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

Lecture - 48
Introduction to Statistical Functions - Introduction, Frequency and Partition Values

Hello friends, welcome to the course Foundations of R Software. Now, from this lecture we are going to begin with a new topic. So, you can see, right from the beginning of this lecture we were talking of different types of command, different types of say the control structure and different ways of data handling etc. So, now the question comes here how are you going to use them? Where are you going to use them? What are the different applications of this R software and why this R became so popular.

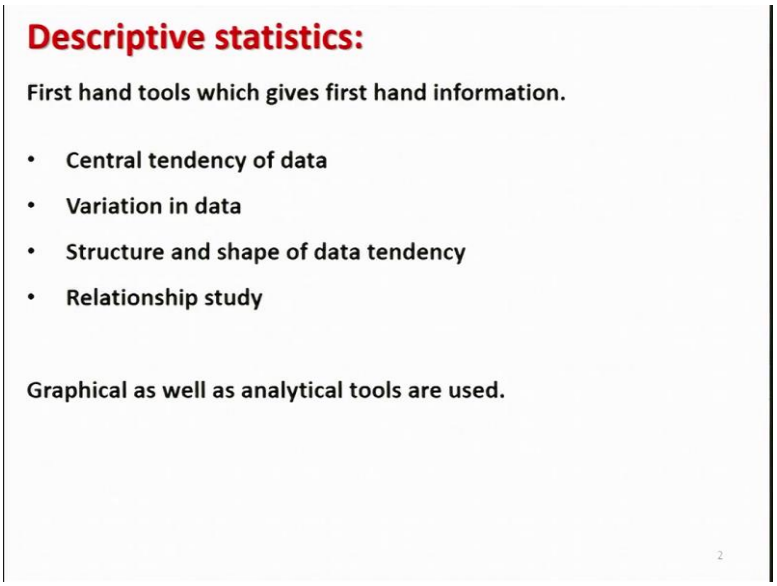
So, one of the reason that why R became so popular was that it was very useful in the statistical programming. Well, in this course we are not going to talk about the statistical programming, but definitely I would like to give you here some basic operations in statistics and which I believe that you have learnt it learnt them in the elementary classes.

So, that should not be a problem for you to understand means anything related to statistics. You already have done the topic like frequency table and then counting the values, partitioning values, quantiles, quartiles etc. So, my objective now from this lecture is that, ok. I would try to give you some idea that whatever you have learnt earlier how it can be executed in the R software.

So, after just giving you some elementary concepts of frequency table quantiles etcetera I will try to take the topic of graphics. Because, in graphics I need these concept for example, if you want to create a histogram or a bar plot, then you need this concept that of what value you are going to create the graphics.

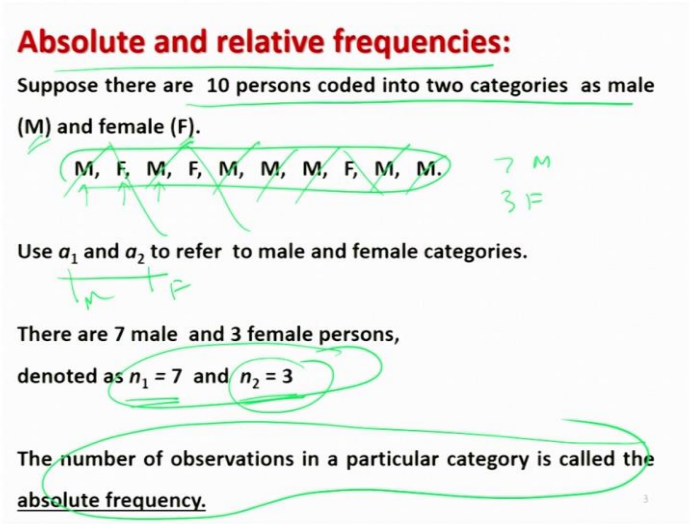
So, that is why I will try to consider some very basic elementary statistical topic in the lecture today. And after that, I will try to take up the topics of graphics and I will try to show you that how you can successfully produce graphics in the R software. So, let us begin our lecture and we try to understand some basic elementary statistical concept, ok.

(Refer Slide Time: 02:40)



So, you know that whenever you get some data, the first job which you are going to do is that you want to employ the tools of descriptive statistics, right. Descriptive statistics means, you want to have the idea about the central tendency of the data, variation in the data, structure and shape of the data, relationship study and in order to understand this we would like to use the graphical as well as these analytical tools, right.

(Refer Slide Time: 03:05)



So, now we are not going into this details in this course, but I would try to simply show you some very basic elementary tools of descriptive statistics we are trying to compile

the data in a statistical way. So, first we try to understand the concept of absolute and relative frequencies, right. So, this is a very elementary concept and I believe that you have been taught it in some elementary classes only, elementary means possibly in class 6, 7, 8, 9, 10 and so on, right.

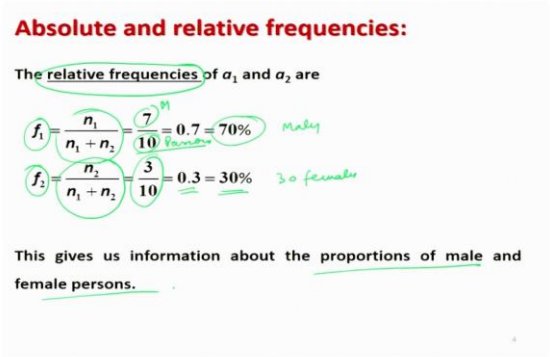
So, let me try to take here an example and try to explain what is this absolute frequency and what is relative frequency. Suppose there are 10 persons and they have been coded as male or female in two categories. So, male is indicated by M and the female is indicated by F. So, now the person has been selected and the person turns out to be male and then the second person turns out to be female, then the third person turns out to be male and so on and the data is recorded here like this M, F, M, F etc.

So now, you would like to know that how many people are in these two categories of male and female. So, let us try to count that how many are here male and how many are here female. So, let me try to count it here 1, 2, 3, 4, 5, 6, 7 so, there are 7 males and then 1, 2, 3 there are here 3 females, right. So now, these two categories say male and female they can be in general indicated by say here a 1 and a 2. For example, a 1 can be male category and a 2 can be a female category.

And the total number of people who are male and female in these two categories are say they are indicated by n 1 is equal to 7 and n 2 is equal to 3, right; that means the number of people in the category a 1 that is n 1 and n 2 is the number of people in the category a 2 which is here n 2 equal to 3, right. So, the number of observation in a particular category this is called as absolute frequency, right.

So, for example here you can see here this n 1 equal to 7 and n 2 equal to 3 they are the absolute frequency of the class a 1 and a 2 which are indicating the two categories male and female, right.

(Refer Slide Time: 05:23)



Similarly, if you try to consider that what is the proportion or percentage of these frequencies? So, that is called as relative frequencies. So for example, we have here two classes a 1 and a 2. So, the relative frequency of class a 1 is indicated by here

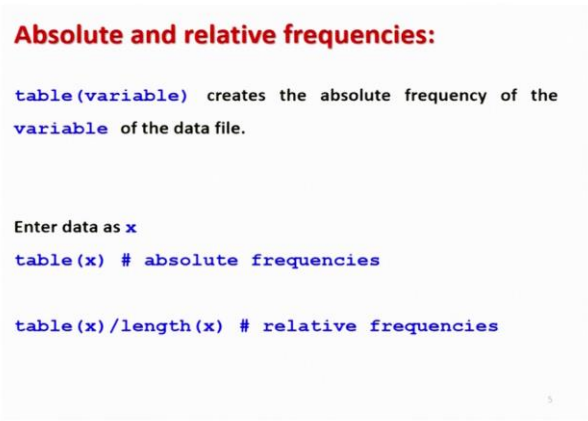
$$f_1 = \frac{n_1}{n_1 + n_2} = \frac{7}{10} = 0.7 = 70\%$$

that is equivalently 70 relative frequency of the second category

$$f_2 = \frac{n_2}{n_1 + n_2} = \frac{3}{10} = 0.3 = 30\%$$

which is equivalent to 30 percent. So, you can see here, I can say here that in these values there are 70 percentage males and there are 30 percentage females, right. So, this relative frequency gives us the information about the proportion of male and female in the data set.

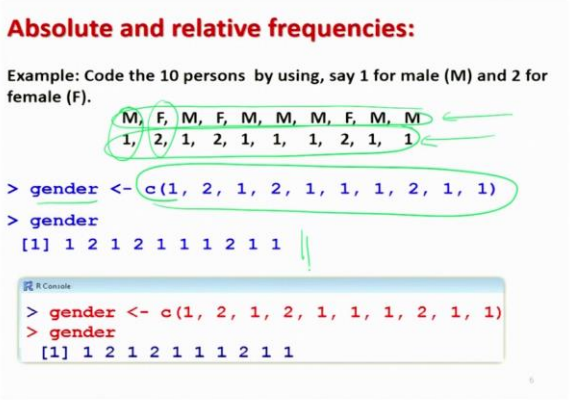
(Refer Slide Time: 06:18)



So, now the question here in which we are interested is that how can you find such absolute and relative frequencies in the R software, right. So, in order to do this first we need to create our table here and this table will give us the information about the absolute frequency of the data set which is stored inside the name variable. So, the command will be here table t a b l e and inside the parenthesis you have to give the name of the variable of which you want to find out the absolute frequency, right.

So obviously, if I if you try to store the data as x then, if you try to write table and inside the parentheses here x then this will give you the absolute frequencies and if you try to divide the table x by the total number of observation in x which can be obtained by the command l e n g t length of x. That means, how many observations are there in the x data vector, then it is going to give you some idea about the relative frequencies, right.

(Refer Slide Time: 07:24)

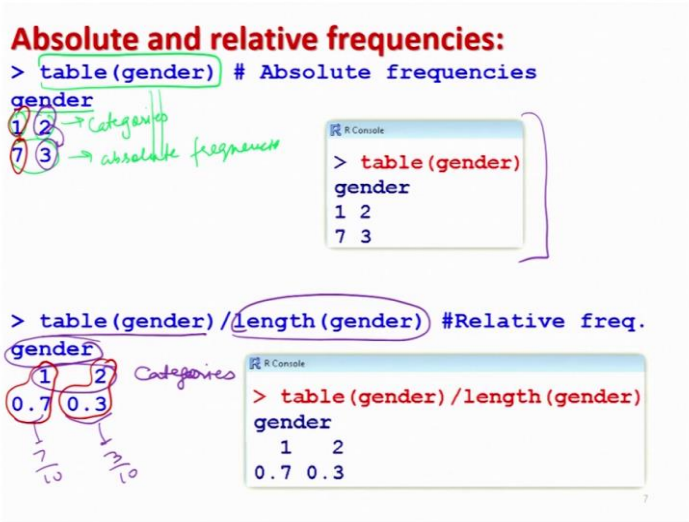


So, for example if you try to take here same example so, we have the data here M, F, M, F etc. So now, definitely when you are trying to consider these values M, F etc. then, they cannot be understood by the R software, but you need to convert them into some numerical value. So, that R can understand it and can do the required mathematical operation. So, what we try to do here that we try to indicate the categories male and female by the numbers 1 and 2. So, male is going to be indicated by 1 and female is going to be indicated by 2.

So, now the same data set which is in the format of male and female that is now indicated in terms of the numbers 1 and 2 and I try to store this data here in terms of 1

and 2 as and g e n d e r gender and this data is stored here in this data vector c, right. So, this is here the data, yeah.

(Refer Slide Time: 08:18)



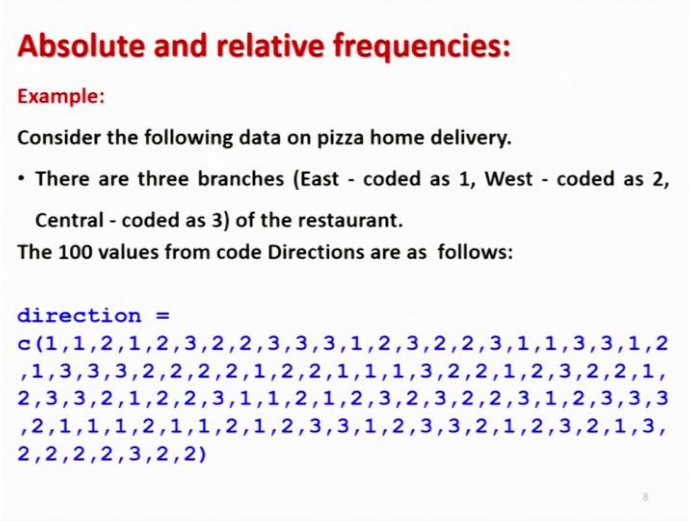
And if you do not want to do it manually you can also know that you can do it through factor also, but anyway that I will leave up to you my objective here is simply that how you want to create the table and find the absolute and relative frequencies.

So, after that if you try to simply say here table gender, then it will give you here this type of outcome. This means here, this is the name of the variable g e n d e r and these are here the 2 categories, right. So, this is indicating the categories 1 and 2. And now it is given here in this second row these are the values of absolute frequencies. So, it is trying to give here the formation that in this category 1 there are 7 people or the absolute frequencies here 7.

And this category 2 the total number of people are here 3, right. So, this will give you the absolute frequency, this is how it is indicated and here you can see the screen shot also. And then, if you want to find out the relative frequency here this is here the table divided by the length of the gender variable. So, that is the total number of observation in the data vector, right. So, now it will give you here the name of the variable, here as a gender 1 and 2 these are the categories and then you have here the values of relative frequencies.

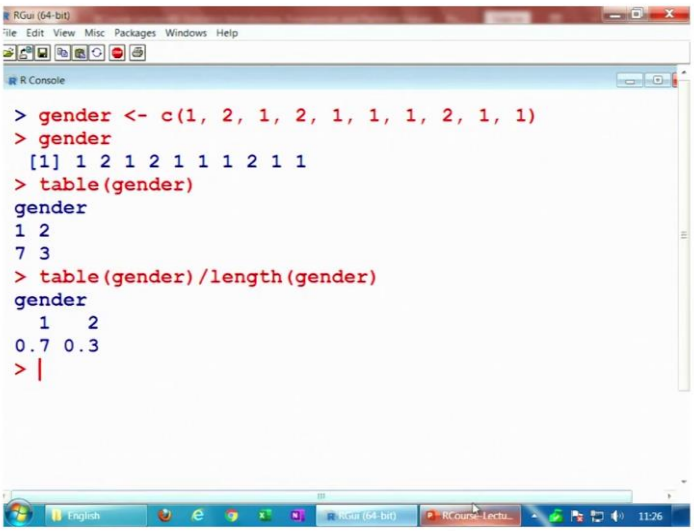
So, this is obtained as 7 upon 10 and this is obtained here as a 3 upon 10. So, it is trying to indicate here that in the category 1 there are 70 percentage people in the category 2 there are 30 percent people, right. So, that is how it actually gives you the values of relative frequencies, ok.

(Refer Slide Time: 10:05)



So, now let me try to show you these things on the R console so, that you get here a fair idea that how the things are working and after that I will try to show you some more examples here, yeah.

(Refer Slide Time: 10:15)



And you please try to keep these examples in mind because; we are going to use it in the further lectures also. So, gender data here is like this and if you try to create here a gender table of this gender variable. So, it will come out to be here like this and if you try to divide it by here the length of gender, it will give you here the relative frequency like this. So, you can see here this is very straightforward and simple to obtain this absolute frequencies and relative frequencies in the R software, right.

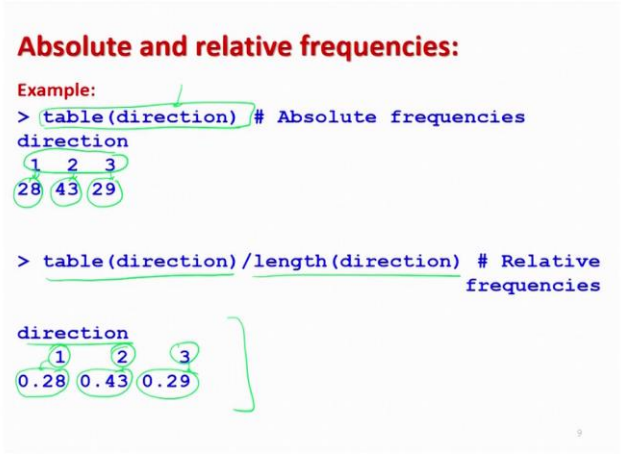
So, now I try to give you here one more example suppose there is a restaurant which deliver the pizza at home, right they have a home delivery system and suppose this restaurant has three branches one is in the east direction of the city, another is in the west direction of the city and third is in the central direction of the city; that means, in the central part of the city, right.

So, the branch which is allocated in the east part of the city this is coded as 1, the branch which is located in the west part of the city that is coded as 2 and the branch which is located in the central part of the city that is coded as 3, right. And suppose they get the orders every day and they try to deliver in different places, right.

So and then, it is like that they are getting the orders at a center place and depending on the location where the order has to be delivered they try to choose the appropriate branch from where the delivery can be made faster correction which are coded as 1, 2 and 3 this is obtained here like this.

So that means, it is simply what was the distribution of coded branches 1, 2 and 3 and we want to know that how many they have delivered from these branches. So, this data has been can see stored as 1, 2 and 3 like this in terms of 1, 2 and 3 and the data is stored in the variable d i r e c t i o n direction and this is stored using a data vector.

(Refer Slide Time: 12:11)



Now, in case if you try to see here what will be the distribution of these values? So, I can find out here the absolute and relative frequencies. So, the absolute frequencies are going to be obtained by t a b l e table and within the parenthesis; I down the variable name direction and then you can see here that there are 3 directions here 1, 2 and 3.

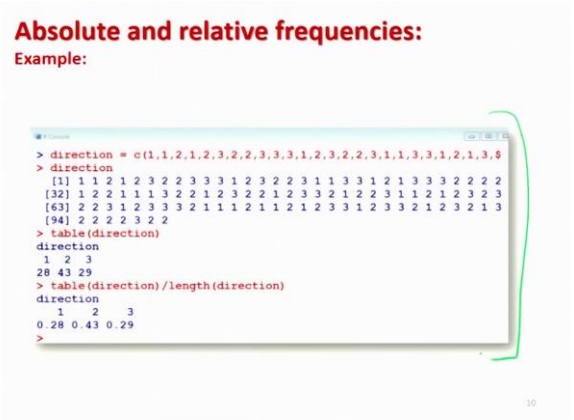
And 1 has made 28 deliveries and direction number 2 has made 43 deliveries and the rest in the direction number 3 that has delivered 29 such deliveries. And if you want to find out their relative frequencies here, so that will be these frequencies divided by the total number of observation that is here 100.

So, if you try to write down here table correction divided by length of the direction that will give you the relative frequencies. So, the outcome will come here like this direction and then these locations means east to west and central. So, you can see here the eastern branch has delivered 28 percent of the orders western branch has delivered 43 percent of the orders and the central branch has delivered 29 percent of the order.

So, this is the thing which you can say get directly from the R software. And the difference between this example and example is that in the first example you can see there are only 10 values. So, you will get confused that I can do this calculation by I can do the manual calculations also, but you can see here as long as you have your 100 values you simply just cannot count by your eyes, but you have to count it and you can think that if there are say big set of values say consisting of 1000 million values then,

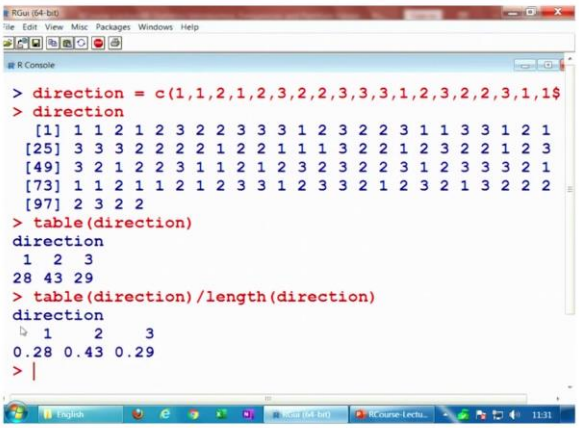
you cannot do such calculations manually. And in that cases this command here in the R software that helps you.

(Refer Slide Time: 13:46)



So, you can see here this is the screenshot of the same outcome. So, let me try to show you this operation on the R console.

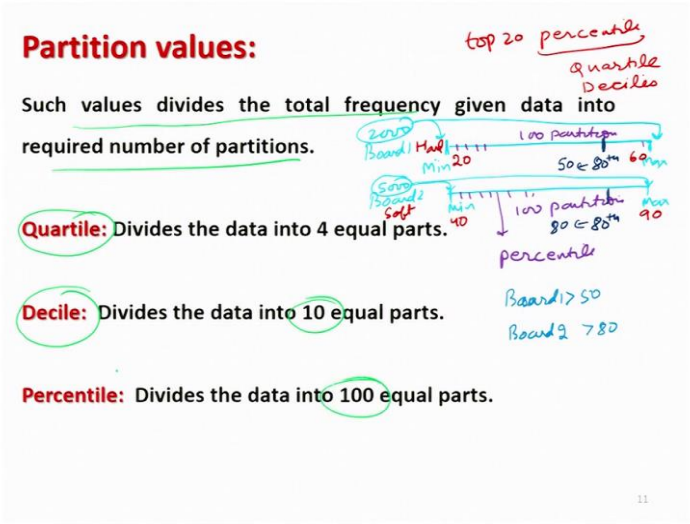
(Refer Slide Time: 13:59)



So, let me try to first create the data set here. So, you can see here this data set here is like this direction and if you try to find out the absolute frequency. So, you execute the command here table; you will see here this is coming out to be like this, right that there are 28 orders are delivered by the eastern branch, 43 orders have been delivered by the western branch and 29 orders have been delivered by the central branch of the pizza restaurant and if you try to find out their relative frequency.

So, you can divide each of these absolute frequency by the total number of observations you get here like this, right. So, that is the same outcome which I just explained you on the R console.

(Refer Slide Time: 14:35)



So, now after this let me try to give you here one more option, one more concept, right. Many times you have heard that in some examinations there is a condition that in order to qualify the examination you have to be in the for example, say top 20 percent quantile or say it is like say you have to be in the top 20 percentile, what does this mean? And then once you talk about the value like percentile then you have the values like here quartile and then deciles etc. So, what are these values?

So, you see look at me try to give you here a very simple example. For example, you have heard many times and yeah that is only a hypothetical example to explain you. Suppose you hear that there are two examination boards and they are trying hard marking.

So called quote unquote hard marking and the other board is doing a soft marking means, quote unquote soft marking. Yeah, because marking is always a marking and you cannot question it, but sometimes we feel that ok there is some board which is very strict in giving the marks if you make a mistake and there is another board which is not so strict in giving the marks if a candidate is trying to make a mistake.

So, this is what we mean by soft marking and say hard marking, right. So now, suppose both the boards give the marks out of 100. So, the minimum is 0 and maximum is here a 100. Now, suppose we call them as board 1 that is giving the hard marks and board 2 that is giving the soft marks.

Now, suppose large number of candidates appear in the examinations and the examination board 1 and examination board 2. Now, when their marks are coming to us and suppose I am a school means, I have a college and I want to take admissions of those students who have appeared in the examinations under board 1 and board 2. Now, my problem is that the student who has got suppose say 40 out of 100 from the board which is giving the hard marking and there is another student who has got say 70 percent from the board which is giving the loose marking how to compare.

I am sure that you will agree with me that these marks are not really comparable you cannot compare the 40 percent marks of a board which is doing the tough marking and 70 percent marks of a board which is the which is giving the loose marking or say soft marking so, how to compare it. So, in order to compare the types of marks we have one option here that what can do; that can simply create here a scale and where I try to give here the minimum and maximum marks, right.

So, this is my here board 1 and then I have here board 2 and I also try to do the same thing here minimum and here maximum, right. And now, yeah this minimum and maximum for the board 1 and board 2 can be different for example, minimum and maximum are indicating the mark of a student which is the minimum among all the students in the board 1 and similarly in the board 2.

So for example, if there are suppose 2000 students appear in board 1 and there are 5000 students appear in board 2. So, whosoever marks is minimum out of 2000 this is here and whosoever marks is maximum out of this 2000 that is here. And similarly, in the board 2 out of this 5000 student whosoever has got the minimum marks, that minimum mark is here and whosoever has got the maximum mark that mark is given here as maximum, right.

So, suppose if I say that board 1 gives in the hard marking and board 2 gives the soft marking. So, it is possible that in the board number 1, the minimum marks are suppose

20 and the maximum marks are suppose 60. And in the board 2, because it is giving the soft marking so it is possible that the minimum marks are here suppose here 40 and the maximum marks are here 90.

So now, in case if a student has got a particular mark it becomes now very difficult to get the correct information about the capability of the student. So what we can do here, that we try to divide the difference between this minimum and maximum marks in to equal parts. So now, this equal part can be 4 that can be 10 that can be 100 or that can be anything whatever you want. So, suppose for example, if I say I try to make here 100 partitions 100 partitions so, 1234 to here 100.

So, similarly here in this case also I try to make here 100 partitions, right. So, these 100 partitions each of the partition is called as quantile and when you are trying to make 100 partition and every partition is called as percentile. So, that is why many examinations they have made a condition that they are going to consider those students who are in the top 20 percentile of their board examinations.

So, what does this mean, that they will always try to look whatever is the top 20 from the maximum; that means the 88th percentile? So, 88th percentile of board 1and 88th percentile of board 2 and they will yeah 88th percentile of board 1 and the 88th percentile of board 2 they are usually expected to have different marks.

So, it is possible that in the board 1 at 88th percentile there can be a student who has got only say 50 marks and in the board number 2 there can be a student at the 88th percentile who has got suppose say 80 marks. So now, they will say they will make a very simple rule that all those students who have got the marks which are more than the 88th percentile of the respective board they are eligible to apply for admission. That means, if a student has appeared in board 1 and if a student has got the marks between say 50 or say more than 50.

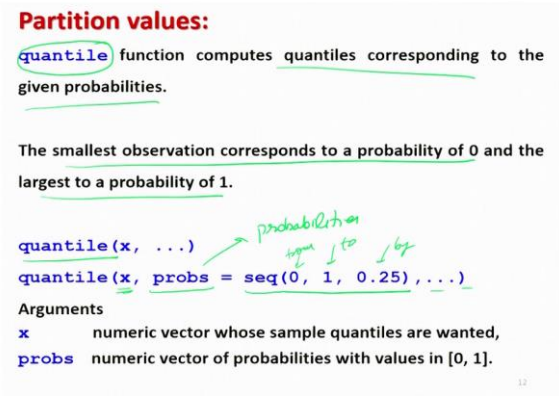
Then, the person that student is eligible and if a student from board 2 has got the marks more than suppose here 80, then the student is eligible, right. So, now, this is how you can possibly compare. This is one possible way which is the application of this quantile, yeah. So, I have taken this example here just to explain you that what are the partition values. So, similarly if you want to make here only 4 partitions instead of 100 then these

partitions are called as quartile, you call it first quartile, second quartile that is median third quartile etc.

And similarly, you can you have an option that you can make as many as partitions you want depending on your wish when you are trying to work in the R software. So, this is the topic which I am going to now explain you. I already have explained you the concept and now my job is to show you basically that how to execute these things in the R software, right.

So, in this partition value the values are divided the total frequency of the is divided into the required number of partition. So, you have the data, you create the frequency table and then you try to divide them into the required partition. So, if you are trying to divide the total frequency into 4 equal parts, then it is called as quartile and if you are trying to divide it into 10 each of the partition is called as decile and if you are trying to divide the total frequency into 100 equal part then you try to then you call it as percentile, alright.

(Refer Slide Time: 22:38)



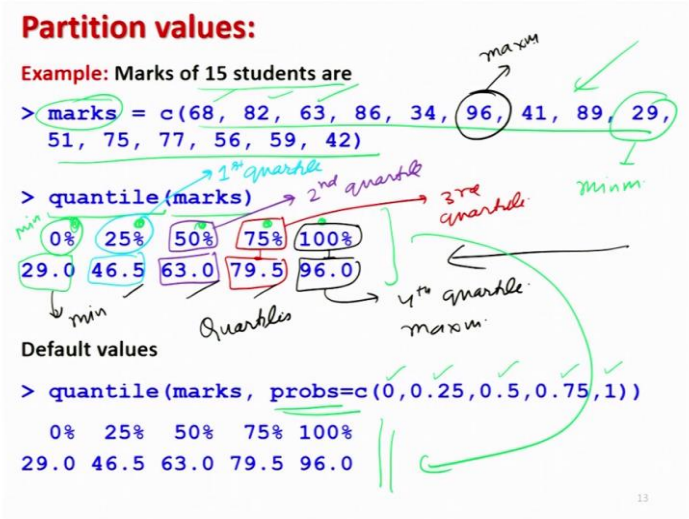
So, now the question here is, how to obtain such partitioning values in the R software? So, in the R software we call them in general as a quantile q u a n t i l e and the function which computes such partition this is the quantile function q u a n t i l e; and this quantile function computes the quantiles corresponding to the given probabilities, right. Because, you can define the partition in terms of the values between 0 and 1 so, probability also lies between 0 and 1.

So, that is why the values of the partition; that means, at what value we want to partition it that is indicated by a probability value. So, how do you get it? I will try to show you in the example through the example, right. So, what happened that? The smallest observation corresponding to probability of 0 and the largest observation is corresponding to the probability of one, right. So, that is how this partition is done.

So, now we have a function here quantile q u a n t i l e and then, inside the parenthesis we try to write down here the data vector x. So, if you do not write any choice then the default here is quartile that it will try to divide the total frequency into 4 equal parts, but if you want to do it according to your wish then you have here an option probs p r o b s which is the short form of the probabilities, right. So, this is the short form of the probabilities. So, probabilities are between 0 and 1.

So for example, now you can choose it in the way you want you can use the data vector and if you want to give the sequence this is up to you. So for example, if you try to write down here sequence 0, 1, 0.25, then you understand what is the meaning of this is here to and this is here by, right and after that there are many options. I would request that you please try to go to the help of this quantile and try to understand other options, right.

(Refer Slide Time: 24:31)



But anyway I will try to take here some examples. So, that I can show you that how these things look like and how you can interpret them. So, for example, I take here a very

simple example that we have got the marks of 15 students out of 100 and these marks are stored in this data vector here mark they are 68 82, 63 and so on. So now, I want to create the quantiles of these marks. So, I simply try to give here the command on the R console as quantile q u a n t i l e all in.

Lower case alphabets and within the parenthesis I write down here the data vector mark and you get here this type of outcome. So, now you have to understand what is this trying to say here. You can see here there are 5 values this is here 1, this is here 2, this is here 3 50, 75 and here 4. So, what you have to understand here that the first value here is the minimum value that is at 0 percent and this value here is 29. And if you try to see in this database here this is value here is 29 which is the minimum value.

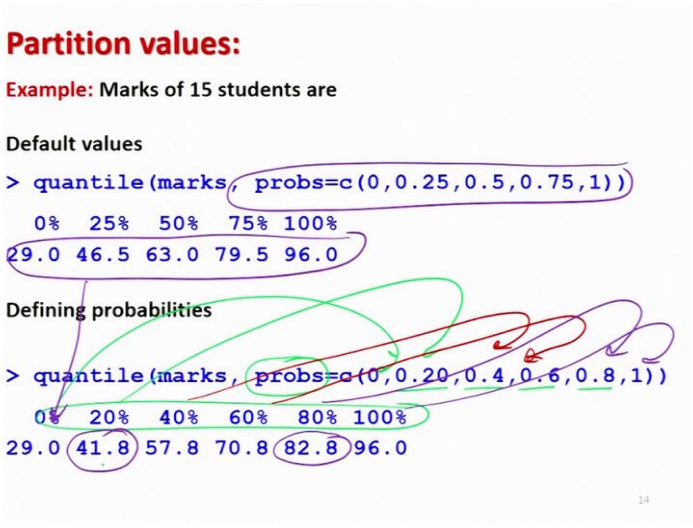
So, that is coming here, right. After this there is here an option here 25 percent. So, that is your here 1st quartile, right. So, it is value here is 46.5. And similarly, you have here the next value here is 50 percent, this is the value of your 2nd quartile and this value here is 63. And similarly, if you try to see here the next value here is 75 percent and this value here is 79.5.

So, this is the value of the 3rd quartile. And after that, you have the last value which is here 100 percent, which is here the 4th quartile whose value here is 96 and this is the 4th quartile, but if you try to see here that data what you have considered here is the 96 and this is also the maximum value. So, that is what I said that this is also here the maximum value. So, the value and this 0 percent is the minimum value.

So, the data between 0 percent and the minimum and maximum that has been divided into 4 equal parts and these partitionings are happening at 46.5, 63, 79.5, right. So, this is how we read this outcome and we try to understand the values of these quantiles which is actually giving you here the quartiles. Now, similarly if you try to see that when I am not using the command here probs, then it is giving you this type of outcome.

But, in case if you try to use here the command here probs and if you try to say that the values are at 0, 0.25, 0.5, 0.75 and 1 that means the values are increasing by 0.25 then, you get here the same outcome which you have obtained here. So, you can see here the default outcome of the quantile function is the quartile, right.

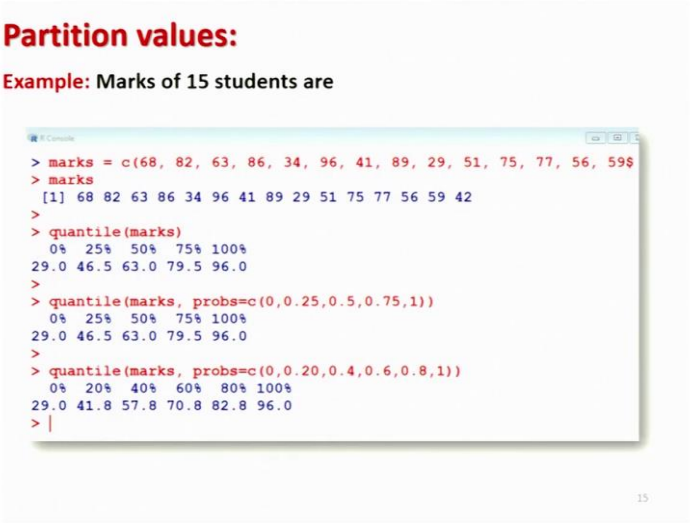
(Refer Slide Time: 27:29)



And similarly, if you want to make here some other value. For example, you want to have the quantiles at say 28th at 20 percent, 40 percent, 60 percent, 80 percent and so on then you can just give it here by probs p r o b s. Probs and then you can try to give these value inside the data vector and then you can see here you are getting here these values here, right. So, this 0 is corresponding to this 0, this 20 percent is corresponding to this 20 percent, 40 percent is corresponding to this 0.4, 60 percent is corresponding to this 0.6.

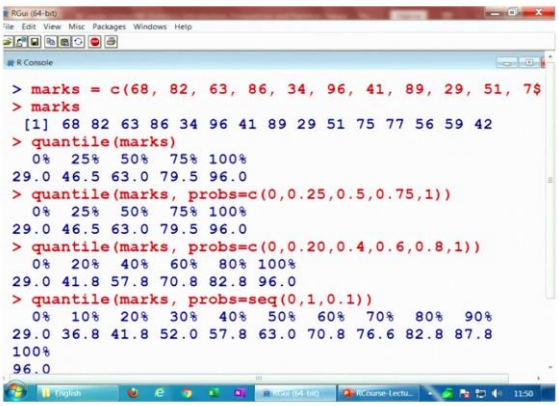
And after this 80 percentage is corresponding to 0.8 and 100 percent is corresponding to say here 1 and these are the values. So the value at that the 20 percent partition is 41.8 and the value at the 80 percent partition is 82.8. So, this is the control the probs. And if you try to take the probs is equal to 0, 0.25, 0.5, 0.75 and 1 you can compare these are your here basically the quartiles and you can see here that these values are quite different than the values which are obtained here, right.

(Refer Slide Time: 28:42)



So, this is how this actually quantile function works here and this is here the screenshots. So, you can see here these are your here quartiles and these are here the quantiles which are partitioned at 20 percent, 40 percent, 60 percent, 80 percent and 100 percent these values are quite different from each other, right. So, let me try to show you these operations on the R console.

(Refer Slide Time: 29:13)

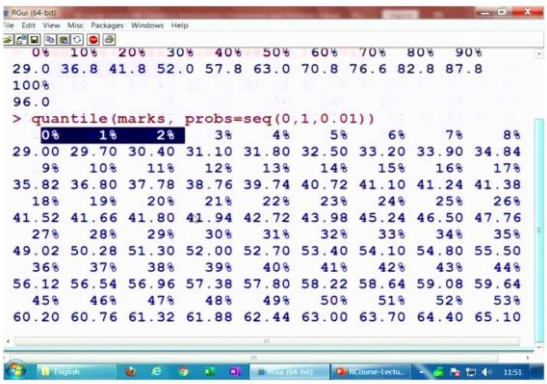


So, that you can understand them very easily and the reason I took it here, because this quantile is very useful in real life. So, you need it many many times and another values table etcetera that are you will see that we will use when we are trying to create the graphics, right. So, this is you are here the data vector marks and if you try to find out

here the quantile of here marks. So, it will give you here this quartiles and if you try to use the probs function here at 25 percent partitions you get here the same value you can see, right.

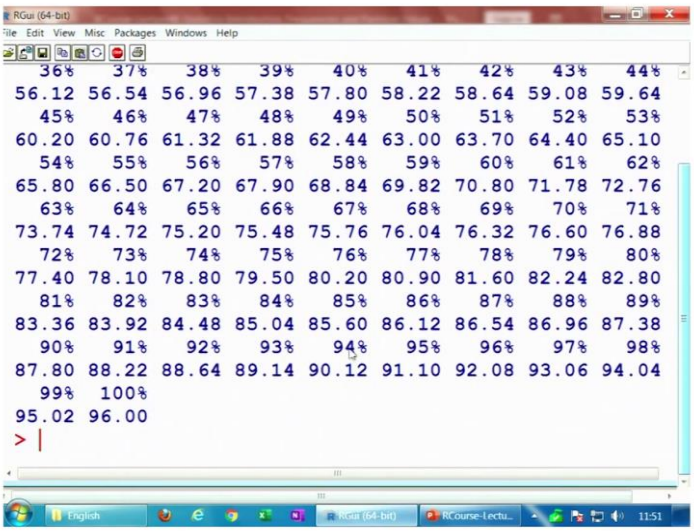
And if you try to use here some other partition here, right possibly you can see here this is the partition in partitioning is at 20 percent. Similarly, if you want to obtain the deciles then, I can give this partitioning as a sequence between 0 to means from 0 to 1 with the increment of 0.1, right. You can see here this comes here like this. So now, you have here 10 partitions the 10 percent, 20 percent 30 percent etc.

(Refer Slide Time: 30:09)



And similarly, if you want to know here what are the values of percentile. Percentile means here very simple you want to make 100 partitions so, you give the sequence starting from 0 to 1 and changes are occurring at a 0.01.

(Refer Slide Time: 30:27)



So, this you can see here these are 100 values starting from 0 1 percent 2 percent and finally, you have a 99 percent and 100 percent value, right. So now, I stop in this lecture and you can see that was a pretty simple and straight forward lecture and my idea was not really to teach you statistics, but my idea was that I want you to consider here some topic which I am going to use in the further lectures.

And I because when I am trying to create the graphics, the graphics are created from some data and in order to understand the data as I said that you have two options; one is the analytical that you try to find out these values frequency table etcetera, but the frequencies table etcetera they can also be indicated by some graphics.

So, that was the reason I took these topics in this lecture. So, why do not you try to take some data set and try to create such frequency tables and remember one thing I have considered here only a variable which is categorical variable for creating the table. And for this quantile you can you have to take any continuous data means the value should not be like that which are taking means some categorical values etc.

The value should be like a their height, weight etc. which can take any real number, right means something like height. Height can be 1.57 meter also, height can be 1.5 meter also, height can be 1.2 meter also, height can be I mean 1meter also and so on, right.

So, try to consider such continuous type of data and then you try to create such a quantile that will make sense, right. So, why do not you try to practice these things and try to understand what they are trying to indicate because, that is going to be useful when I am going to consider the topics in the next lecture. So, you try to revise and I will see you in the next lecture, till then goodbye.