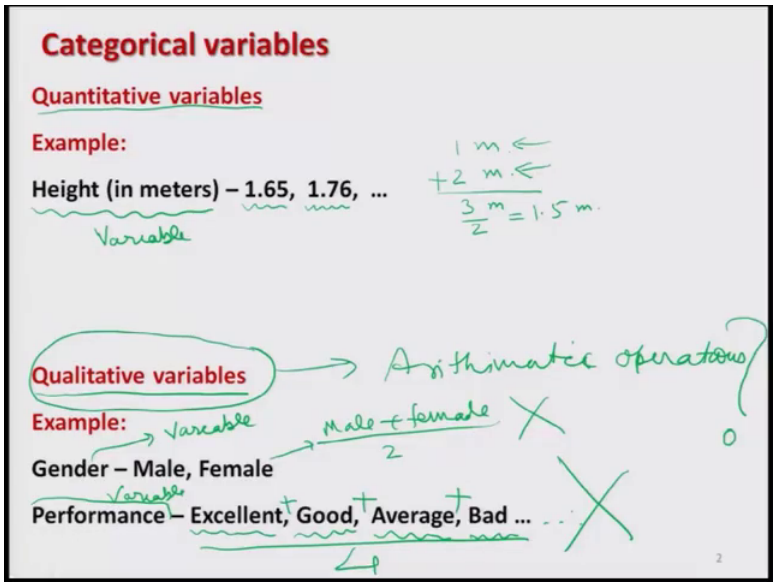
Introduction to R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 22
Factors

Welcome to the next lecture on Introduction to R Software. Now in this lecture and in the next lecture we are going to take up a topic on Factors.

So, we will try to understand what are factors, and how they are implemented in r in this lecture with some examples, and in the next lecture I will try to continue the same topic with some more example, and some other aspects of factors. So, first question comes what is a factor, but before that we have to understand some background, there are different types of variables and in particular when we are talking in terms of statistics, then there are 2 types of variables, one is quantitative variables, and another is qualitative variables.

(Refer Slide Time: 01:18)



Variables are the 1 on which we try to collect some data or the numerical values for example, if I say my variable is height, then I try to collect the data on height of the persons for example, suppose if I say my variable here is height, this is my variable and I am measuring the height in meters, I measure the first person height comes out to be 1.65 meters, then I measure the second person this comes out to be 1.76 meters and so on.

In this case you can see that all the observations, which are coming out for this variables are quantitative, means I can express them in some form of numerical values and those numerical values have interpretation.

For example, if I say there are two persons height of one person is 1 meter, and height of another person is 2 meters, then I can say that the height of the second person with 2 meter is twice of the height of the first person whose height is 1 meter, or if I try to find out the total of the height is I can say 1 plus 2 this is equal to here 3 meters, and in case if I want to find out the average height, I can divide the 3 meter by here 2, and I can see that here this is 1 point 5 meter.

So, such variables are called quantitative variables means, I can obtain the observations in some quantified way, another type of variables are qualitative variables for example, suppose if I have a group of persons, then I can classify them into 2 groups 1 male and say another female.

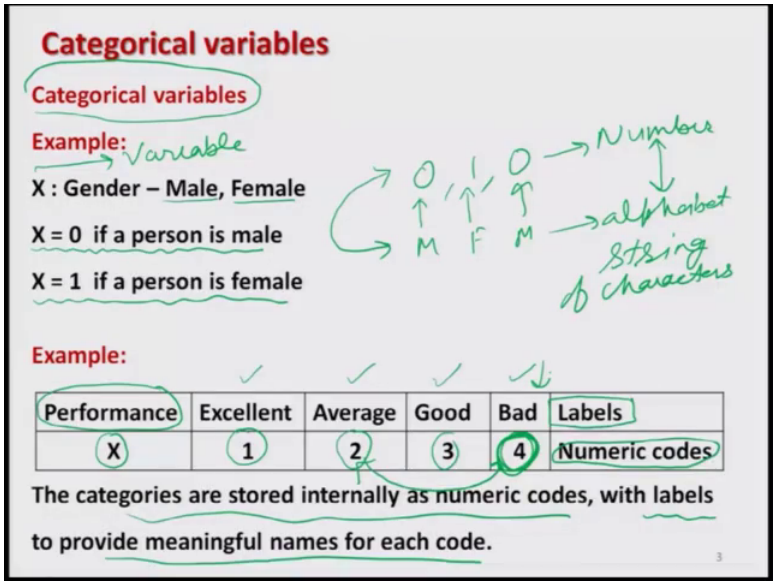
So, in this case my variable here is gender, and that is giving me 2 types of values 1 is male and say another is female. Similarly I can take another example, in which I try to define my variable here as say performance, this performance can be in exam performance can be in any say sports event or say anywhere.

The performance cannot be measured, but that is classified as performance is excellent good, or average, or bad, or something else. In this case if you try to understand, the variable gender is male and female the performance is excellent good average bad these are very well understood, what do we mean for example, in a simple race if a child is running faster than another child, then we say that his performance is better than the earlier one, but we cannot say this is 1 time better, or 2 times better, or 3 times better.

So, we cannot quantify it this type of variable they are called qualitative variable, but my problem is that whenever we have a quantitative variable I can do all sorts of mathematical manipulations for them finding out arithmetic mean or say anything else, but when we have this qualitative variables, then if I say I want to have here arithmetic operations, then does it make any sense, can I take here; male plus female, divided by 2 as the average of gender, it has no meaning. Similarly if I try to take here excellent plus, good plus, average plus, bad divided by here 4, does this make any sense no this is garbage this has no meaning.

So, the difference in the operations of a quantitative and qualitative variables is that in quantitative variables, we can do all sorts of mathematical manipulations whereas, in the qualitative variables we cannot do directly all the mathematical manipulations, but is still these qualitative variables make sense, they have certain meaning they have certain interpretations. So, the question is this how should we handle them 1 option is to use the categorical variable, and this is our objective to understand what is a categorical variable right.

(Refer Slide Time: 07:05)



Now, let me take here the same example, and suppose if I say my variable is denoted by here x this is my variable, and it takes here 2 values male and female and we define this variable x here as say x takes value 0 if a person is male, and x takes value 1 if a person is female.

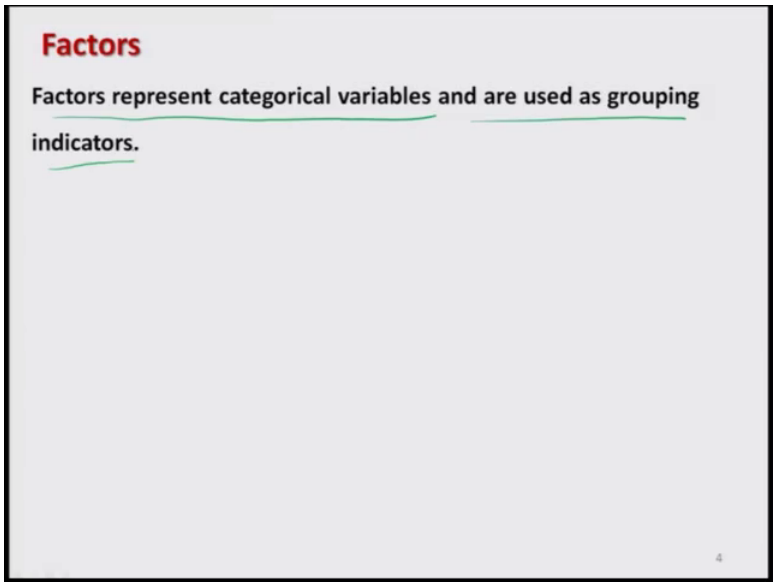
So, what we have to do we simply have to take a person, and then write whether he is male or female and instead of writing male or female I will write 0 and 1. If I get a data see here 0 and see here 1 and then here 0. That means, this mean that first of all we had a male person, then a female person, and then again a male person, this is a categorical variable. So, if you try to see what are we doing here this is here a number and this is here a sort of alphabet, or in our language this is a string of characters, and I am trying to make a 1 to 1 correspondence between the number and string, I have made here 1 to 1 correspondence between 0 or 1 and male and female.

So, now I have mapped the string character into a numerical value, but you have to keep in mind that this numerical value 0 or 1 that is only a sort of indicator, that is only indicating the absence and presence of some qualitative variable. Similarly if I try to take another example where I am trying to classify my variable performance that we are denoting here by here X into 4 categories, excellent by 1, average by 2, good by 3 and bad by here 4, but it does not mean that, when I compare 2 and 4 this does not mean, that the average is half of the bad it has no meaning, this is not the interpretation, but 1 2 3 4 they are simply some numerical codes, which are indicating the presence or absence of particular type of quality in terms of excellence average good or bad.

So, this is strings such as excellent, average, good or bad, they are called labels and these numbers they are called as numerical codes or numeric codes, and what is really happening if you try to see, these categories are actually internally stored as numeric code, and these labels are chosen in such a way such that they provide a meaningful interpretation and name for each of the code for example, I take here the numeric code 4, and then I am choosing here a proper label say here bad; that means, 4 is indicating bad.

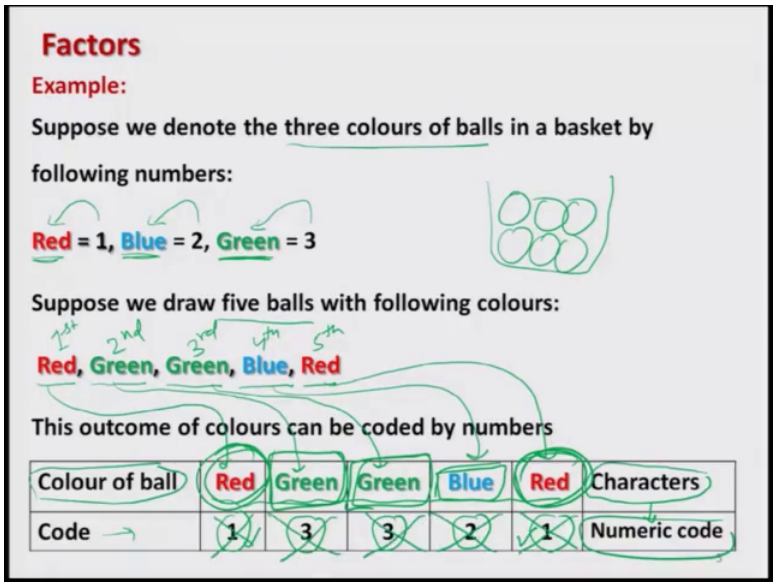
So, if you observe here again I have made here 1 to 1 correspondence between the string of characters like as excellent, average, good and bad, with respect to the numeric codes 1 2 3 4. Now, there is a 1 to 1 mapping between the swing character and a numerical value right, ok.

(Refer Slide Time: 11:25)



So, now we come to what are the factors; factors, represent the categorical variables and are used as grouping indicators, whatever we have understood what is categorical variable in the language of r this is called as factor, right.

(Refer Slide Time: 11:54)



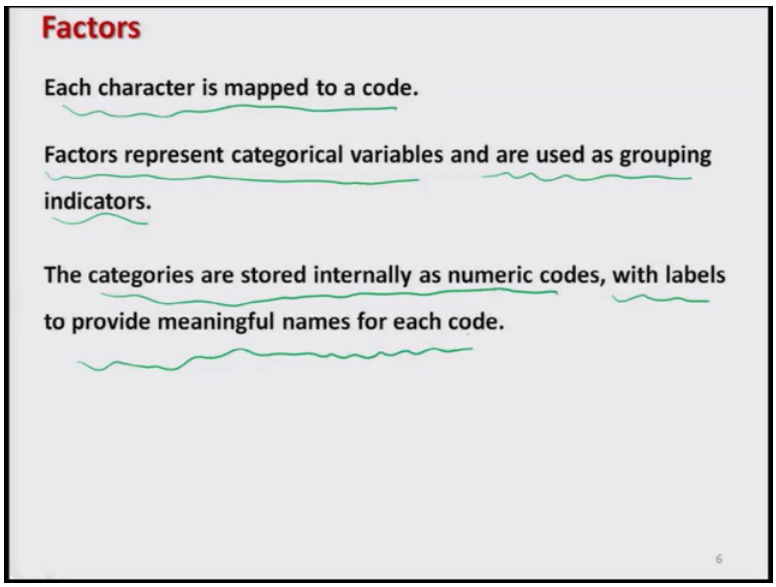
So, now we try to understand what do we really mean by factors, in the context of this categorical variables to understand it better, let us try to take a very simple example, suppose I have a basket which has 3 colors of balls, and these colors are red blue and green which I have written here in the same colors also for a better understanding, now red blue and green these are the 3 character they are the character strings, we cannot do any mathematical manipulations over this, I cannot say red plus green plus blue divided by 3. So, we try to give a numerical code, so I give number 1 as numerical code to red, number 2 as numerical code to blue, and number 3 as numerical code to green.

Now, we have this basket in which there are, so many balls are there and we try to draw here 5 balls, and suppose I get the 5 balls of the following colors red color first ball, second ball green color, third ball green color, fourth ball blue color, and fifth ball red color. So, this is first, this is second, this green is third, fourth is blue, and red is here fifth ball, now I try to code this color using the numerical codes. So, I try to define here a color of the ball which is a character string, and then I define here the code.

So, we have got first the red color ball it has numerical code 1, then we have got the green color ball whose numerical code is 3, then we have got the third ball green color

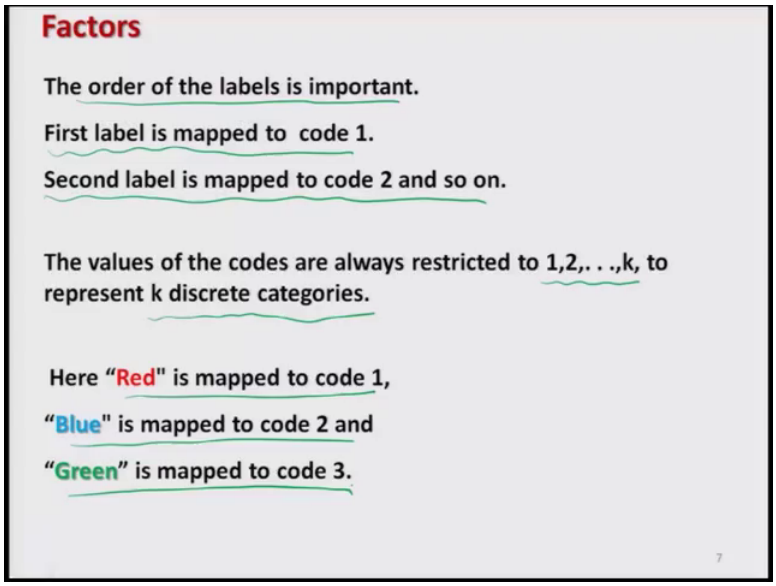
whose numerical code is again 3, then we have obtained the fourth ball blue color ball whose code is 2, and then finally, we have got the red color ball whose numerical code here is 1. So, you can see here by this red green blue and red colors we have defined a the string of characters, which has some meaning, and then they are connected or map with a numerical code 1 2 and 3 right. And again I would say 1 2 3 they are only the indicators means, if I try to make in this case as 1 plus 2 plus 3 divided by 3 this does not have any meaning right.

(Refer Slide Time: 14:49)



So, you can see from this example here that each of the character is mapped to a code, and factors represent the categorical variables, and these categorical variables are used as grouping indicators for example, if you see in this earlier example we have got here 2 red balls, and 2 green balls, and 1 blue ball. You can see here I can group that I have got 2 times numerical code 1, 1 time numerical code 2 and 2 times numerical code 3, you can see here 1; 1, time 1 2 times or here I will make it more clear by crossing it, then 2 only 1 time 3 1 2 two times right. And these categories are numerically stored internally using the numerical codes, and the corresponding labels they provide the meaningful names for each of the code, this is the characteristic of a factor.

(Refer Slide Time: 16:20)

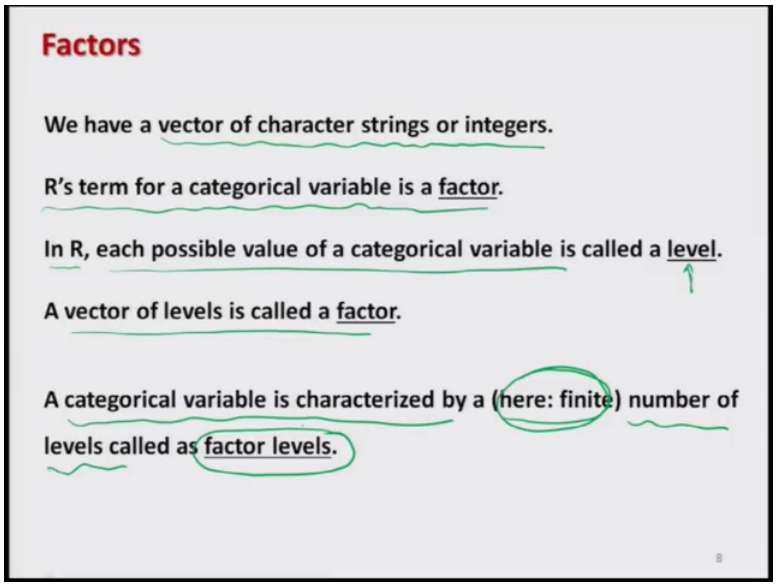


And in this case the order of the label is also important many times for example, if I want to know that which of the ball came first, then order is important, but if I am not interested that which of the ball came first or which of the ball came in the second row, then order is not important, but when we are talking of the 1 to 1 mapping of the character string with the numerical codes. Then in that context the ordering is also important for example, in this case the first label is mapped to code 1, second label is mapped to code 2, and so on.

So similarly in general, these codes 1 2 3 up to here is see here k they denote the k discreet categories.

For example we had here 3 colors ball red blue and green, so I have used 1 2 and 3. So, these numbers indicate the category for example, here you have seen the red is mapped to code 1, blue is mapped to code 2, and green is mapped to code 3, in the example that we have considered.

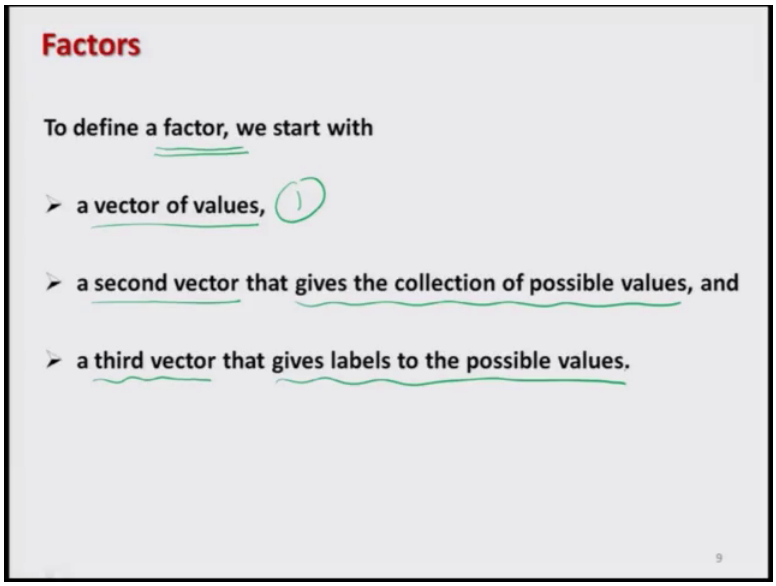
(Refer Slide Time: 17:44)



Now, after this we try to combine the concept of R and the factor as well as categorical variable, whatever we have understood in r usually we have a vector of character string or say or integers, 1 basic fundamental what you have to understand that R's term for a categorical variable is a factor, in R that is termed as factor ok, and in R each possible value of a categorical variable is called level, and a vector of levels is called as factor that is a more precise definition.

Suppose we have finite number of groups, what we have consider in this case. Now a categorical variable is characterized by the number of levels, which are termed as say factor levels do not worry do not get confused we will try to understand all this terminology with a simple example then it will be more clear.

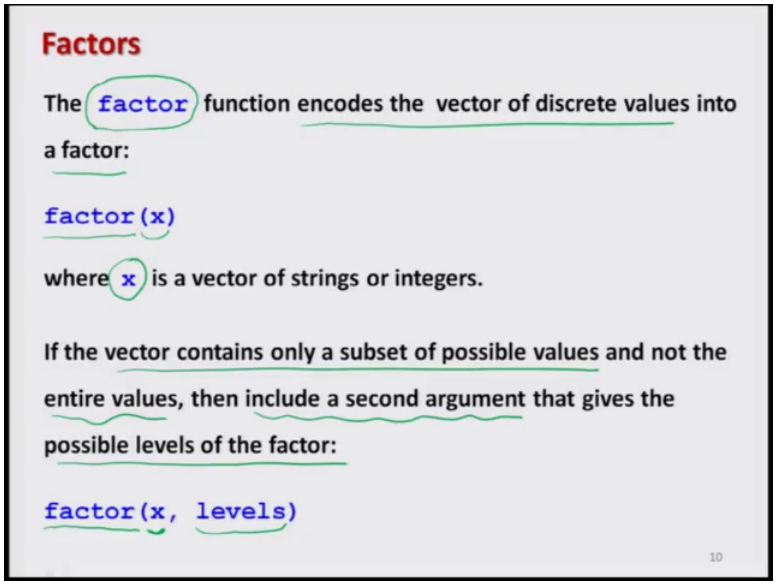
(Refer Slide Time: 19:17)



So, in order to define a factor in R we have the following steps. First step is take a vector of values; then choose a second vector that gives the collection of all possible values, and then choose our third vector that gives labels to the possible values. And if you follow these 3 steps you can get a factor this again I would say do not worry we will try to take an example, and we will try to understand these concepts.

But before that let us try to understand how the factor is dealt in R. So, first we briefly try to give you an idea of the syntax and then I will come up with an example.

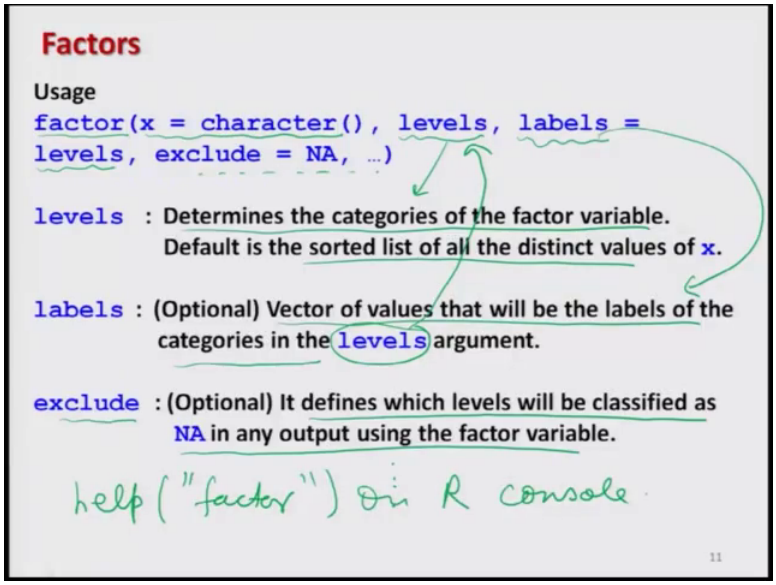
(Refer Slide Time: 20:13)



This word factor is the command to encode the vector of discrete values into a factor. So, factor is a function that is well defined in the base package of R, and suppose I have a vector here denoted by here x which is a vector of string or some numerical values integers, then the syntax is we write factor all in small letters and inside the bracket we write the x vector on which we want to create the factors right.

Now there are several possibilities, in case if the vector contains only a subset of possible values not the not all the values, then we include a second argument, and this second argument, gives the possible levels of the factor, and this is denoted by the syntax factor inside the bracket the vector of string of integer say x, and then followed by here another name or say argument levels.

(Refer Slide Time: 21:41)



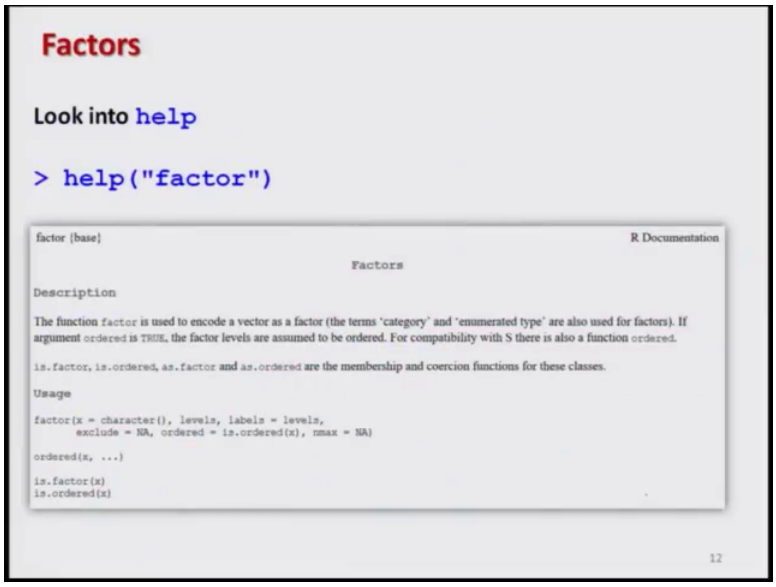
A more general syntax for the usage of this factor here is that we try to write down here the factor, then the data inside numerical or is a character strings, and then we try to see here levels we define the labels, and then we define the labels please try to differentiate between levels and labels; and this labels, are equal to the levels their number has to be same, and then we have here some more options something like exclude NA and so on.

So, this levels what we have given here, they determine the categories of the factor variable, and the default value is that the they try to consider the sorted list of all the distinct values in x, and then this here level; label, is an optional argument, and this is a vector of values that will be the labels for the categories in the levels argument which is

given here. And similarly we have here exclude option this is this is again an optional argument which defines that which of the level will be classified as not available in any output using the factor variable.

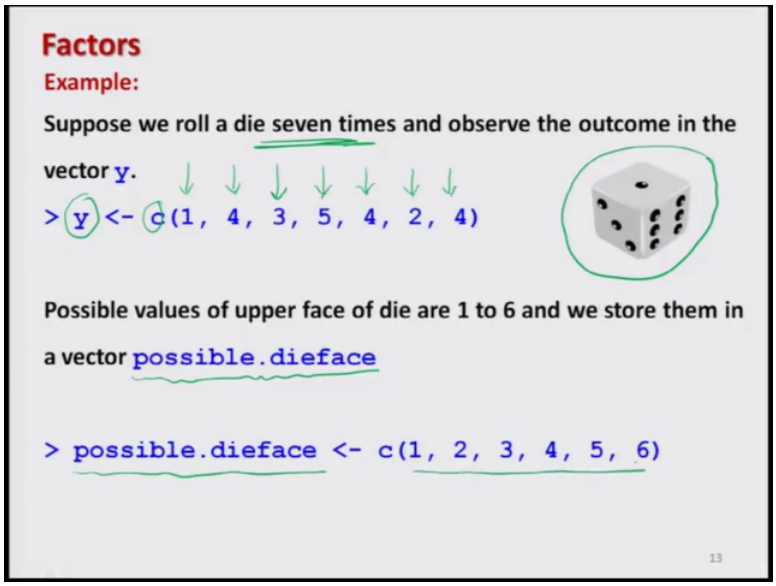
But, it is not the end there are more things about this and if you really want to have some more details on the factor, I will say simply type help say factor on the R console, and you will get more details.

(Refer Slide Time: 23:32)



But instead of going into those details, I will now try to take up an example and here you have seen that I have given you a us a brief snob snapshot, but it is continuing. Further even, I would suggest you that you please try to go through this help and then try to see what do you obtain, but basics we have covered.

(Refer Slide Time: 23:50)

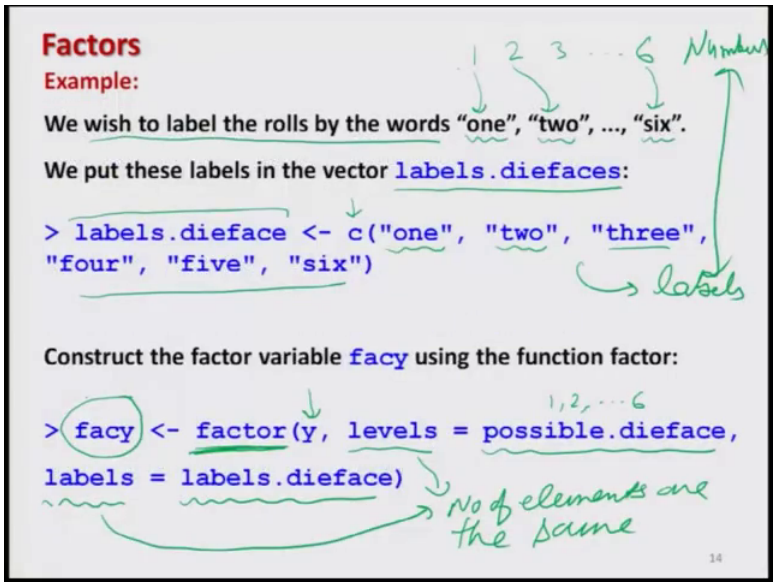


So, now let us try to take a very simple example of a dice what is the dice that we have seen the in our childhood we have played, various games using this type of structure if we try to roll it and on the upper face there comes a number and we try to get the reading.

So, there are 6 faces of this dice and the possible numbers that come on the upper face that can be 1 2 3 4 5 or 6, now suppose we roll this die 7 times, and we observe the outcome and whatever is the outcome that is stored in the vector y. So, in the first throw I get the value 1, in the second throw I get the value 4, in the third throw I get the value 3, and in the fourth throw we get the value 5, in the fifth throw we get the value 4, 6th throw we get 2 and finally, in the 7th throw we get the value 4, on the upper face of the dice.

So, all these values they are combined using the c operator, now there are 6 possible values that can appear on the upper face of the dice, and I try to store all those values 1 2 3 4 5 and 6 in a different vector, and we call this vector as a possible dot die face I am just trying to take a longer name which has more meaning, so that you can understand it you can keep in mind easily. So, this vector possible die phase contains 6 values 1 2 3 4 5 and 6 ok.

(Refer Slide Time: 25:38)



Now, suppose we wish to label the rolls of the die which are actually 1 2 3 4 up to 6 by the words by the character strings like as one 1 two 2 and up to 6 six.

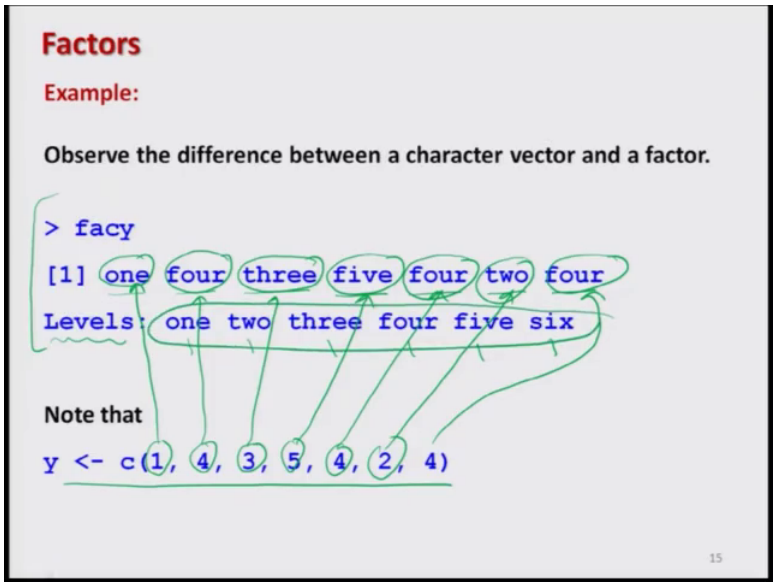
So, this 1 is indicating this thing this 2 is indicating this 2, and up to here this 6 is indicating this 6 right. Now we try to put all these labels inside a vector and we call this vector as labels dot die faces this is again a longer name but I have taken it, so that you can understand the meaning.

So, I try to write down all those characters one two and so on inside this vector, and I combine them by here c. Now you can see here I have here 2 things. one is these are my labels, and these are my here numbers, and I want to make a 1 to 1 mapping between them and based on that I want to do all my manipulations.

So, I use the command factor, now the question is where you want to use your factor command. We use the factor command over the values which are given by here y vector, and what are the levels that you want to define these levels are your numbers 1 2 3 4 5 and 6 there are 6 levels of a roll of a die. So, I am saying that these levels are the possible die phase, which are taking the value 1 2 3 4 up to here 6 and what are the labels what label I have given it, I have given the label number 1 as one 1, number 2 has a label two, number 3 has a label the double e and so on.

So, this labels are going to be defined by here this vector labels dot die face, and now you can see here that the number of elements in levels and labels are the same right, now let us try to implement it, and means all this values they have been stored in this variable here, facy I have given the name facy; that means, these are the factors of y.

(Refer Slide Time: 28:35)

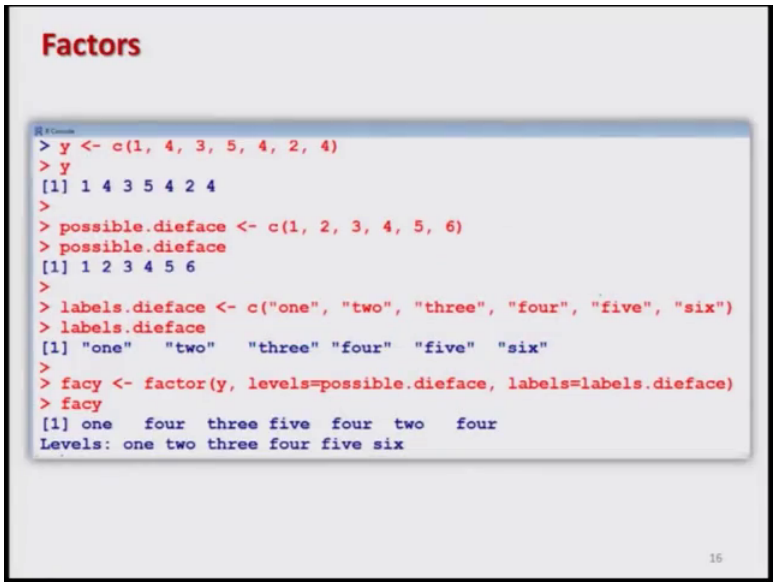


And as soon as you do it here you get this type of outcome, facy gives you here 1 4 3 5 4 2 and 4, and it also gives you the levels say here as say 1 2 3 4 5 and 6 what does this mean.

Now, if you try to recall how you started and what you wanted to do, you had this vector y which had the values 1 4 3 5 4 2 and 4, now you have converted the values which are the numerical values in y into a character string, how this 1 is denoted here by this 1, this 4 here is denoted by here, this 4 four this 3 here is denoted by here this thr double e this 5 is denoted here by here this five 5, this 4 is denoted by here four here 4, 2 is denoted by here two 2, and 4 is denoted here by four 4 so, you have converted a number into a factor.

And what are the levels of the factor, they are given over here that there are 1 2 3 4 5 and 6 these are the different factors which are use in this factorization right.

(Refer Slide Time: 30:18)



So, let us try to do this thing in on the R a console also, but you can see here this type of outcome we are going to get here right, but you let us try to do it here. So, first of all I try to define my here y, so this y comes out to be here like this, and then we try to create the vector possible die phase which are here like this, and you can see here the value of this vector is given by here this thing, and then we try to create another vector for the labels of the die phase which we have done here.

And you can see here that what are the different labels, that you have want to give this are the correct characters strings, and then I try to create here the factors of my here y based on my requirement, and this gives me here this outcome, and this is the same outcome which is given on my slide. So, you can see here that when you try to define here label; label, is a character which is given here inside the inverted commas so; that means, this is a character and what are the different levels; levels, are the numerical values which are given here as numerical values and finally, you started with some numerical values here, and you have converted them into here the into strings.

So, we have done what we wanted to do, and this is here the screenshot of the same thing, now, we stop here well they were several concept they were several links starting from variable to qualitative variable; quantitative variable, and from qualitative variables to categorical variable; categorical variable, to see here a factor, and then we took a

several examples to understand the meaning and interpretation of each of the terminology.

Please try to revise this thing try to settle them inside your mind, and take this type of some more example try to create some more example yourself. And then try to see are you getting the same outcome and practice it. And we will continue again with some more example and some other aspects in the next lecture, till then goodbye.