a column, where the age of persons are mentioned.

So, from that big data set, I want to choose a particular column related to the age and then from those ages, I would like to draw a sample. So, here in this case, my population will be defined as the data which has already been collected through some online

automated resources. But that data, but that data might be of very huge size and I would like to draw a sample from there. So, this is the question which I would like to address today using the sampling package sample, right.

(Refer Slide Time: 03:33)

**Notations:**

Following notations will be used:

- $N$  : Number of sampling units in the population (Population size).
- $n$  : Number of sampling units in the sample (Sample size)
- $Y$  : The characteristic under consideration
- $Y_i$  : Value of characteristic for the  $i^{\text{th}}$  unit of the population  
( $i = 1, 2, \dots, N$ )
- $y_i$  : Value of the characteristic for the  $i^{\text{th}}$  unit of the sample  
( $i = 1, 2, \dots, n$ )

But before that, let us quickly have a review of the symbols and notations. And I would also try to address that, what are the meanings and interpretation of these symbols. So, we are going to use here following symbols,  $N$  as we had discussed earlier; this is going to indicate the population size that is the total number of sampling units in the population. Similarly,  $n$  that is our sample size, which is the total number of sampling units in the sample.

Now, there are two symbols  $Y$  and  $y$ . So, here one has to be careful; sometimes students get confused with the symbols. And so, whenever I am trying to use here the symbol  $Y$ ; that is the letter alphabet that is going to denote a random variable. And under the setup of simple random sampling, this is going to indicate the characteristic under study.

For example, if I have a data set here say, on several variable; suppose here some identification tag is given say, roll number of the students, then there are here marks in say physics, marks in mathematics and marks in chemistry and those marks are given over here, right.

So, in this case, suppose I want to sample the observations on the marks on physics; then my  $Y$  is going to be the marks in physics. And if I want to study the marks in mathematics; then my  $Y$  is going to be the marks in mathematics. And now under this variable, there are observations; for example, there are some students numbered 1, 2, 3, 4 and so on and suppose there are  $N$  number of students.

So, now the value of  $Y$  for those students, this is denoted by  $Y_i$ . So, this is essentially the value of the characteristic for the  $i$ th unit in the population. So, for example, suppose I say that, I am considering the height of the students. So, now, suppose there are say here 1, 2 up to here  $N$  number of students.

So, the value of the height for the first student that will be denoted by  $Y_1$ , the value of the height for the second student that will be denoted by  $Y_2$ , and similarly the height of the  $N$ th student will be denoted by  $Y_N$ . So, these are some numerical values.

Now, from this population, I try to draw here a simple random sample. In general as sample; but since we are considering here only the simple random sampling, so I will say this is a simple random sample. Otherwise this definition is valid for all types of sampling scheme in the further lectures.

Now, some of these  $y_i$ 's will be selected in this sample and we are assuming here that, we want to draw a  $n$  number of units in the sample; that means out of this  $N$  number of values  $Y_1, Y_2, \dots, Y_N$ , I want to choose here some  $Y_i$ 's which are  $n$  in number. So, number of  $Y_i$ 's will be equal to  $n$ .

Now, those  $Y_i$ 's, those  $y_i$ 's which are going to be selected in the sample and which are in  $n$  numbers, their values will be denoted by  $y_i$ . So, this is essentially the value of the characteristic for the  $i$ th unit in the sample. Now, here comes a confusion.

Now, in this population there will be units like  $y_1, y_2$ ; there will be some  $y_n$ , then  $y_{n+1}$  and says and  $y_N$ , because a  $n$  will always be less than  $N$ . Sometimes students try to think that we are trying to choose only first  $n$  number of values of variable; this is wrong, this is a wrong interpretation.

(Refer Slide Time: 08:39)

**Notations: Example**

$Y$ : 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)

$Y_i$ : Height of  $i^{\text{th}}$  student in the population

$y_i$ : Height of  $i^{\text{th}}$  student in the sample

WOR  $\binom{10}{3}$   $\binom{10}{3}$   
WR  $10^3$   $10^3$

So, now in order to explain this confusion, let me try to take here a example and from there I will try to make it more clear. So, suppose  $Y$  is the height of the students in the class and suppose there are 10 students in the class. So, my population size is here  $N$  is equal to 10. And suppose I want to draw here a sample of size  $n$ , which is equal to 3. So, I want to draw 3 units out of 10 units.

So, now I have here two types of notations; one will be here  $Y_i$  and another will be here a  $y_i$ . So,  $Y_i$  is going to be the height of the  $i^{\text{th}}$  student in the population. So, here I will be going from 1, 2 up to here 10 say; that means student number 1, student number 2 and student number 10 and their heights.

Now, from this  $Y_i$ , I am going to draw here only three observations and those three observation can be any three values out of  $Y_1, Y_2, Y_{10}$ . Why I am calling it any three? Because we had learnt earlier that, if you are trying to use the without replacement SRS,

WOR; then the total number of samples are going to be  $\binom{10}{3}$ .

And in case if you try to use here with replacement, then the total number of samples are going to be  $10^3$ ; let me write here more clearly  $\binom{10}{3}$  or  $10^3$ . So, these many samples you are going to obtain under WOR or WR.

So, now you are going to choose one of the sample over here. And this value which you are going to choose here, I am writing here; say  $y_i$ , where  $i$  goes from 1, 2, 3. Now, the confusion comes here; that in this case I am trying to take here  $i$  goes from 1 to 10, but in, but in the second case, in case of a  $y_i$ , I am trying to take the values say  $i$  goes from 1 to 3.

So, that is why sometimes students get confused that, how the range is changing; but remember one thing, these are the population value and these are the sample value. For example, my sample can be the  $Y_1$ , say  $Y_3$  and here  $Y_5$ . So, in or the sample can be  $Y_1$ ,  $Y_5$ ,  $Y_{10}$  or even my sample can be  $Y_1$ ,  $Y_2$ ,  $Y_3$ . So, these are the three values from the population which are obtained here.

Now, what I am doing how this; how these symbols are changing? Now, if you try to take here the first sample. So, I have drawn here  $Y_1$ . So, now, this becomes my here  $Y_1$ ; then I have taken  $Y_3$  the height of the third student, this becomes my here  $y_2$  and then I have taken here the height of the 5<sup>th</sup> student, so this becomes my here  $y_3$ .

Now, in the second sample, I have taken here the first value. So, this remains as here say,  $y_1$ . Then I choose fifth value, so now this becomes here  $y_2$ . And then I choose third value, which is here that height of the 10<sup>th</sup> student, so this becomes here  $y_3$ .

So, now, you can see here  $Y_5$  and  $Y_5$  both are chosen; but in the first sample, this is denoted by  $y_3$  and in the second case, it is denoted by  $y_2$ . And it is also possible that that whatever are the three, whatever are the first three values they are just obtain as such; so in this case  $Y_1$  will become a  $y_1$ ,  $Y_2$  will become  $y_2$ ,  $Y_3$  will become  $y_3$ , right. So, first for the heights of the first three student, they are simply; the they are also in the sample, ok.

(Refer Slide Time: 13:13)

**Example**

**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	$Y_5 = 155$
F	$Y_6 = 156$
G	$Y_7 = 157$
H	$Y_8 = 158$
I	$Y_9 = 159$
J	$Y_{10} = 160$

So, now suppose the values of height of the students in the population and their names are now given in this table. So, I am using here the names here A, B, C, D up to here J; there are 10 students that is their name and the  $Y_i$  is my height of the  $i$ th student.

So, this is my first student. So, the height of the first student is a  $Y_1$  is equal to 151 centimeters; height of the second student is  $Y_2$  is equal to 152 centimeter; height of the third student is  $Y_3$  is equal to 153 centimeter and similarly the height of the tenth student is  $Y_{10}$  is equal to 160 centimeter, ok.

Just for the sake of understanding and you do not forget what is your population and what are the value of the unit; what I have done here that, I have taken the name of the students in the order of alphabets.

Then I have defined the units here  $Y_1, Y_2, \dots, Y_{10}$  and their corresponding heights are starting from 151, 152 up to here 159; so that means the first value will correspond to 151, second value correspond to 152, third value correspond to 153 and similarly the ninth value correspond to here 159 and tenth value corresponds to 160. So, it is easy for you to remember that if I say that,  $Y_6$  is selected; that means 156 value has been selected, ok.

(Refer Slide Time: 15:03)

**Notations: Example**

Suppose Popn Value

$Y_1 = 151 \text{ cms.}, Y_2 = 152 \text{ cms.}, Y_3 = 153 \text{ cms.}, Y_4 = 154 \text{ cms.}, Y_5 = 155 \text{ cms.},$   
 $Y_6 = 156 \text{ cms.}, Y_7 = 157 \text{ cms.}, Y_8 = 158 \text{ cms.}, Y_9 = 159 \text{ cms.}, Y_{10} = 160 \text{ cms.},$

---

$y_i$ : Height of  $i^{\text{th}}$  student in the sample

Selected sample = 3<sup>rd</sup>, 7<sup>th</sup> and 9<sup>th</sup> student

$y_1 = Y_3 = 153 \text{ cms.}, y_2 = Y_7 = 157 \text{ cms.}, y_3 = Y_9 = 159 \text{ cms.}$

So, now you can see here that these are my population values, these are my population values. So, I have denoted  $Y$  to be the height and these are the values of  $Y_1, Y_{10}$ . Now, I try to do one thing that, I would like to draw here a sample of size 3. So, now, this  $y_i$ , this is going to be the height of the  $i^{\text{th}}$  student which is selected in the sample; this is very very important to keep in mind.

So, now suppose in my sample, we take the help of random number tables or software whatever it is. Suppose the three numbers which you select are 3, 7 and 9; that mean the 3<sup>rd</sup>, 7<sup>th</sup>, and 9<sup>th</sup> students are going to be in my sample.

So, that means, the 3<sup>rd</sup> student whose height is 153 centimeter, the 7<sup>th</sup> student whose height is 157 centimeter and the 9<sup>th</sup> student whose height is 159 centimeter will be in my sample. And the first observation will be denoted by  $y_1$  which is the value of  $Y_3$ .

Similarly, the second observation will be denoted by here a  $y_2$ , which is essentially the value of the  $Y_7$ . And thirdly, the 3<sup>rd</sup> value in the sample that is denoted by  $y_3$ , which is essentially the 9<sup>th</sup> value of the student in the population. So, now, I hope that after this explanation, there should not be any confusion in the notations and symbols of a  $y, Y, y_i, Y_i$ .

(Refer Slide Time: 16:51)

**Drawing of sample**

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)
```

Now, I come to my job, where I would like to draw a sample. So, now you can consider here suppose this is the data which is given to us and you can imagine that this data is very huge; means although I have taken a very value to explain you, but you can imagine that this is a very big data set, which you just cannot see from your eyes.

And suppose there are many more variables here also; means there can be weight, there can be age and whatever you want. And your objective is that, out of this big data set; you want to choose here one particular column, one particular variable and you would like to draw the sample from this population, the population is the height of the students.

So, now, I am considering here two columns; one is here name of the student and say another is the height of the student. And what I am going to do? From this given data set, I will try to draw a sample of the name of the students and a sample of the height of the students, ok.

So, now, first we have to define this data set. So, well I am not doing here; but I had done it in the introductory part of our software that how to create a data frame. So, there can be a CSV file, there can be a excel file or whatever it is, this data will be given to us.

So, now, since we are creating the data; so let me try to create this data into the format of a data frame. So, I have entered here the heights in this data vector, name in this data vector as a name and then I am trying to create here a data frame.



And the and that we had discussed that in order to create a data frame, the command is data dot frame and inside the parenthesis, you have to write all the variables. And one condition in defining the data frame is that, all the variable should be of the same length. So, this is possible here, because height and name they are of the length 10, ok.

(Refer Slide Time: 19:01)

```
Drawing of sample using R
> heightdata
  name height
1    A   151
2    B   152
3    C   153
4    D   154
5    E   155
6    F   156
7    G   157
8    H   158
9    I   159
10   J   160

> names=heightdata$name
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J

> heights=heightdata$height
> heights
[1] 151 152 153 154 155 156 157 158 159 160
```

*dataframe \$ name of variable*

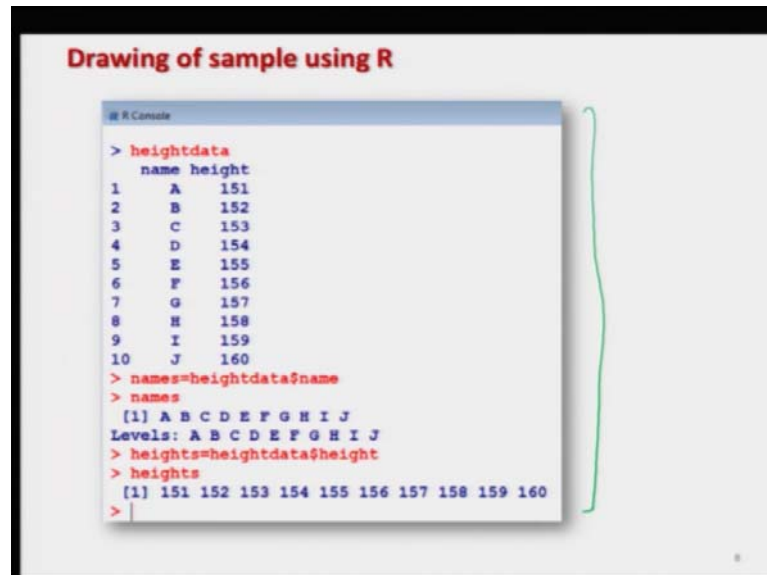
So, I do this thing and then means, the data frame which has been renamed as height data; it will look like this, I will try to show you on the R console also, but this will look like this. And we also had discussed earlier that, if you want to take out the data only on one of the variables say for example, name or say height; then the and the command here is, you try to write down the name of the data frame and then dollar and then name of the variable, this is the rule.

So, what I try to do here because for the sake of illustration; I try to take out this data set on the name of the students, and I try to store it here as a names n a m e s; whereas the variable name is name. So, the name of the data frame is height data, dollar and then name; name is coming from here, ok. So, now, if you try to see, you will get this data.

So, definitely this is a factor data in the form of A, B, C, D; so it will come like this. And similarly, the height is the quantitative variable; so I try to define here a name heights. So, what I have done, I simply have added s in the variable names. So, this will be the

name of the data frame heightdata, dollar and then the name of the variable height and this data will come like this, ok.

(Refer Slide Time: 20:36)



**Drawing of sample using R**

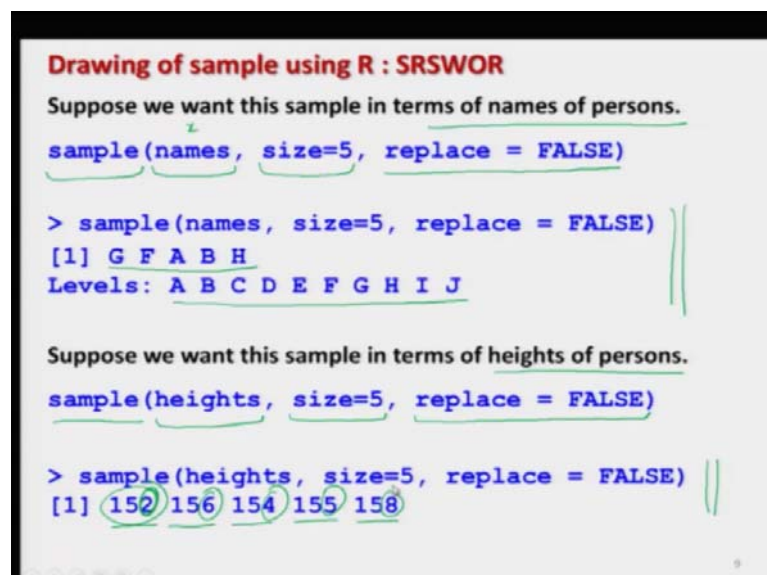
```
> heightdata
  name height
1    A   151
2    B   152
3    C   153
4    D   154
5    E   155
6    F   156
7    G   157
8    H   158
9    I   159
10   J   160

> names=heightdata$name
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J

> heights=heightdata$height
> heights
[1] 151 152 153 154 155 156 157 158 159 160
```

And here is the screenshot means, at the moment you can be sure that that this will work, right, ok.

(Refer Slide Time: 20:46)



**Drawing of sample using R : SRSWOR**

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

```
sample(names, size=5, replace = FALSE)

> sample(names, size=5, replace = FALSE)
[1] G F A B H
Levels: A B C D E F G H I J
```

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

```
sample(heights, size=5, replace = FALSE)

> sample(heights, size=5, replace = FALSE)
[1] 152 156 154 155 158
```

Now, I come to drawing the of the sample. So, in this case, once you have stored the variable names in your R console as names or height; then the life becomes very simple,

you simply have to use the command `sample`. And wherever you had given the data vector `x`, there you have to give the name of the variable. So, for example, suppose we want to draw a sample in terms of the names of the students or name names of the persons.

So, I will simply write here the command to draw the sample, `sample`; then earlier we had written here `x`, but now I am writing here the name of the variable in which the data has been stored, which is here `names`. And suppose I want to draw a sample of size 5. So, I give here `size` is equal to 5. And then suppose I want here simple random sampling without a replacement.

So I use here `replace` is equal to `false`. And then if you try to execute it for example, here is one out one such outcome and you can see here that we have got here the names of five students G, F, A, B, H and this is telling that what are the levels in which the data is available, ok.

Now, suppose I decide that ok, I want to have the data in terms of heights of the person. So, what I have to do, I simply have to use the same command; but now instead of the variable name `names`, I can use here another variable name `heights`.

And then the size of the sample which is here 5 and then SRS WOR which is `replace` equal to `false`. And once you try to use this command over here, you will get this type of say outcome. Say you are getting here 152, 156 and 154, 155, 158; so that mean the second unit which I can identify from the 3<sup>rd</sup> digit, 6<sup>th</sup> unit, 4<sup>th</sup> unit, 5<sup>th</sup> unit, and 8<sup>th</sup> unit have come to your sample and now here is the screenshot.

(Refer Slide Time: 23:10)

### Drawing of sample using R : SRSWOR

```
R Console
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J
> heights
[1] 151 152 153 154 155 156 157 158 159 160
> sample(names, size=5, replace = FALSE)
[1] G F A B H
Levels: A B C D E F G H I J
>
> sample(heights, size=5, replace = FALSE)
[1] 152 156 154 155 158
> |
```

(Refer Slide Time: 23:13)

### Drawing of sample using R : SRSWR

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

Sample of size 5

```
> sample(names, size=5, replace = TRUE)
[1] F F I E A
Levels: A B C D E F G H I J
```

Sample of size 8

```
> sample(names, size=8, replace = TRUE)
[1] C C D D J H G E
Levels: A B C D E F G H I J
```

So, now before going into the further details, let me first try to show you it on the R console, ok. So, first let me create the data frame. So, I copy this commands over here to save the time.

(Refer Slide Time: 23:30)

```

> 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)
> heightdata
  name height
1    A    151
2    B    152
3    C    153
4    D    154
5    E    155
6    F    156
7    G    157
8    H    158
9    I    159
10   J    160
> names=heightdata$name
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J
> height=heightdata$height
> height
[1] 151 152 153 154 155 156 157 158 159 160
> sample(names, size=5, replace = FALSE)
[1] G D I F J
Levels: A B C D E F G H I J
> sample(names, size=8, replace = FALSE)
[1] H G A J C D F I
Levels: A B C D E F G H I J
> sample(heights, size=8, replace = FALSE)
[1] 152 158 152 157 159 154 155 154
> sample(heights, size=8, replace = FALSE)
[1] 156 160 157 153 158 151 155 154
> sample(heights, size=12, replace = FALSE)
Error in sample.int(lengths, size, replace, prob) :
cannot take a sample larger than the population when 'replace = FALSE'
> ]

```

But I would request you to at least type the data with your own hand, because then you will be making couple of mistakes. And once you make those mistakes, you will understand that where you have to be careful, right. So, you can see here I have stored the values of heights, name and then the I have created the data frame over here and this data frame looks like this.

And from here if I try to store call the data on names; so this is going to be renamed as names and this is your here name of the data frame, height data followed by dollar and then name and you can see here. So, you can see here the data under this here names column, try to observe my cursor; this is here given here like this.

And similarly if I try to take here the data on heights; so I add here heights is equal to height data, dollar, heights. And you can see here that, this data is stored here; you can see whatever is the data under this height column here, this data is here.

So, now I try to execute my sample commands, right ok. So, now, I try to suppose, I want to take here a sample of here with respect to names. So, I will say here, try to take here sample of size 5 from the names; you can see here that these names are here, obviously means this outcome will not match with the outcome which is given in my slides, because this is a random number generation.

So, every time you generate it, with the probability that it will be repeated so soon that is very less. And similarly if you want to have a sample of size here 8; you can see here, this will come out to be of a different sample of size 8, right.

And in case now suppose if you want to have this sample with respect to heights; so I try to use the same command over here and I use here the variable height. So, you can see here now, this will draw a sample of size here 8 like this. And if you try to repeat it, this will be here like this, right. And if you try to say here size is equal to say here 12, you will get an error that we discuss in the last lecture; because it is simple random sampling without replacement, so the sample size cannot exceed the population size, well ok.

Now, you can see that it is not so difficult, which earlier looks to be. So, now, I try to consider the SRSWR and try to solve the same problem that we have the same data set; but now we would like to draw the sample by simple random sampling with replacement.

So, suppose I want to draw the sample with respect to the names of the person. So, I have to use the same command; the only difference will be here that, the replace will become here true. And once you try to do it here, you will get here this type of outcome and you can see here that these two values F and F they are being repeated here.

Similarly, if you try to obtain here a size sample of size 8; then use the same command, except that you change the size is equal to 8. And you can see here this is the outcome, where you can see here C is being repeated two time, D is being repeated two times and so on and the total number of sample size is 1, 2, 3, 4, 5, 6, 7, 8.

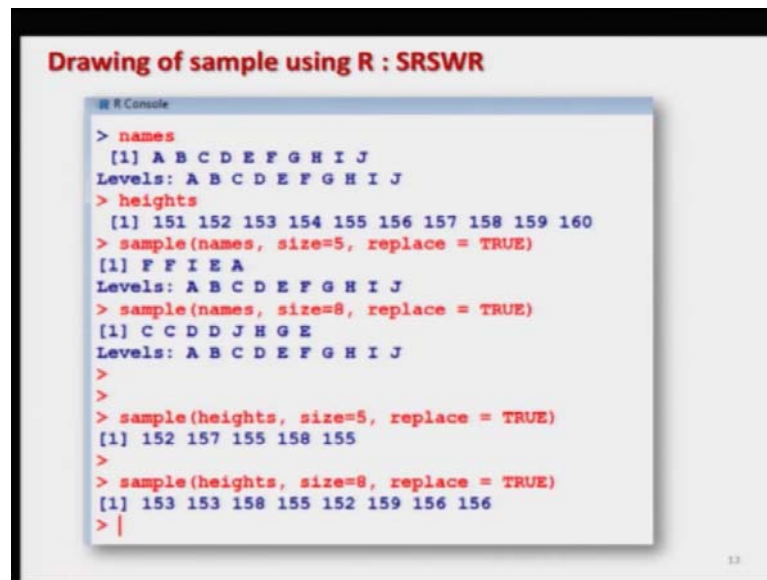
(Refer Slide Time: 27:28)

```
Drawing of sample using R : SRSWR  
Suppose we want this sample in terms of heights of persons.  
Sample of size 5  
> sample(heights, size=5, replace = TRUE)  
[1] 152 157 155 158 155  
  
Sample of size 8  
> sample(heights, size=8, replace = TRUE) ||  
[1] 153 153 158 155 152 159 156 156
```

We do the same exercise once again that, now we would like to draw the sample with respect to the heights of the person or heights of the student. So, I simply have to use the same command sample and I have to change the name of the variable here to be heights, size is your choice; suppose we are, we would like to find here a sample of size 5, so size is equal to 5.

And then you have to make here one change here, replace is equal to true and you can get this type of outcome and you can see here that this value 155 is being repeated two times. Similarly, if you want to have a sample of size 8, then one possible outcome is like this, where you can see here this value 153 is being repeated two times, 156 is being repeated two times and so on. So, this is your SRSWR.

(Refer Slide Time: 28:23)

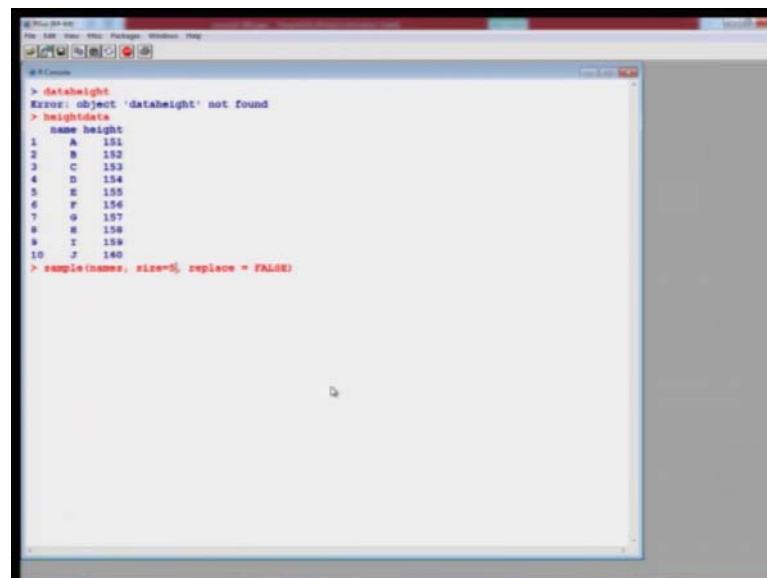


**Drawing of sample using R : SRSWR**

```
R Console
> names
[1] A B C D E F G H I J
Levels: A B C D E F G H I J
> heights
[1] 151 152 153 154 155 156 157 158 159 160
> sample(names, size=5, replace = TRUE)
[1] F F I E A
Levels: A B C D E F G H I J
> sample(names, size=8, replace = TRUE)
[1] C C D D J H G E
Levels: A B C D E F G H I J
>
>
> sample(heights, size=5, replace = TRUE)
[1] 152 157 155 158 155
>
> sample(heights, size=8, replace = TRUE)
[1] 153 153 158 155 152 159 156 156
> |
```

And the screenshot corresponding to this observation, you can see here; it is given here, where I am trying to give all this outcomes of names and heights.

(Refer Slide Time: 28:40)



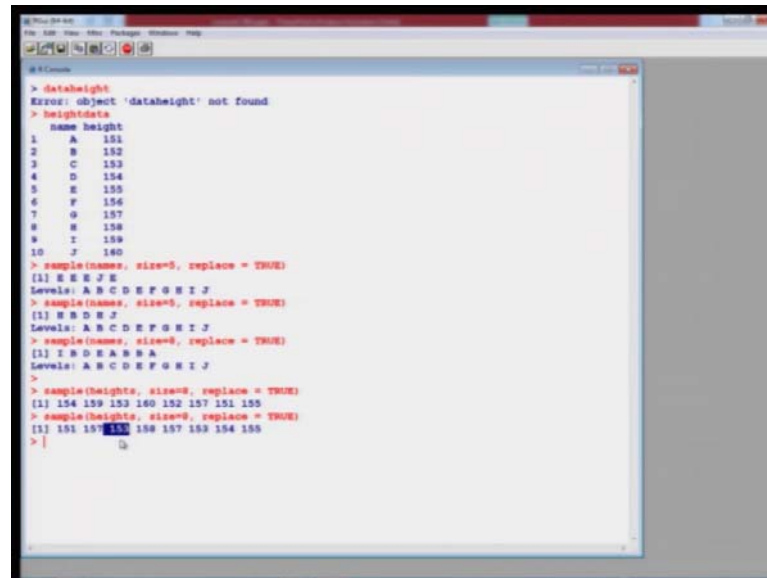
```
R Console
> dataheight
Error: object 'dataheight' not found
> heightdata
  name height
1.  A    151
2.  B    152
3.  C    153
4.  D    154
5.  E    155
6.  F    156
7.  G    157
8.  H    158
9.  I    159
10 J    160
> sample(names, size=5, replace = FALSE)
```

Now, let us try to do the same exercise on the R console over here. I try to clear my data and then just for your information I will keep this data before your eyes. So, it was height data, right.



And now what we have to do? I have to use the same command; suppose I want to find a sample of here names and suppose I want to find here a sample of size 5. So, I have to give write down sample name size equal to 5 and now replace will be here true.

(Refer Slide Time: 29:03)



```
> dataheight
ERROR: object 'dataheight' not found
> heightdata
  name height
1    A    151
2    B    152
3    C    153
4    D    154
5    E    155
6    F    156
7    G    157
8    H    158
9    I    159
10   J    160
> sample(names, size=5, replace = TRUE)
[1] E E E E E
Levels: A B C D E F G H I J
> sample(names, size=5, replace = TRUE)
[1] H H D H F
Levels: A B C D E F G H I J
> sample(names, size=8, replace = TRUE)
[1] I B D A B B A
Levels: A B C D E F G H I J
>
> sample(heights, size=8, replace = TRUE)
[1] 154 159 153 160 152 157 151 155
> sample(heights, size=8, replace = TRUE)
[1] 151 157 153 159 157 153 154 155
> |
```

So, you can see here that, this is here thus size 5 and you see you can see here very surprisingly four times E has been repeated, right. Well legally you are correct and similarly if you try to repeat it once again; then you can see here the H is repeated only two times. And if you try to obtain here a sample of size 8, for the names it will be something like this. And similarly, if you want to have the sample on the heights; suppose you want to have sample of size 8.

So, by SRSWR, so you can see here this is here, here the outcome and surprisingly once again no value is getting here repeated. So, now, if you try to look at only these values; can you really judge whether this data is from the SRSWR or WOR? But if you try to repeat it, possibly you can see here in the next outcome this 157 has been repeated here, 153 is also repeated here two times and so on, right.

So, you can see that it is not really difficult to handle the big databases. In data sciences whatever the data has been collected, that is stored automatically in some format; those can be text file, they can be some spreadsheets, or they can be some comma separated value file CSV files. But they are going to be very very huge, the number of variables

will also be very very large and usually you will not be interested in sampling only one or two something like this.

So, now this lecture gives you an idea that how you can sample from the required column or the required subsection of the data set. Now, the next question is this; this is possible that, once you have selected some values on a given variable, you would like to obtain the entire data set corresponding to those values. That means suppose, I have obtained here the name of the student as say here A and B in my sample.

So, now, I would like to have the complete row corresponding to the data of student A and B. Those things are not difficult; in R all this manipulations related to the data values, tabular values they are possible. Well, this is not the place where I can discuss all the things; because that is simply related to choosing some of the rows or columns of a data frame or a particular type of file inside the R.

So, for that my recommendation will be that, you try to look into the R software and help and how do you handle different types of files, and how do you do different types of data manipulation. So, that will complete your objective. But I have given you an idea; I have initiated the thought process. And I also have given you a hint that how you can start, how you can proceed. Well there can be different approaches; but this is one possible way.

So, you try to look into these things and try to take some more example. Although I am not doing here, but I request you that you try to take a table; then try to select some values and try to select the entire rows or the entire data corresponding to those selected sample values. So, you practice and I will see you in the next lecture, till then; good bye.