Introduction to R Software

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

Lecture - 35
Introduction, Frequencies and Partition Values

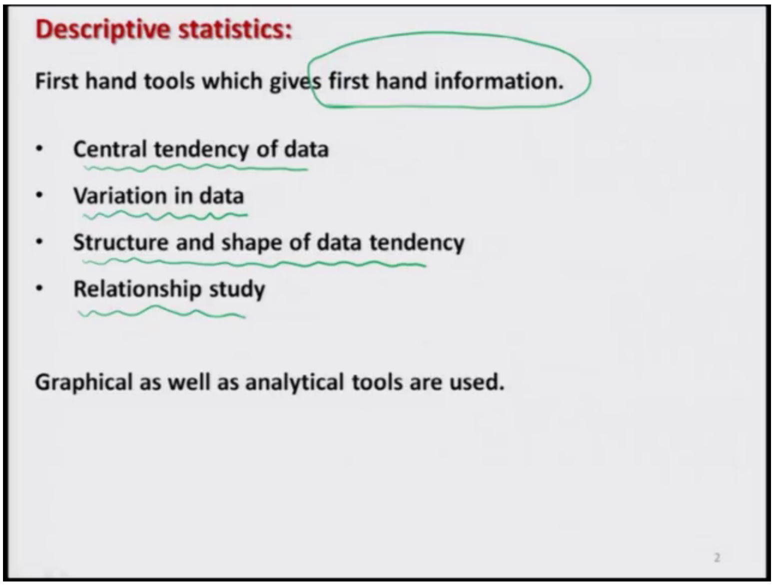
Welcome to the next lecture on the course introduction to R software and we are here with the next lecture and a new topic. You may recall that up to now we have done the basic functions which are mainly related to the programming, different option, different syntax, different commands and various examples. Now my objective is to show you that there are some built in functions in R and they can be used to know something without writing a new program and these functions helps us in computations. In order to illustrate this aspect we are going to consider some statistical functions while do not worry for statistic I am not going to use any hi-fi statistics.

But I am simply going to use very elementary functions which all of you know for example, how to compute mean, how to compute variants and so on. And with these things I would also like to give you some idea that how to draw various types of graphics. So, what we are going to do here that we are going to consider the statistics which can be used with one variable and statistics that can be used with 2 variables and when I am doing this statistical function, there is another aspect graphics, how to draw different types of graphics in R.

So, when I am taking an example my objective is to use that example to expose you with the different statistical functions available in R and to introduce you with the graphics - what type of graphics are there, how to create those graphics well definitely I cannot take care all the statistical functions and all the graphics, but I would try to take some good number of functions and good number of graphics so that you get confidents and you can create any type of graphics after that yourself.

So, we start here and then this lecture we are essentially going to understand about the say these absolute frequencies, relative frequencies and some partitioning value. But before that what you really we try to do in statistics usually we have a set of data that is collected from anywhere from some external sources or somebody has collected it himself or say herself or they can be various sources from which the data can be obtained the role of statistics is to provide different types of informations which are hidden inside the data and statistics is a language to explain the features or to extract the features which are hidden inside the data.

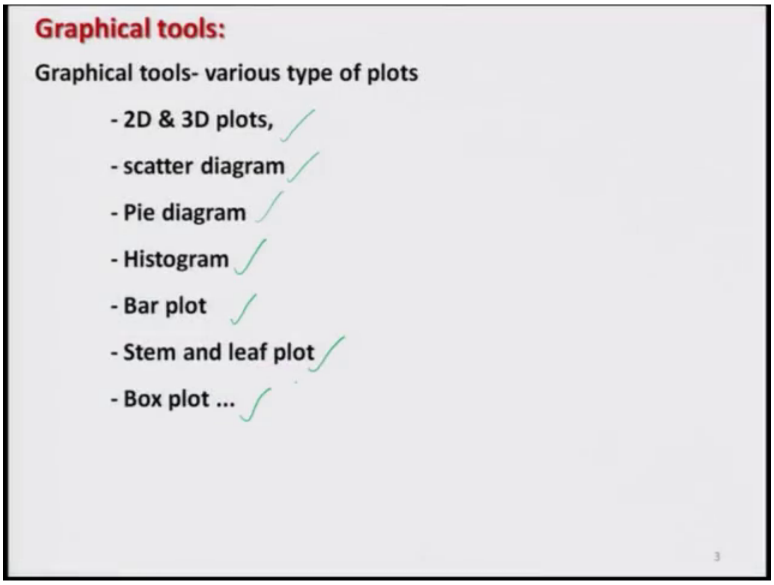
(Refer Slide Time: 03:47)



So, whenever we are talking of the features of data there can be various aspects. Those different features can be like for example, to know what is the central tendency of the data; that means, where the data is mostly concentrated then what is the variation in data; that means, how disperse is the data usually from the central value then we would like to study also the structure and shape of the data and what is its tendency then in case if we have say more than one variable, then we would also like to study what is their relationship and usually with these features we get the first hand information about the data and for that we have different types of statistical tools which are used and whenever we are trying to get the first hand information.

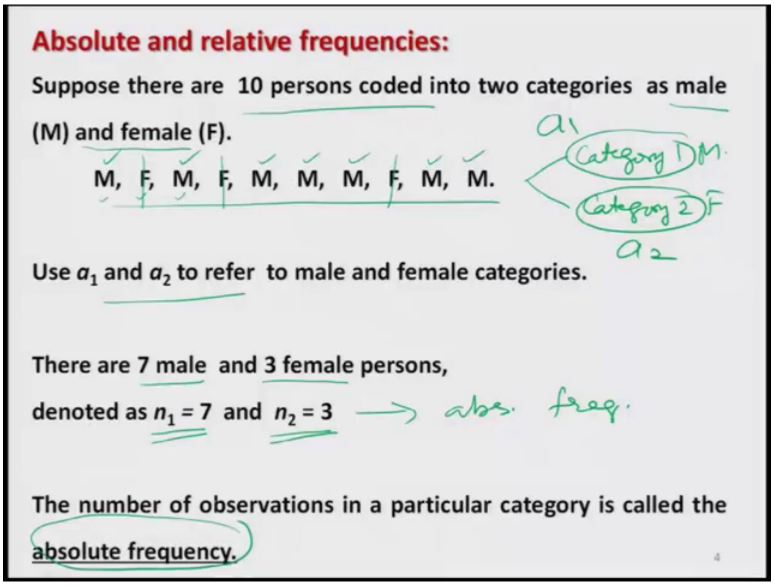
We used the analytical tools as well as the graphical tools and in order to take the final conclusion we try to combine the information gathered from analytical tool as well as the graphical tool to reach to a fruitful and good conclusion. So, in R beside the analytical function there are various types of graphics which can be plotted.

(Refer Slide Time: 05:11)



For example 2 dimensional plots 3 dimensional plots scatter diagram pie diagram histogram bar plot stem and leaf plot box plot and so on and there is a long list. So, we are going to understand how to create all this graphics here in this course also.

(Refer Slide Time: 05:25)

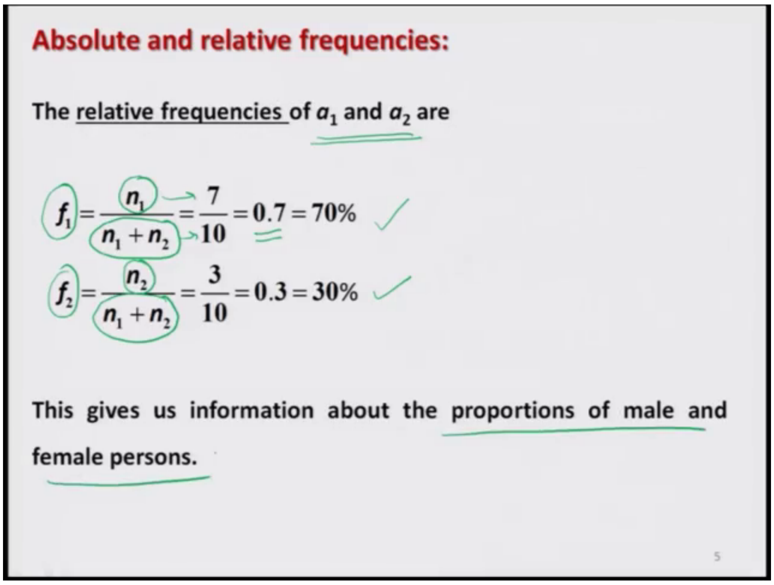


Now, we start first with the some analytical function and we start with some basics which you already have done, but possibly you have forgotten. So, I would try to give you say an example and from there I would try to introduce you with the different analytical function.

So, first topic I am going to consider is about the absolute and relative frequencies. What is the frequency? Frequency is occurrence. So, we would like to see whether a value is occurring in a data or not and in case if it is occurring then how many times. So, for example, if I try to take an example here suppose I collect the gender of 10 persons and I code them as say male and female. So, I denote capital M for male and capital F by female and suppose I get this type of data the first person is male second person is female third person is male and so on. So, now, I can divide this entire data into 2 category there can be category 1 and there can be category 2, category 1 can be say male and category 2 can be female. So, what we try to count here and we try to denote this category one by say here a 1 and category 2 by here a 2, right.

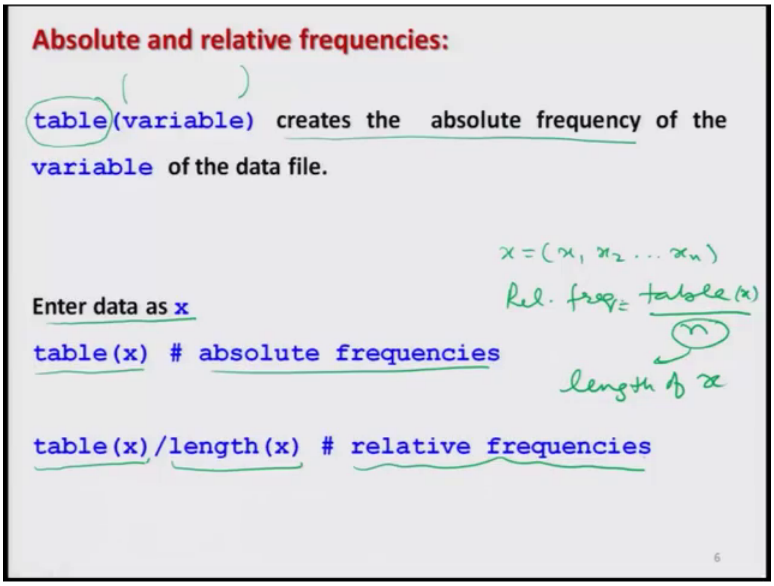
So, now we try to count here 1 2 3 4 5 6 7. So, there are 7 males and there are say 1 2 and 3, 3 females. So, I can denote these numbers the number of males and females by here n 1 and n 2. So, n 1 become seven and n 2 becomes here 3 right. So, these numbers are actually the absolute frequencies what is written over here they are called the absolute frequency. So, they are simply trying to tell as that how many male and how many female persons are available in that data set and similarly I can also define the relative frequencies.

(Refer Slide Time: 07:48)



So, I have here 2 categories a 1 and a 2 and in order to find out the relative frequency of the first category denoted here say f 1 I have to simply find out the proportion of the male in the entire data. So, this is n 1, n 1 is the number of male persons and n 1 plus n 2 is the total number of persons in my data. So, n 1 here is 7 and n 1 plus n 2 here is 10. So, this comes out to be here 0.7 or that is equivalently seventy percent and similarly the relative frequency of the second category is f 2 which is here say number of persons in the female category divided by the total number of persons and this is 3 upon 10 which is 0.3 and that is equivalently 30 percent. So, these f 1 and f 2 they are called as relative frequency and they give us an idea about the proportion of male and female persons in the data.

(Refer Slide Time: 08:55)



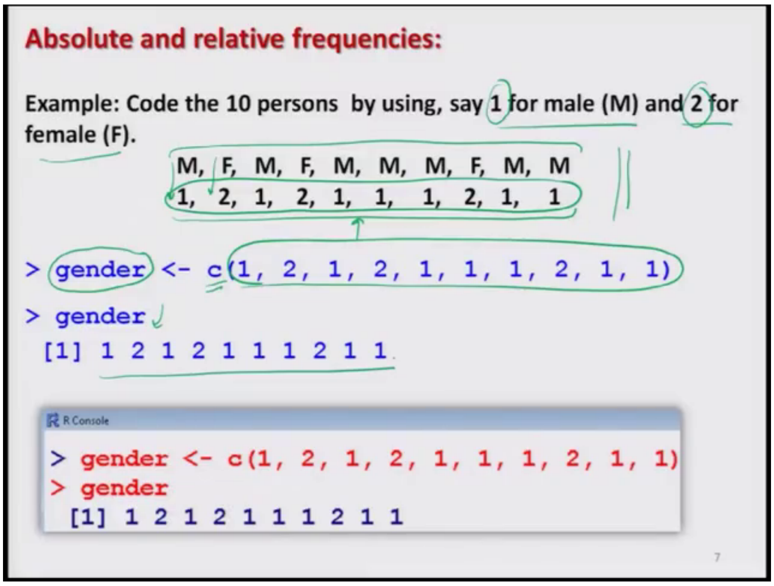
So, that is the first step whenever we are trying to say this start any statistical analysis whenever we get a data first I would like to explore different possibilities. So, possibly this is the first step that we try to do.

So, in order to obtain this absolute frequency in R what we need to do that is the question which we are going to entertain first. So, our next question is how to compute this absolute frequency in R and how to compute the relative frequency in R. In order to calculate the absolute frequency there is a function table in R software and this function creates the absolute frequency of the variable on which we have obtain the data variable is something on which we try to obtain the data for example, if I say my variable is height; that means, I am going to measure the height of persons here in this case my variable is gender. So, I am going to collect the information on the gender of the person right. So, the syntax here is that you try to write down table then inside the argument you have to write down the data or the variable name whatever you want.

So, suppose I say that I have entered the data in a vector say x then in order to find out the absolute frequency the command becomes table and inside the argument x. Now, if you try to see what is the relative frequency relative frequency is simply the absolute frequency divided by the total number of data available in x vector. So, if I say x is my data vector where I have got say small n number of values denoted as x 1 x 2 x n. So, if I try to see this relative frequency is simply here say table of x divided by n and n is simply the length of x.

So, in order to find out the relative frequency what we do we try to find out the absolute frequency by table of x and then we try to find out the length of x and then we take the ratio. So, this will give us relative frequencies of the given set of data. So, we try to use it over this given data set and we try to see what do we get.

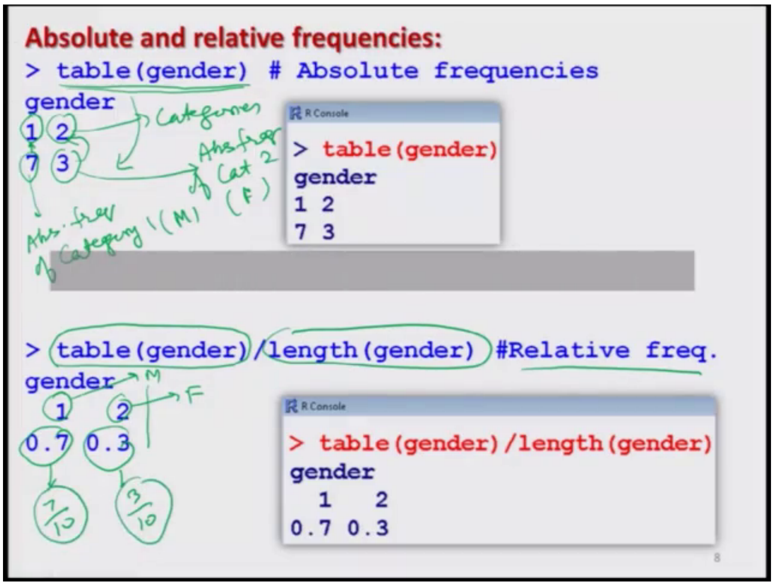
(Refer Slide Time: 11:39)



Now, you see when you want to make any mathematical computations on a functions you cannot do with the category say which is express in terms of male and female because these are alphabets right. So, I need to convert it into a indicator variables. So, what I try to say here that for the male for the letter M I try to use an indicator 1 and for female for the letter F I try to use an indicator 2. So, whatever data I have here in terms of male and female in terms of letters M and capital F I can convert it into in terms of 1 and 2.

So, M is denoted by 1 and F is denoted by here 2 and that you can see I have written here right. After this I try to collect this data set this numerical data set into a vector and I try to create here this vector this is the same data set and I try to combine them by here see and try to create here a variable say gender. So, now, this gender variable consist of these many values and you please keep this conversion of data into your mind because we are going to use this example repeatedly later on also.

(Refer Slide Time: 13:17)

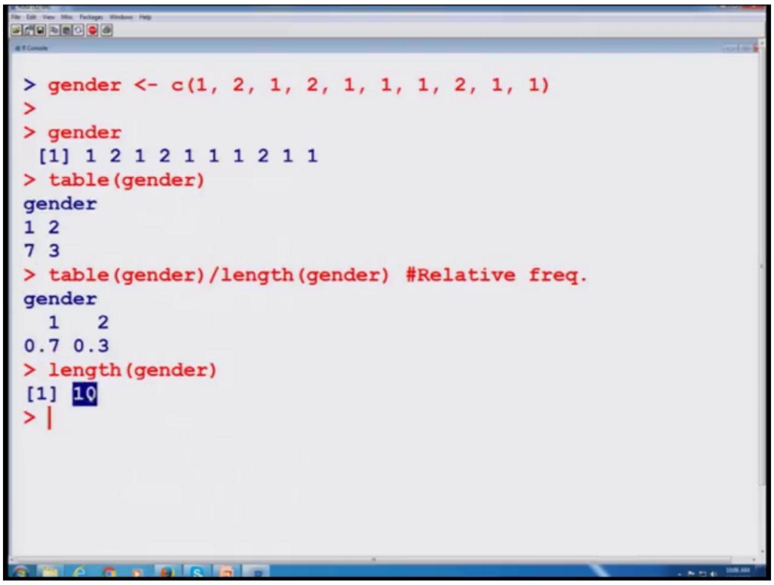


So, now what I do here that I try to find out here table of gender and you can see here I get this type of outcome and this and the meaning of this outcome is this these 1 and 2 are the 2 categories.

And this here 7 this is going to give the absolute frequency of category 1 which is here male and this 3 is going to give the absolute frequency of category 2. So, 7 is absolute frequency of category 1 that is here male and 3 is going to give us the absolute frequency of category 2 that is here female. And similarly if I want to find out the relative frequency then I simply have to use here table gender divided by length of gender and you will get this type of here outcome we are again this 1 and 2 these are denoting the male and female categories and this 0.7 is nothing 7 divided by 10 and 0.3 is nothing 3 divided by 10.

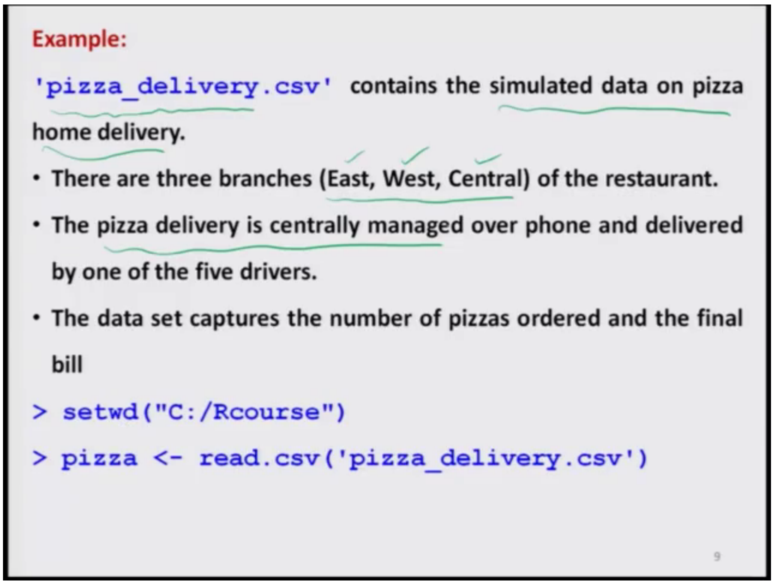
So, because there are 10 observations there are 7 male and there are 3 females. So, this number comes out to be 7 upon 10 and this is comes out to be 3 by 10 which is reported here. But let us try to do this function over the R console and let us try to see what do we obtain. So, first I try to create here a vector gender and then you can see here.

(Refer Slide Time: 15:09)



So, you can see here this comes out to be here gender and now I try to create here table, table of gender this gives you here this thing. And similarly if you try to find out the relative frequency you have to simply divide the table gender by the length of gender and you can see here what is length of gender this is here ten. So, this 0.7 is actually this highlighted 7 divided by 10 right. So, now, we come back to our slides. And here you can see that these are the screenshots of the same outcome.

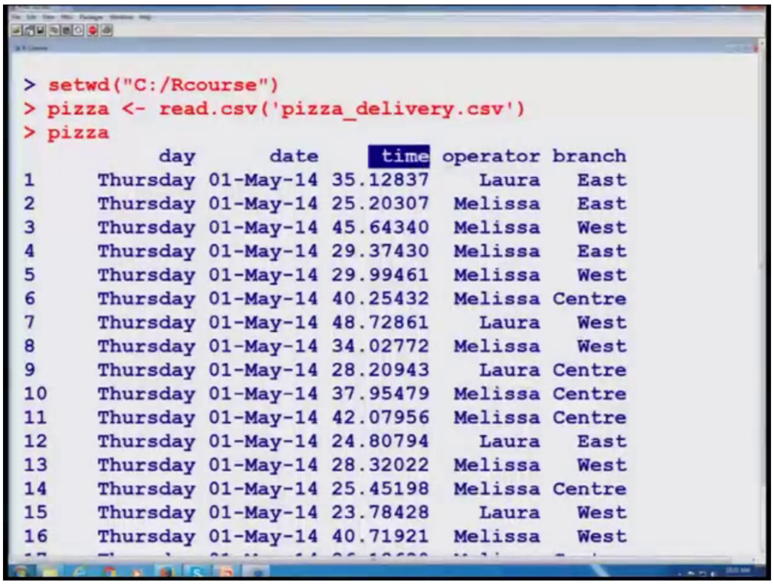
(Refer Slide Time: 16:09)



Now, before going further to consider more example first let me explain you the following. We are going to consider here a data set which is named as pizza underscore delivery dot csv and this contains a sort of simulated data set on the pizza delivery at home and in this data set there are 3 branches of the shop in the eastern part of the city, western part of the city and center part of the city. And this pizza delivery is centrally managed over phone, for example, if somebody calls the shop at some central location and order the pizza over phone and the pizza is delivered at their place and for that there are 5 drivers.

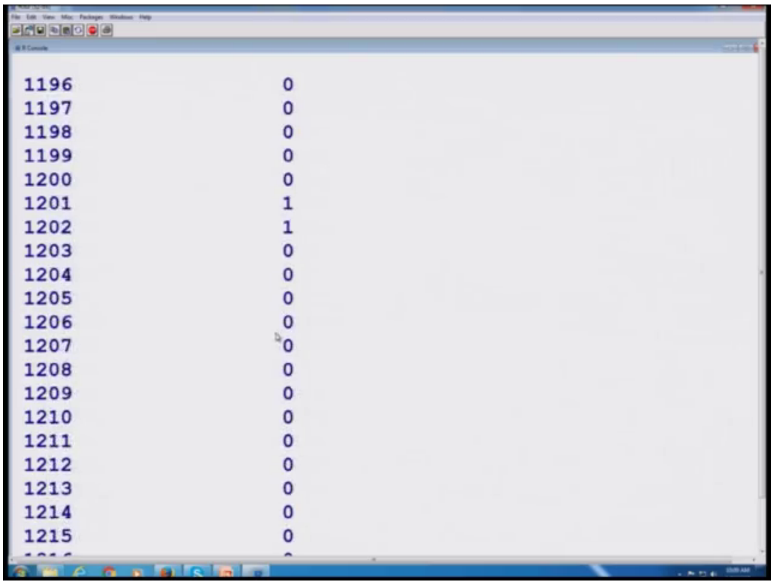
So, the data is collected on the number of pizzas ordered and their final bill means what type of pizza and how many pizza and there are different types of thing and this data is saved in a file pizza underscore delivery dot comma separated value that is csv file. The question comes why should I use this data if you try to see in the earlier example I have simply taken 10 values and I am trying to show you how the things are happening, but in this pizza delivery this is are large data set I will try to show you. And, so, I would also like to show you that what really happens when we are going to deal with the larger data set because in a larger data set you cannot see all the values yourself. So, why not to first load this data set before we go further.

(Refer Slide Time: 17:49)



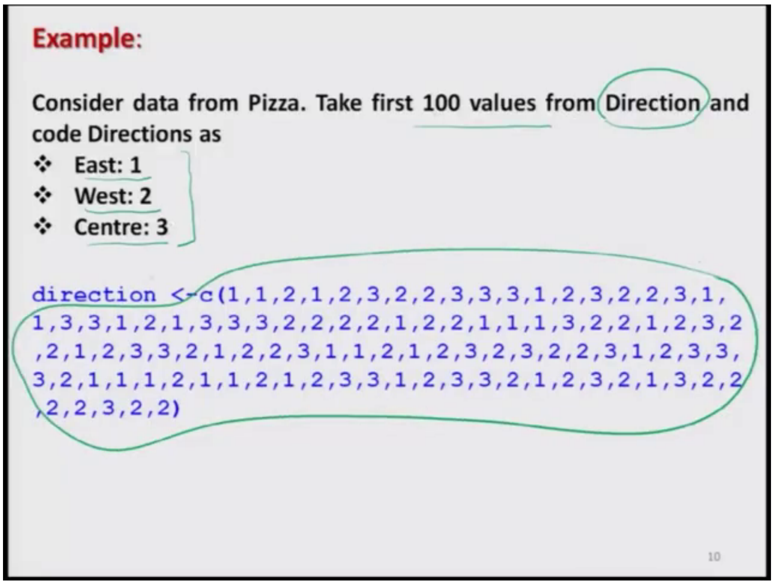
So, you remember the first thing in order to load a data set is that we have to copy it in a specified directory and we have to set the working directory that we discuss earlier in the lectures and by this command say w d I can locate my file. And then whatever is my file here this I am going to denote by a name p i double z a pizza. So, now, I would like to show you what is contain in a pizza. Please try to look into this screen because the screen is going to scroll very fast. So, you can see here.

(Refer Slide Time: 18:22)



So, this is a huge data set which is here right. So, it is giving me an idea of you see all detail when the data was collected and everything and you can observe here that there are different types of variable here day, date, time operator and here branch and similarly here there is another variable which is here direction right.

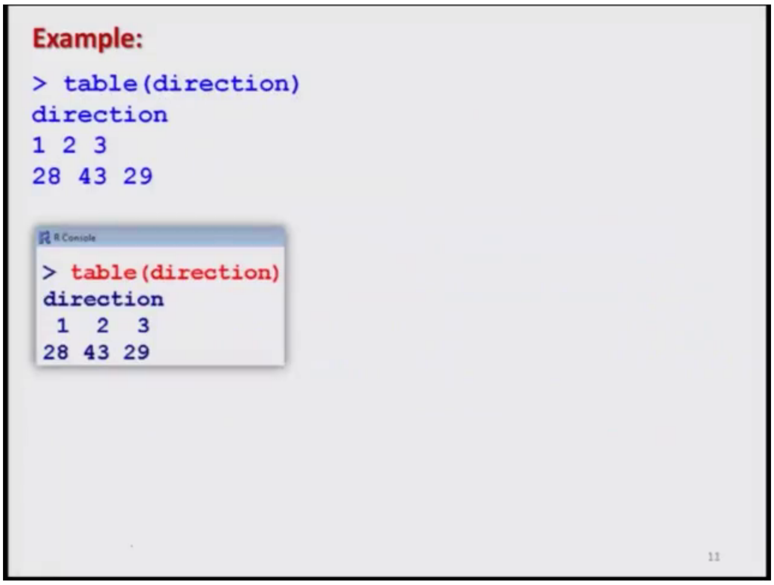
(Refer Slide Time: 18:59)



So, now let us come back to our slides and we try to continue. So, I take this data set and I am simply trying to take first 100 values means from 10 to I have gone to 100 there is a variable direction and this directions are east west center. So, I try to code east direction by number 1 west by number 2 and center by number 3. So, you can see here that this data which is collected that is in terms of east west and center, but definitely you need to give it a category in terms of 1 2 and 3 and if you remember we already have done this exercise that we try to take a variable and then we try to give it a category by a number.

So, I would request you please try to have a look in the earlier lecture and try to use that concept and try to convert this data east west and center in to the categories 1 2 and 3. So, what we have done here that we have taken here 100, first 100 values from this variable and we have stored it in the say variable direction right.

(Refer Slide Time: 20:07)

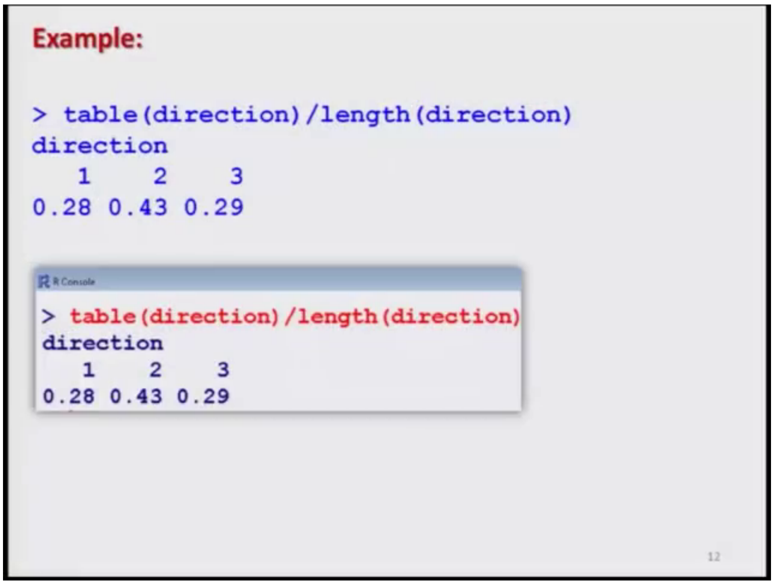


So, now we try to find out the absolute and relative frequencies of this variable direction and for that we try to use here the function table and with the variable here direction and as soon as you enter you get this type of outcome. So, you can see here this data is saying here 1 2 and here 3 and 1 is indicating that there are 28 numbers, for 2 there are 43 numbers and for 3 there are 29 numbers right.

So, this is how you can know that what are the number of east west and these directions in your data set without looking into the data. Well I have taken here 10, so that if you want to be more confident you can compute yourself, but definitely you can extract one variable from this data set and you can find out the absolute frequency directly for the variable because you already have now done that how to extract a particular variable from a data set using the dollar sign or attach function.

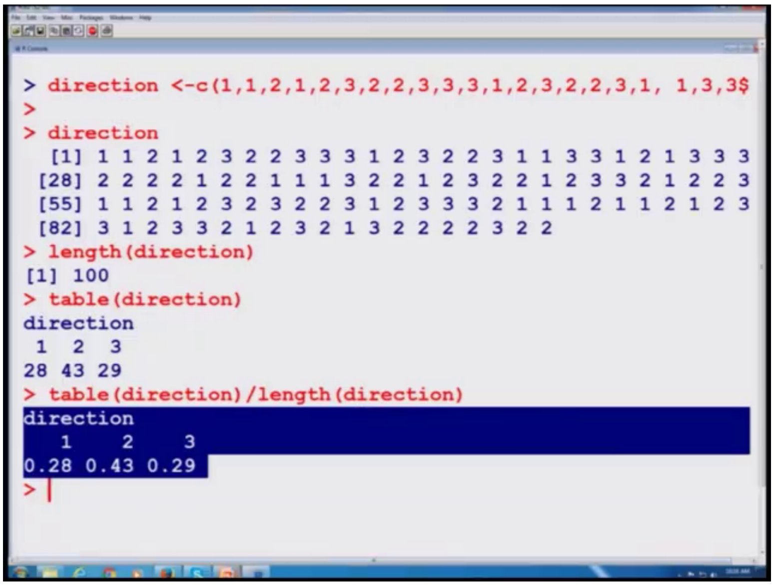
And similarly I try to find out here the relative frequencies. So, for that I try to take here table direction divided by the length of the vector direction and I get here this type of outcome.

(Refer Slide Time: 21:14)



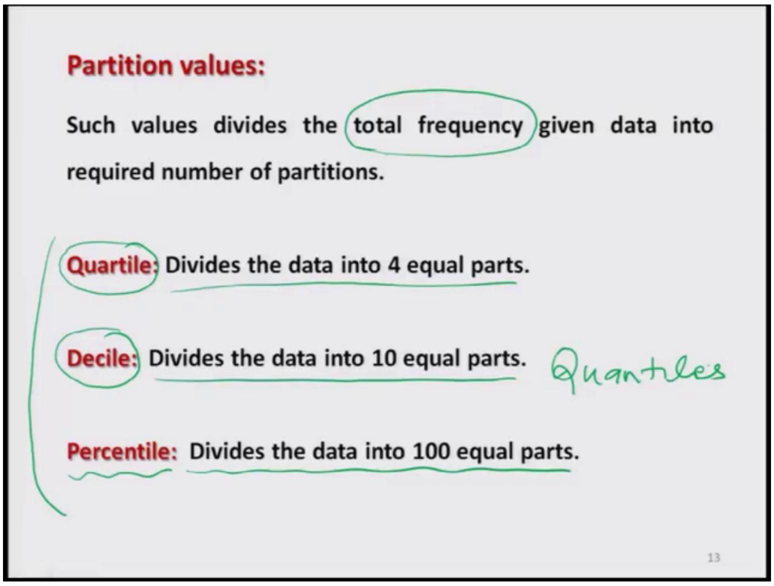
So, you can see here, here is the screenshot, but yeah we would also try to do it ourselves.

(Refer Slide Time: 21:35)



So, I try to create this vector here direction over the R. So, you can see here this is your here direction right and if you try to see here what is the length of direction this is here 100. So, there are 100 values in this vector. Now I would try to find out here the absolute frequencies. So, this is table that is absolute frequencies and in the next line I would write to table divided by length of direction which is coming here the like this which is the relative frequency and the same screenshot I have given in my slides also for better understanding. So, we come back to our slides.

(Refer Slide Time: 22:22)

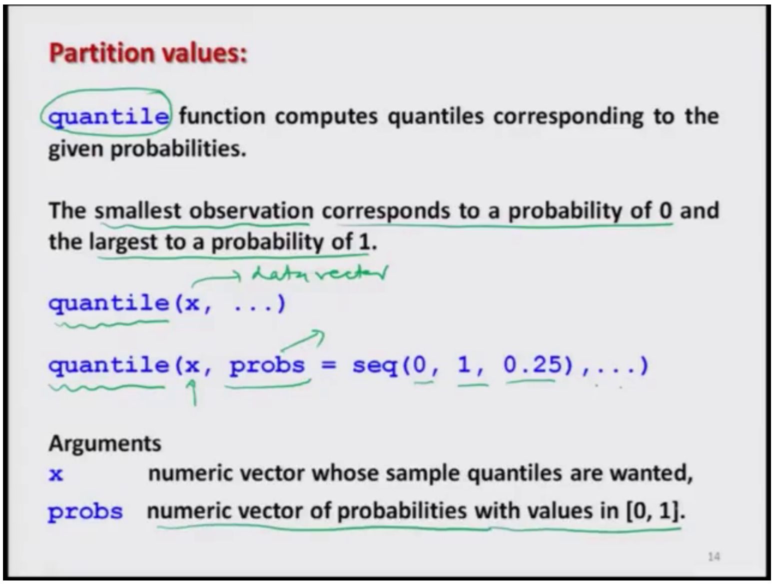


So, this is how we try to compute the absolute and relative frequencies now we come on the other aspect. Whenever we have got a data set we would like to divide that data set in to several partition and one of the approach to partition the data set is to use the total frequency that means, all the data. And this data can be divided into 4 equal parts, say 10 equal parts or say 100 equal parts.

So, when I am trying to divide the entire data into 4 equal parts based on the frequency then every partition is called as quartile, when I am trying to partition the entire data into 10 equal parts every part is called as decile and now when we try to divide the entire data into 100 equal parts then every part is called as percentile right. And nowadays you know that in various examination they have different types of criteria that in order to appear in exam you should be in the top 20 percentile or top 30 percentile and so on. So, these are the same value.

In general these partitions they are called in general as say quantiles, when this quantiles are 4 in number that means, 4 equal parts they are called as quartile when they are divided into 10 equal part they are called as decile and when they are partition in to hundred equal part they are called percentile.

(Refer Slide Time: 24:00)

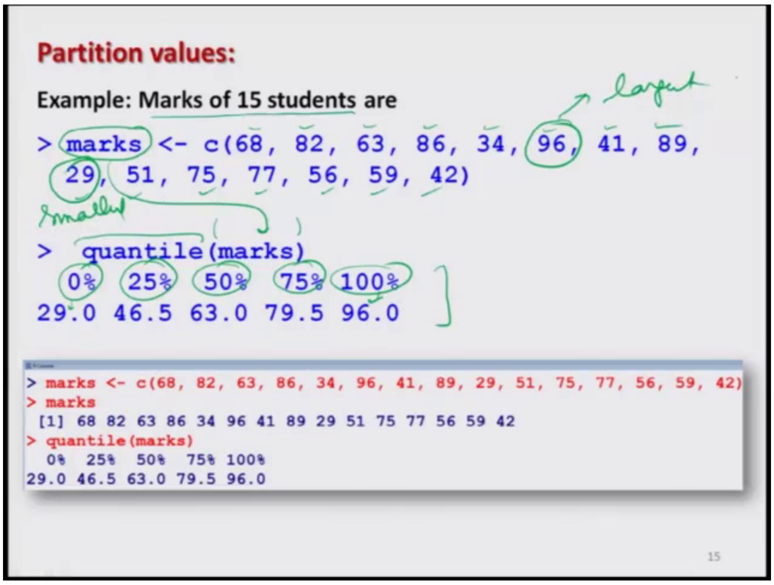


So, in R there is a function say call quantile and this is the function which computes different types of partitions and this partition in this quantile function is based on the probability. Whatever is the smallest observation this correspond to a probability of 0 and whatever is the largest observation this correspond to a probability one because the probability value of any even lies between 0 and 1. So, they are trying to match it with the smallest and largest observation in the data and then they are trying to partition the entire data into different parts.

So, now, in order to find out the quantiles we have 2 option - first is that you simply try to use the function quantile and data vector which is denoted here by x and just enter. The second option is that you can define the probabilities also probabilities means you can divide what percentage of partitions you want or at what values you want to partition the data. For example, I can use this function say quantile with the data vector x and I can give another option say probs p r o b s which is probability and suppose I am writing here this is the sequence of 0, 1 and say 0.25 and there are some other options also.

So, we let us try to take some examples and try to understand that how do we compute the percentiles in the default way and with our own choice. So, and remember one thing this probs, probs is a numeric vector of probabilities which have got all the values in between 0 and 1 right.

(Refer Slide Time: 26:05)

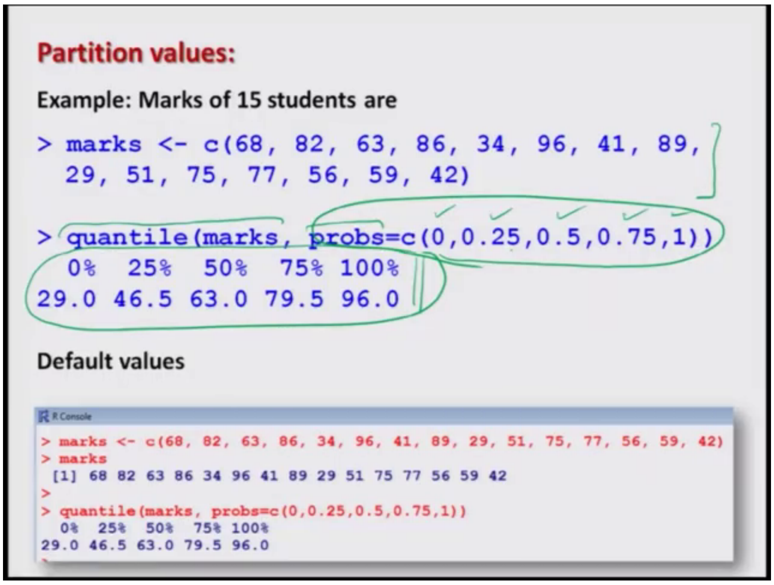


So, now I try to take say this another example in which I have collected the marks of 15 students 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 and now I am storing them inside the vector marks and then I try to compute here their quantiles. So, first of all I try to use the default partitioning.

So, I try to use here quantile q u a n t i l e and inside the argument I give this name of the data vector which is here the marks, you can see here is the outcome this is simply trying to divide it into partitioning with 0 percent; that means, the minimum value is 29. Then 25 percent, 50 percent, 75 percent and then 100 percent, 100 percent which is maximum value which is 96 and you can see also here also that the maximum value here is 96 and the smallest value here is 29. This is the smallest and this is here largest right.

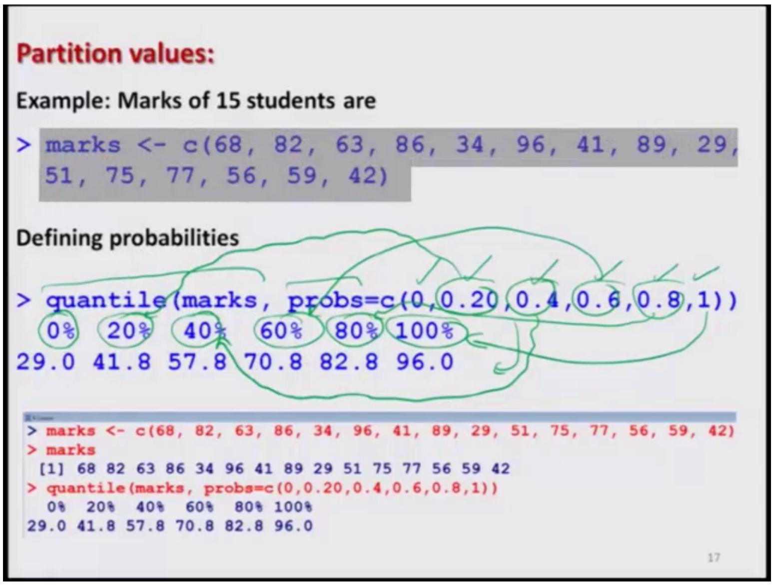
Now in case if you really want to see that what really happens with the function probs then what we try to do here that is the following. I try to take here the same data set and now I try to use the same function quantile over the marks, but I try to define my option probs as say combine vector of 0.25, 0.5, 0.75 and 1.

(Refer Slide Time: 27:30)



And again you see here you get the same outcome. So, you can compare this outcome and the outcome that you obtain in the earlier slide say here this outcome they are the same. So, you can see here that this probs function here in this function is the default function. But now suppose you want to give here a choice and you do not want to use, but you want to compute the quantiles at different partition.

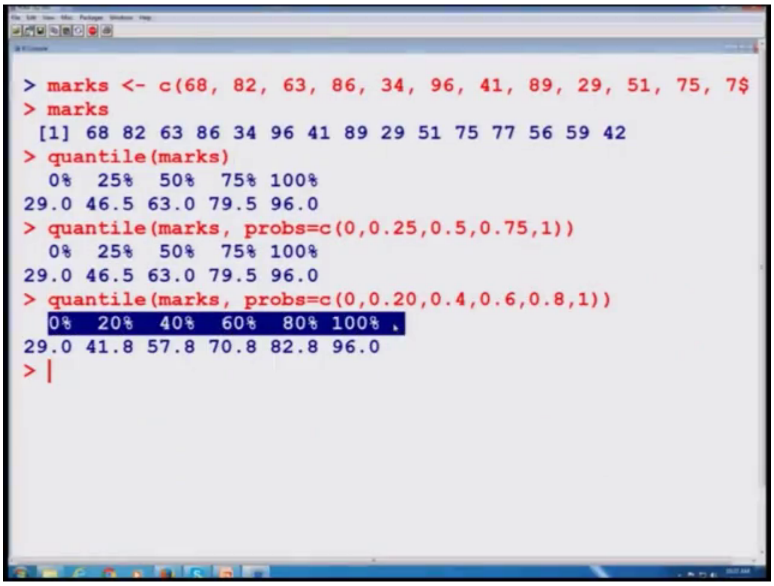
(Refer Slide Time: 28:21)



Now, suppose I want to find out the quantiles at our own choice means I do not want this 25 percent, 50 percent, 75 percent and so on. So, what I try to do here that I try to take the same data set and I use the same function quantile.

But I try to define the partitioning value using the option probs which is different 0, 0.2, 0.4, 0.6, 0.8 and 1 and once you try to enter it you get this type of data. So, you can see here that the partitioning is happening here at 0 percent, 20 percent, 40 percent, 60 percent, 80 percent and 100 percent. So, this 1 is corresponding to 100 percent, this 0.8 is corresponding to say here 80 percent, 0.6 is corresponding to 60 percent, 0.4 is corresponding to 40 percent 0.2 is corresponding to say this here 20 percent and 0 is corresponding to 0 percent. So, this is how you can compute different types of say this here quartiles, but let us try to do it over the R console.

(Refer Slide Time: 29:30)



So, here is my data you can see marks and now I try to find out the quantile that you can see here I am getting the values as 0, 25 percent, 50 percent, 75 percent and 100 percent. Now I try to repeat my examples and suppose I try to compute where there were probs function which is the default function you can get here I am going to get here the same outcome here and here. So, this probs c this function this value these are they are not written here, but still they are by default taken. Now in case if I try to repeat this example where I want to choose the partition ourselves then I can do here like this and you can see here I am getting here the partitions like 0 percent, 20 percent, 40 percent, 60 percent, 80 and 100 percents.

So, here I would like to stop and in this lecture I have try to give you very basic fundamental ideas about the frequencies and partitioning value, you please try to revise these concepts and try to experiment it with some other data set try to create your own data set a smaller data set where you try to do this computation yourself and try to match it with the outcome of the software so that you can gain some confidence. And in the next lecture we will try to deal with some more topics and see you in the next lecture, till then goodbye.