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

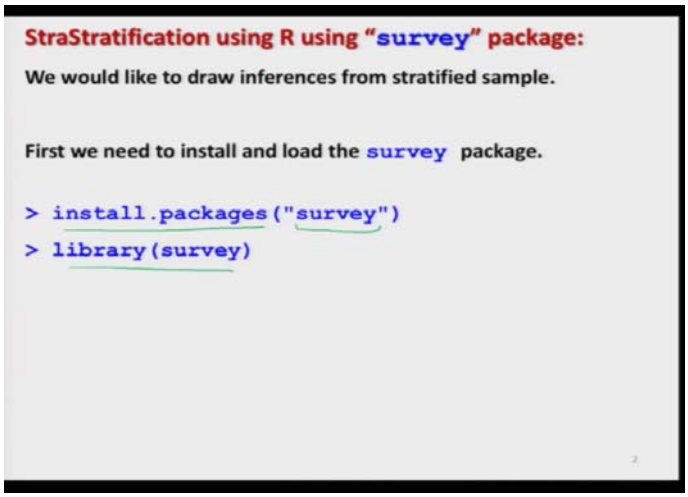
Sampling Theory with R Software
Lecture - 32
Stratified Random Sampling
Drawing of Sample Using sample survey Package in R

Hello, friends. Welcome to the course Essentials of the 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 the Sampling Theory with R Software we are going to learn the Stratified Random Sampling with R software and in this lecture I will continue from the earlier lecture.

In the earlier lecture, I had considered two possible ways to draw the sample in a setup of stratified random sampling. And in this lecture, I will use another package what is called as survey package to draw the stratified random sample. So, I will try to give you briefly this overview because there are many many options which are available in this package and that is a very vast package.

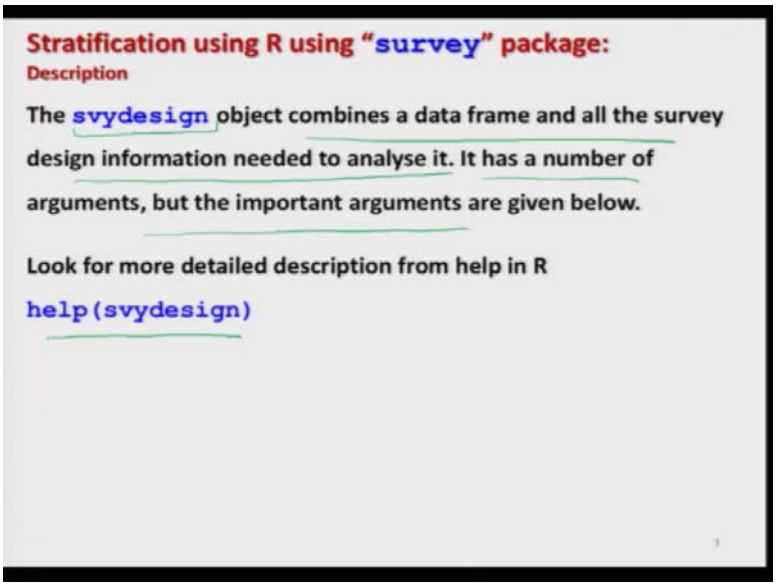
So, I will try to give you only a limited information which is related to the stratified random sampling and rest I would request you that you go to the help and try to look into all other options because all other types of sampling schemes are also incorporated in this package, ok.

(Refer Slide Time: 01:26)



So, let us begin our lecture. So, in this case first you need to install the package whose name is survey s u r v e y. So, for that you know the command here is install dot packages and then you need to load it using the command library as simple as that.

(Refer Slide Time: 01:45)

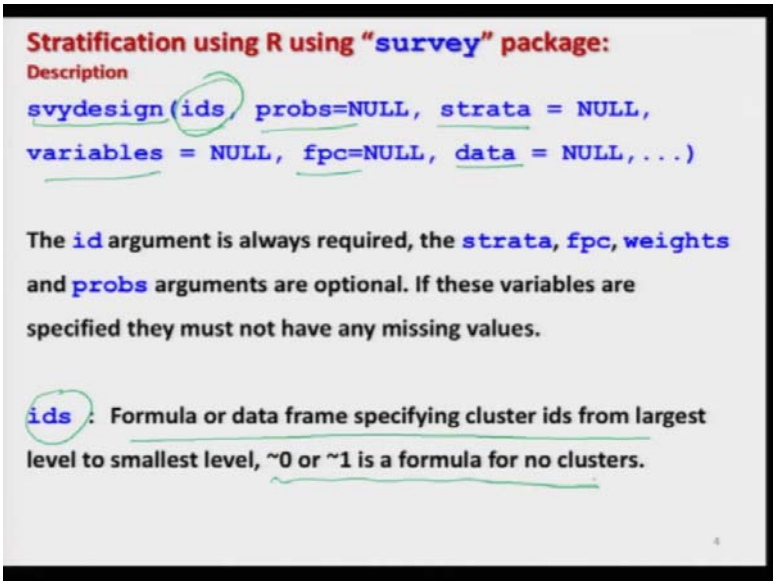


Now, then in this package there is a command s v y d e s i g n survey design that is the short form which is written here as a s v y that is survey and d e s i g n design. So, this command will be used to obtain the sample and actually that is a very general command and many things can be done through this command in this survey package.

And, actually this command this object combines the data frame and all the survey design information needed to analyze it and it has a number of argument, but we are considering here only the some argument which are needed for us. And, I would request you that if you really want to learn it more then just use the command here help svydesign and you will get here more information, right.

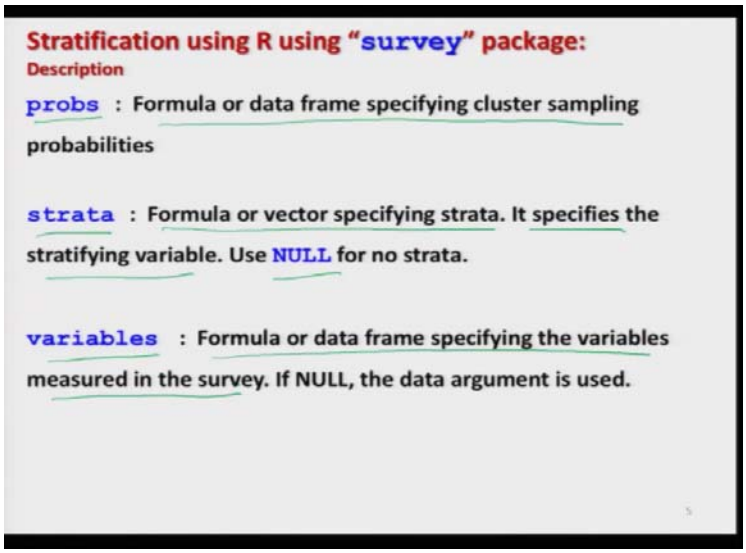
So, I will not go into those thing, but I will just give you a quick review of this of the means how to use it.

(Refer Slide Time: 02:44)



So, the command here is svydesign and then here there are some there are actually lots of option, but we are going to consider here ids, probs, strata, variables, fpc and data. So, these are important and ids is actually compulsory and all other things are optional. So, ids will actually indicate a formula or a data frame specifying the cluster ids from largest level to a smallest level and if you try to use this here 0 or 1, this is a formula for no clusters, right.

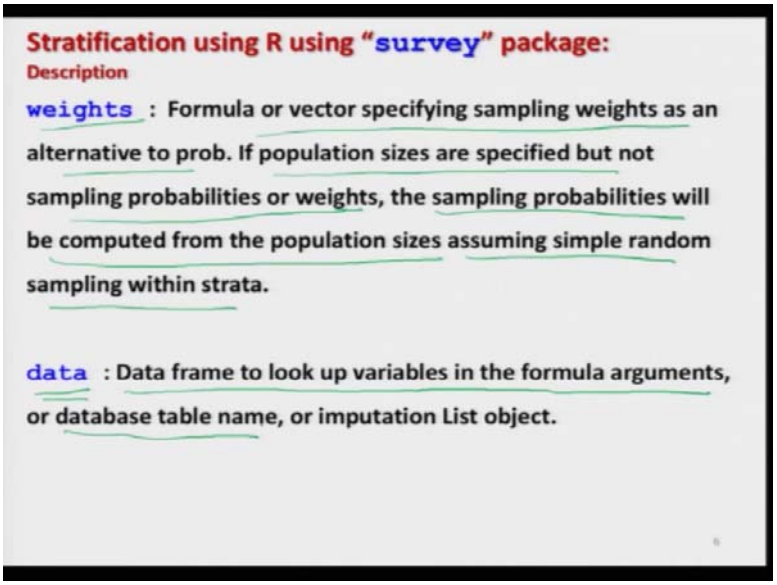
(Refer Slide Time: 03:32)



And, there is a function here there is an option here prob p r o b s – this will give actually this is a formula or a data frame which is specified the cluster sampling probabilities, but we have not considered the cluster sampling here. So, I am not going into the details. And, there is another option here strata.

So, strata is a formula or vector specifying the strata and it specifies the stratifying variable and if there are no strata in other types of sampling you can use here NULL, right. Next option here is variables. This variables is actually a formula or a data frame that specifies the variables which are measured in the survey, right.

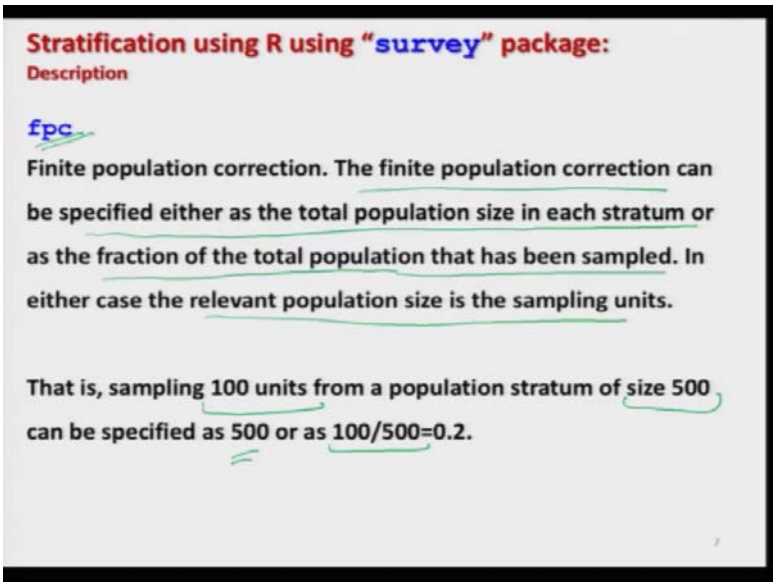
(Refer Slide Time: 04:17)



So, and then there is here option here weights and this is a formula or vector that is specify the sampling weights as an alternative to the prob which we have used for the for computing the probability. And, if the population size are specified, but sampling probabilities always are not specified then the sampling probabilities will be computed automatically based on the population sizes and using the simple random sampling within the strata.

So, that will automatically compute its. And, then there is a option here data this will specify the data, data frame to look up variables in the formula arguments where we have to compute or this will be a database table name or imputation list whatever you want.

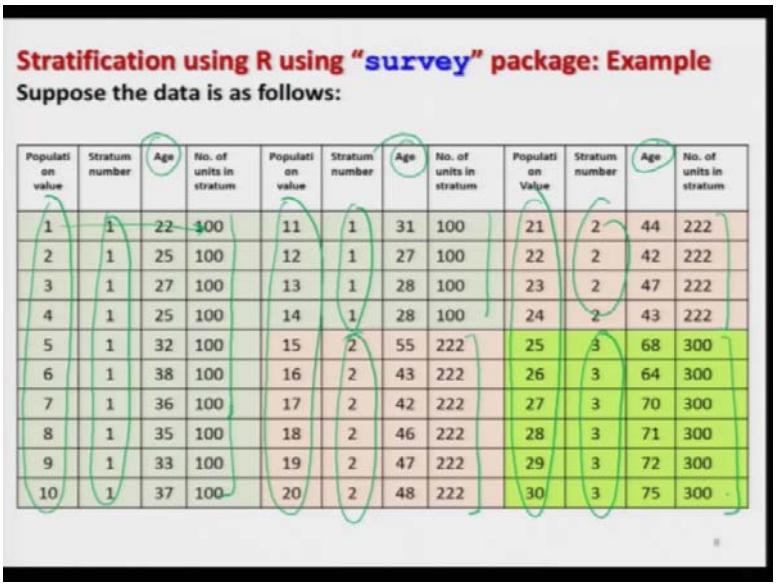
(Refer Slide Time: 05:10)



So, and then another option here is fpc. It is a finite population correction. If you remember we had discussed it in the case of simple random sampling. So, this is important for us the finite population correction can be specified either as a total population size in each stratum or this can also be expressed as a fraction of total population that has been sampled.

And, in either of the case the relevant population size is the sampling units, right. For example, if I say if you want to draw a sample of size 100 units from a population of size 500 then fpc can be specified as here 500 or 100 upon 500 which is 0.2. So, this is the way the information is given in this.

(Refer Slide Time: 06:02)



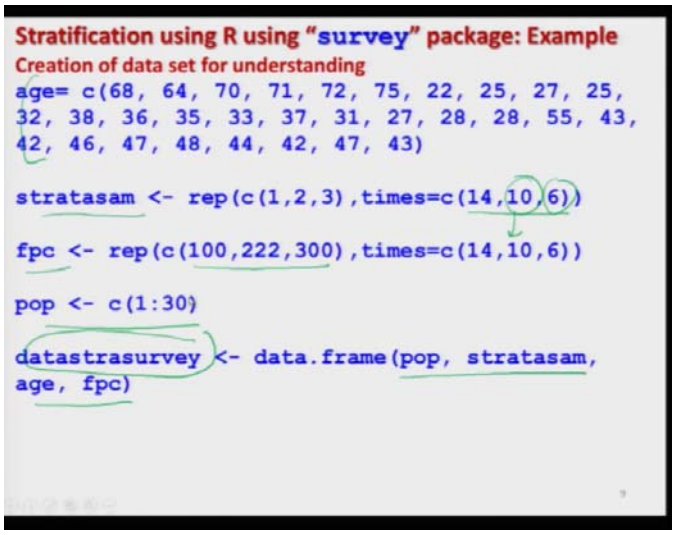
Population value	Stratum number	Age	No. of units in stratum	Population value	Stratum number	Age	No. of units in stratum	Population Value	Stratum number	Age	No. of units in stratum
1	1	22	100	11	1	31	100	21	2	44	222
2	1	25	100	12	1	27	100	22	2	42	222
3	1	27	100	13	1	28	100	23	2	47	222
4	1	25	100	14	1	28	100	24	2	43	222
5	1	32	100	15	2	55	222	25	3	68	300
6	1	38	100	16	2	43	222	26	3	64	300
7	1	36	100	17	2	42	222	27	3	70	300
8	1	35	100	18	2	46	222	28	3	71	300
9	1	33	100	19	2	47	222	29	3	72	300
10	1	37	100	20	2	48	222	30	3	75	300

So, here I am taking here a simple example in which I have constructed the data set artificially. So, I am taking here a population of size N is equal to 30. So, you can see here those units have been identified the number 1 to 30 here, you can see here. And, then I have some observations on age which are given in this column here for the 30 persons and these persons have been classified into different strata and you can see here this is for stratum number 1.

These many people are in the stratum number 1, these many people are in stratum number 2 and these many people are in stratum number 3. And, then I have to define here the n_1, n_2, n_3 for each of the units. So, I have to define here suppose I say in the stratum number 1 there are 100 units which are given here.

So, every observation has to be given a value here say 100. This is the way this package works. And, then in the stratum number 2 just want to make a difference I am taking here the value 222 in each of the corresponding to each of the values and stratum number 3 has suppose 300. So, n_1 is equal to 100, n_2 is equal to 222 and n_3 is equal to 300.

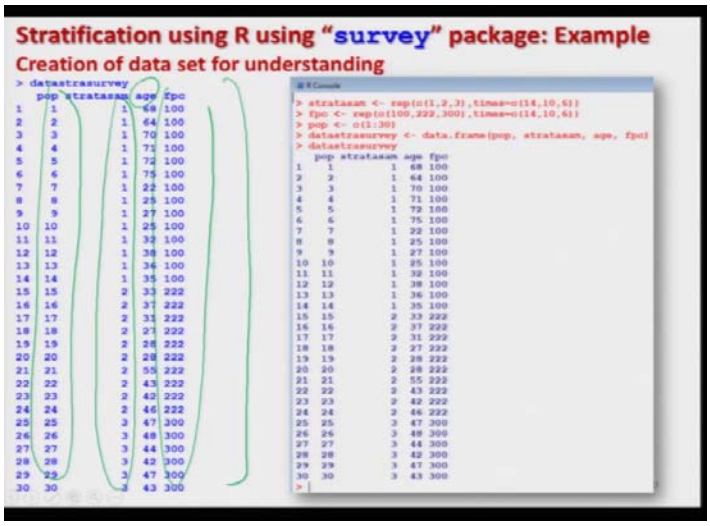
(Refer Slide Time: 07:26)



Now, I need to create a data frame for this data set. So, I have entered here the ages in this variable age and then I have defined the stratum number means 1, 2, 3. So, you can see there are 14 values corresponding to 1, 10 values corresponding to 2 and 6 values corresponding to 3.

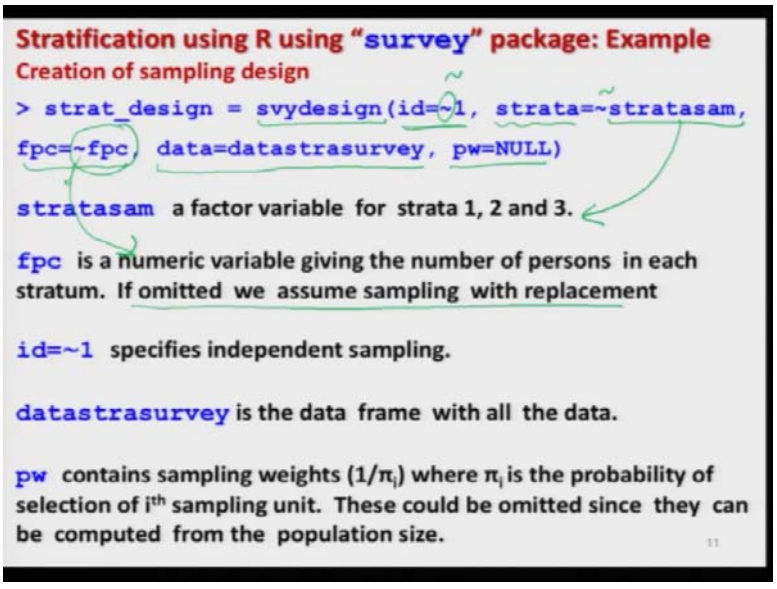
And, then I have defined here the fpc as 100, 222 and 300 times and they will be incorporated with the same number times as the stratum numbers and the population here is the numbers 1 to 30. And, all these data has been combined in the framework of a data frame and the name of this data frame is given here as a datastrasurvey. So, that means, data for the stratified sampling under the survey package.

(Refer Slide Time: 08:14)



And, if you try to see this is how the data frame will look like. So, here is the population, here is the stratum number 1, 2 and 3 and here is the fpc value and here is the value of wage age, right.

(Refer Slide Time: 09:29)



Now, the question come how we are going to draw the sample from this package under the stratified sampling. So, remember one thing this survey package is actually for constructing the design of a survey, right. Design of a survey has different things which we are not considering here, but from the output of this one we will try to extract the information on the stratified sampling. This is the idea in the survey package.

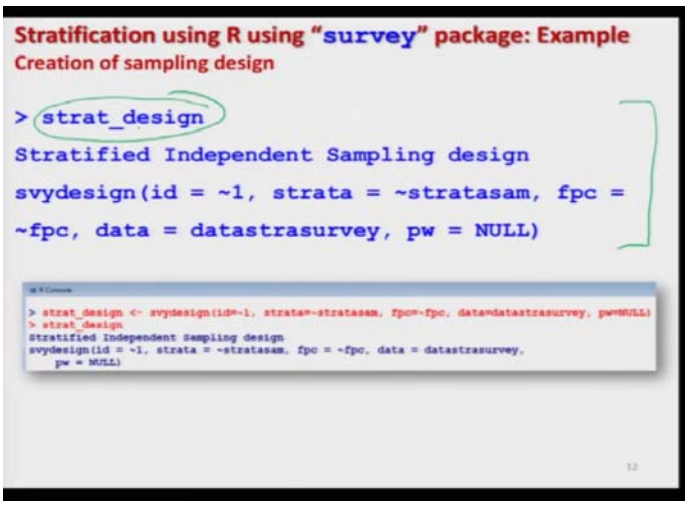
So, the command here is we have to use the command svydesign and then we have to give here id is equal to here 1 and this id equal id is equal to 1 with this equivalent sign which is available on your keyboard. This will indicate that we are going to use the independent sampling.

And, whatever are the strata means how the strata have to be created, this I am trying to indicate by this equivalent sign followed by the value of the variable or the name of the variable. And, then whatever is my fpc either that can be computed automatically or we are supplying it from outside.

So, here I am trying to say that ok this has been supplied from outside and this is stored under the variable name say fpc and this has to be specified with this equivalent sign. And, then what is my data? From where we have to draw the sample? This is datastrasurvey and . So, this is we are and the probability and pw is equal to here NULL because this contains the sampling weights though we are not using it here, right.

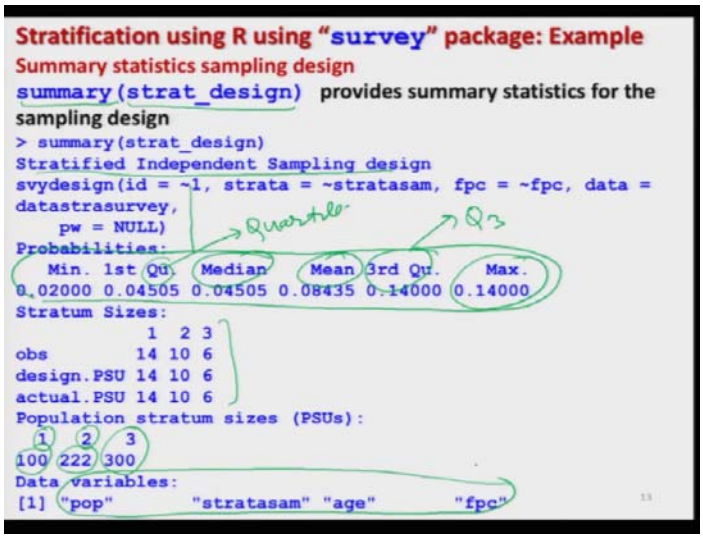
So, now this stratasam will take the value here 1, 2, 1, 2 and 3 and fpc here this will be a numerical variable giving the number of person in each stratum and if this is omitted we assume that sampling is with replacement, right and so, now we try to execute it.

(Refer Slide Time: 10:29)



And, once you execute it you will get here an outcome like this one. This will not give you the sample directly. So, it will look like this. So, and then from this outcome you have to extract the information.

(Refer Slide Time: 10:44)



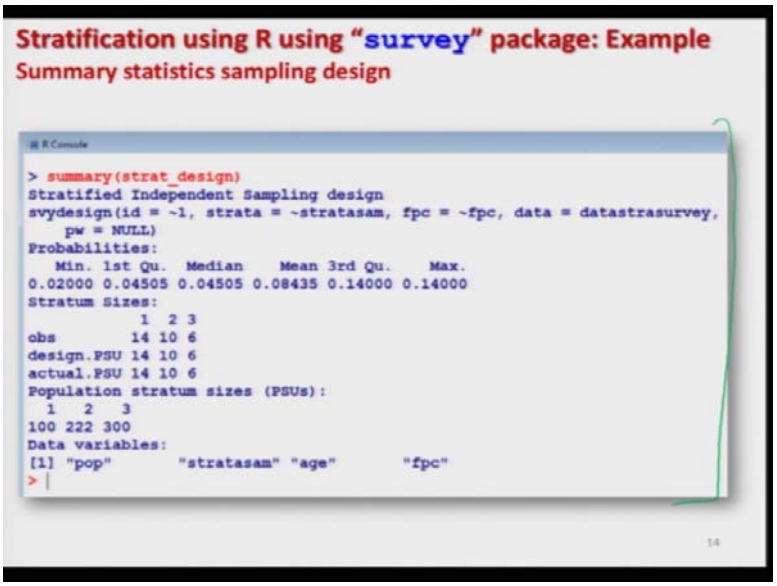
So, for that whatever is the outcome here this I have stored in a variable name say strat underscore design. So, that is the stratified design what we have constructed ok. Now, I use here a command summary; summary and then you have to write down the name of the design that you have constructed within the parenthesis and this will give you the summary statistics for this design.

So, you can see here it is giving you the name that which sampling design has been used this is the command which you have used and here you can see here it is giving you the summary statistics that the minimum value first quartile. This is here quartile and this is the median that is the second quartile, this is his arithmetic mean, this is here the third quartile say Q_3 which is the denoted as in general and this is here the maximum value.

So, this is giving you lots of information about the sampling design and then it is trying to give you here what are the number of observations which have been considered in the construction of strata and what is the value of strata sizes. This is also given here for example, it is giving the strata number 1 has 100 units, strata number 2 has 222 units and strata number 3 has 300 units.

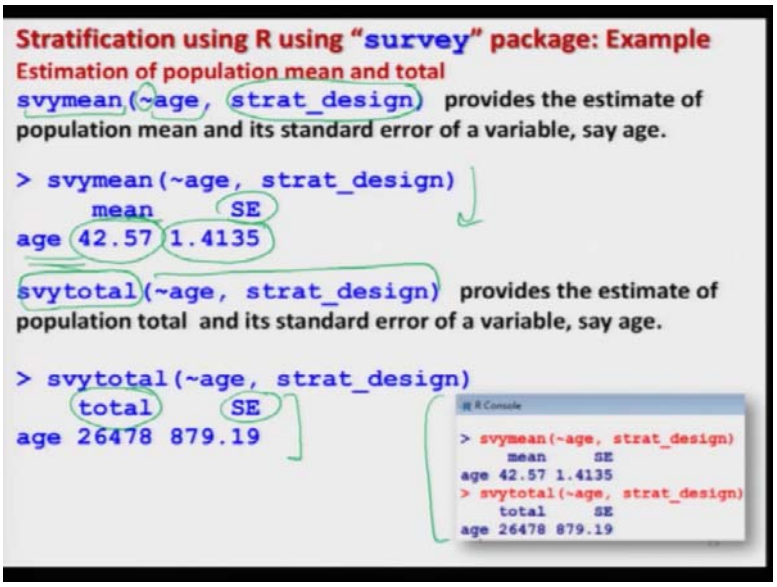
And, what are the variables in the data set? This is these are given here population stratasam, age, fpc.

(Refer Slide Time: 12:14)



And, this is here the screenshot of the same outcome, right.

(Refer Slide Time: 12:22)



Now, in case if you want to find out the mean that is the stratified mean then here in this package we have this option that we can use here a command svymean. It is a something like survey mean, that is, the mean from the survey package. So, mean of here what? Suppose, if I want to suppose we have a variable here age, although we have considered

here only one variable just for the sake of understanding. So, I try to give here the name of the variable.

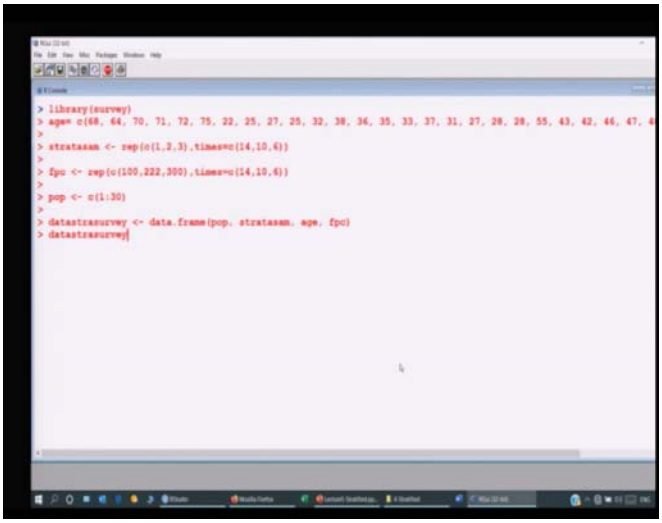
And, remember one thing here in this package thus the rule is to write down the name of the variable using the equivalent sign. So, you write down the name of the variable on which you have drawn the sample and you would like to find out the arithmetic mean or the stratified mean of which data which has been stored in the design that you have obtained under the name strat design.

So, if you try to execute it here you will see here it is not only the outcome of a mean, but it will give you the mean as well as standard error of the sample which you have obtained. So, they can see here this is the arithmetic mean or say $\overline{y}_{st}$ actually weighted arithmetic mean of the data that you have obtained or that you have sampled and this is the corresponding standard error.

And, instead of mean if you are interested in the say this is the total of the values then you have the command here s v y t o t a l. So, this will also give you means the style here is the same thing what you have to use here and then if you try to execute it this will give you here outcome like this one. So, it will give you here the total estimate of total and this is the estimate of variance of total, right.

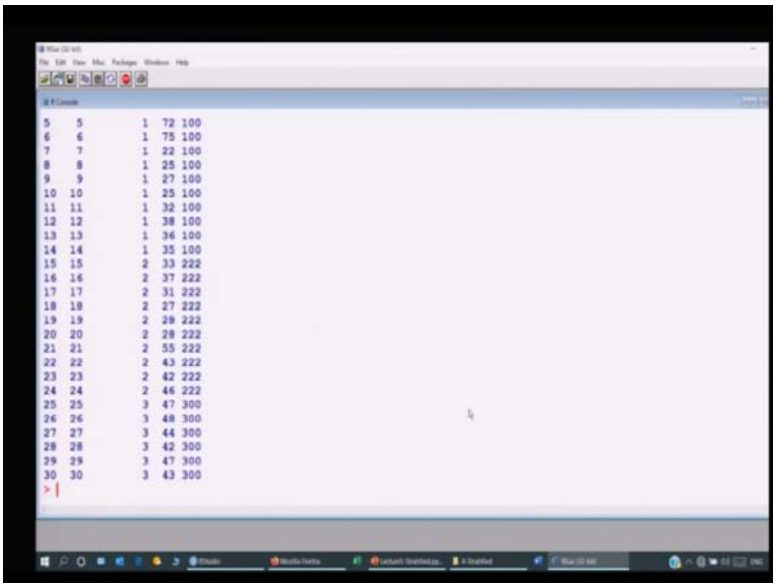
So, this is how it works and you can see here now this is the screenshot of the outcome over here and now I try to come on the R console.

(Refer Slide Time: 14:20)

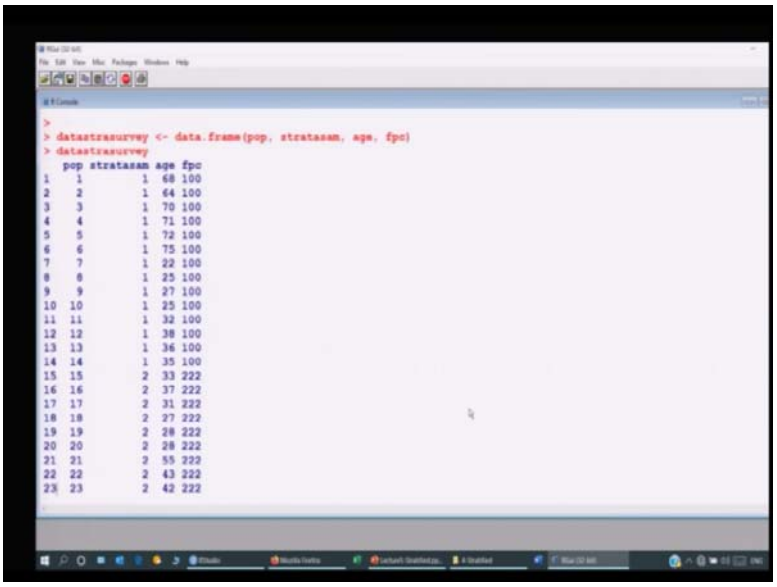


So, let me first up load the package survey which I already which was already there on my computer and I tried to create here this data frame. So, you can see here that is the datastrasurvey.

(Refer Slide Time: 14:38)

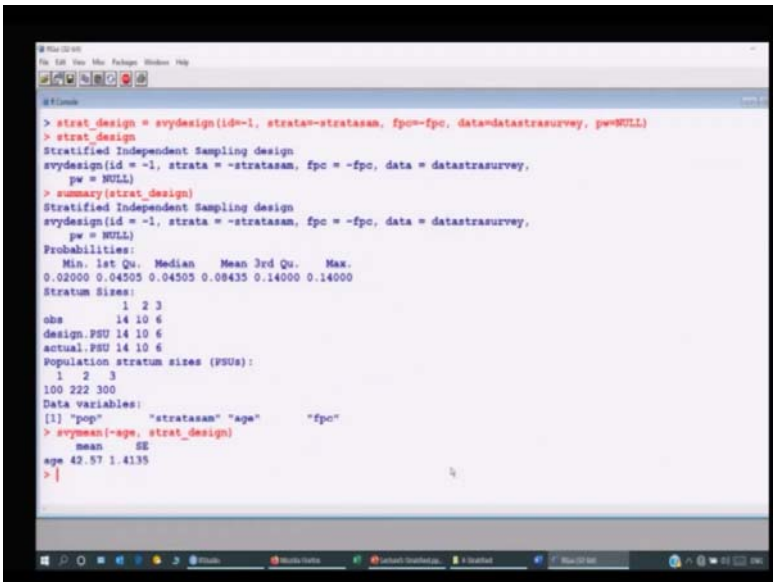


(Refer Slide Time: 14:39)



So, you can see here this will look like this. So, this is here the same data set which you have obtained here, right. And now, I try to use here this command and I try to see what is my outcome.

(Refer Slide Time: 14:59)

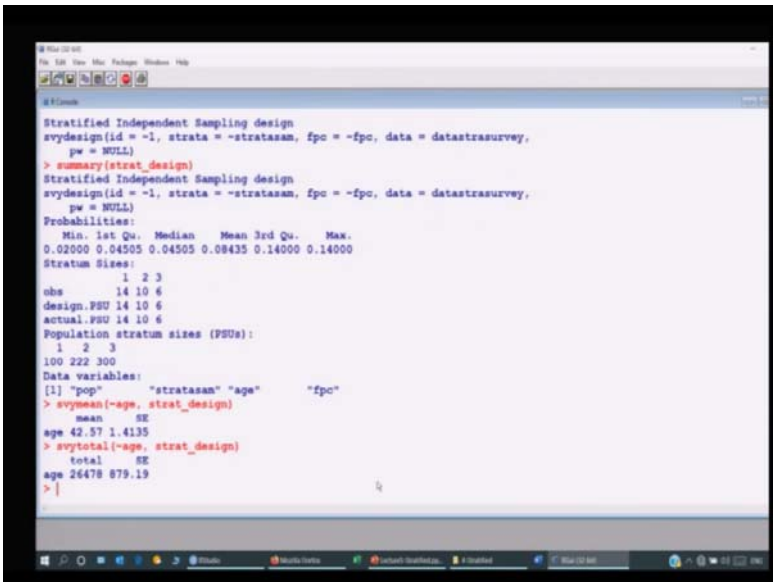


So, you can see here I copy and paste this command over here. So, you can see now here what is the outcome? strat underscore design is the variable in which I have stored the outcome. So, you can see here this is the same thing what we had obtained and, now I would like to see what is contained in it.

So, I try to go for here say the command here summary and you can see here it is giving me a similar information. Well, this these values will be different from the values which are reported in my slides because it means every time you draw a sample, the sample is going to be different.

If you want to find out the survey mean then you have to use here the same command here on the R console. You can see here this is survey mean and if you want to use here the survey total here. So, it will come out to be here like this, right.

(Refer Slide Time: 15:51)



So, you can see now here this is the same outcome which you have obtained in the slides, ok, right. Now, I have completed that how will you try to execute the drawing of the stratified random sample using the R software. So, you have seen that in R different people are contributing different packages with different objectives. So, whenever you want to use a package first you have to go through with the help menu and try to see what the authors are trying to provide through the software, then you have to match what is your need.

Sometime those needs can be fulfilled directly using the direct commands and sometimes you have to do something more as writing a small function, a small program which is not difficult. Means, if you ask me to find out or write a program to find out the $\overline{y}_{st}$ or estimator $\overline{y}_{st}$ possibly means I can do it in 15 minutes also.

I will although I am not a very good programmer right, but a good programmer will take possibly only 5 minutes. Because means everything is there mean is there the command mean will provide you the mean of the sample, the variance command is there that will provide you the variance of the sample. You simply have to just sum them in the required way.

So, I will now request you that you please try to take some examples and try to practice it. Now, I will be stopping with the topic of sampling theory and as I said if I continue on

the topics of sampling theory, possibly that will be a complete course on sampling theory. But, my idea was that I wanted to show you that how the classical statistics can be connected to the data sciences and what are it is utility and without these tools you cannot work in data sciences.

So, I explain you that how this stratified you sampling can be used in getting a representative sample in different types of huge data sets. So, now, R is the possible way out which can give you the direct implementations for the computation different types of computation, different types of sample drawing and means other things.

So, I will be stopping here on the chapter of stratified random sampling. At every stage I was telling you that ok because of mathematical complexities I am unable to do this unable to do that. But now here I will try to present a computational procedure with which you can compute many many things in spite of the fact that you may not handle them theoretically. So, you try to revise these thing, you try to practice and I will see you in the next time.

Till then, good bye.