

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

Lecture - 36
Graphics and Plots

Welcome to the next lecture on the course Introduction to R Software. You may recall that in the earlier lecture, we started discussion on the statistical function. And we had talked about how to obtain absolute and relative frequencies, as well as we also learnt how to obtain different types of partitions in terms of quantiles of a function. Now after giving you that elementary introduction on the analytical tools in this lecture, I would try to give you an elementary introduction to the graphical tools right.

As you are aware that there are different types of graphics for example, some of them you already have done something like histogram pie diagram and so on. So, these graphics can also be created in R software, and we would like to see how this can be done. So, let us try to start our discussion there are various types of plots and graphics which gives us information about the data, than formation that is hidden inside the data right. For example, if I try to make here one plot like this.

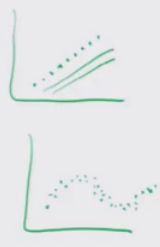
(Refer Slide Time: 01:26)

Graphical tools:

Graphical tools- various type of plots

- 2D & 3D plots, ✓
- scatter diagram ✓
- Pie diagram ✓
- Histogram ✓
- Bar plot ✓
- Stem and leaf plot ✓
- Box plot ... ✓

Appropriate number and choice of plots in analysis provides better inferences.



These points are indicating some data values you can see here, that they are lying on a straight line nobody is telling us, but this information is giving us the same thing. On the

other hand, if I have a data which is like something like this. You can see here this is indicating that there is a particular type of function which is hidden inside the data.

So, this type of information can also be retrieved using the different types of plots. So, there are various types of plots in statistics we use 2 dimensional plots 3 dimensional plots scatter diagram, pie diagram, histogram, bar plot, stem and leaf plot, box plot etcetera. And there is a long list and one thing you have to keep in mind that sometime we feel that a statistical analysis good only when there are large number of graphs, no. Just use appropriate number of graphs and appropriate graphs.

That is the moral of the story, but here we would like to see that what can we do in R. In R there are various types of graphics that can be created. Bar plot, pie chart, box plot grouped, box plot scatter plot coplots histogram, normal quantile, quantile plots and everything. There is a long list, but here we are going to learn about some of the things some elementary graphics rather.

(Refer Slide Time: 02:55)

Bar plots:
Visualize the relative or absolute frequencies of observed values of a variable.

It consists of one bar for each category.

The height of each bar is determined by either the absolute frequency or the relative frequency of the respective category and is shown on the *y-axis*.

```
barplot(x, width = 1, space = NULL,...)  
> barplot(table(x))  
> barplot(table(x)/length(x))
```

4

So, first of all we start with simple plot which is the bar plot. What is the use of bar plot? We have understood; what is the concept of absolute frequency and relative frequency. So, this bar plots gives us the same information in the graphic mode, and they try to create one bar for one category right.

(Refer Slide Time: 03:28)

Bar plots:
Visualize the relative or absolute frequencies of observed values of a variable.
It consists of one bar for each category.
The height of each bar is determined by either the absolute frequency or the relative frequency of the respective category and is shown on the y-axis.

```
barplot(x, width = 1, space = NULL, ...)  
> barplot(table(x)) help("barplot")  
> barplot(table(x)/length(x))
```

So, this bar plot helps us in visualizing the relative and absolute frequencies of the data. And this consists of one bar for each category. So, if there are 2 categories there will be 2 bars if there are 3 category there will be 3 bars and so on. And how this bar has been created? The identification tag is height, height of the bar. The height of the bar is determined by the absolute frequency or the relative frequency of the respective category; that means, if the absolute frequency is higher the height of the bar will be higher.

And similarly if the relative frequency is higher, then the height of the bar will also be higher. This is how we try to interpret it and this relative frequency or the absolute frequency that is denoted on the y axis and these bars are constructed on the x axis. The syntax for creating bar plot in R is `barplot`, that is simply bar plot and inside the arguments first option is to write the data vector and there are some other options also.

But here we would like to restrict our discussion to an elementary level. So, I am not going to discuss about other options, but you can simply use help something like help. Inside double quotes write bar plot and this will take you to the internet site and then you can get all the information about bar plot. In order to create a bar, plot we use the function bar plot and inside the arguments we use the function table x. And this will give us bar plot which is based on the absolute frequency. And in case if I want to create a bar plot with respect to the relative frequency, then inside the argument I have to give the

option for relative frequency that is table of x divided by length of x that we did in the last lecture.

So, now I try to take some example.

(Refer Slide Time: 05:55)

```
Bar plots:  
> help("barplot")  
  
barplot(height, width = 1, space = NULL,  
names.arg = NULL, legend.text = NULL, beside  
= FALSE, horiz = FALSE, density = NULL, angle  
= 45, col = NULL, border = par("fg"), main =  
NULL, sub = NULL, xlab = NULL, ylab = NULL,  
xlim = NULL, ylim = NULL, xpd = TRUE, log =  
"", axes = TRUE, axisnames = TRUE, cex.axis =  
par("cex.axis"), cex.names = par("cex.axis"),  
inside = TRUE, plot = TRUE, axis.lty = 0,  
offset = 0, add = FALSE, args.legend = NULL,  
...)
```

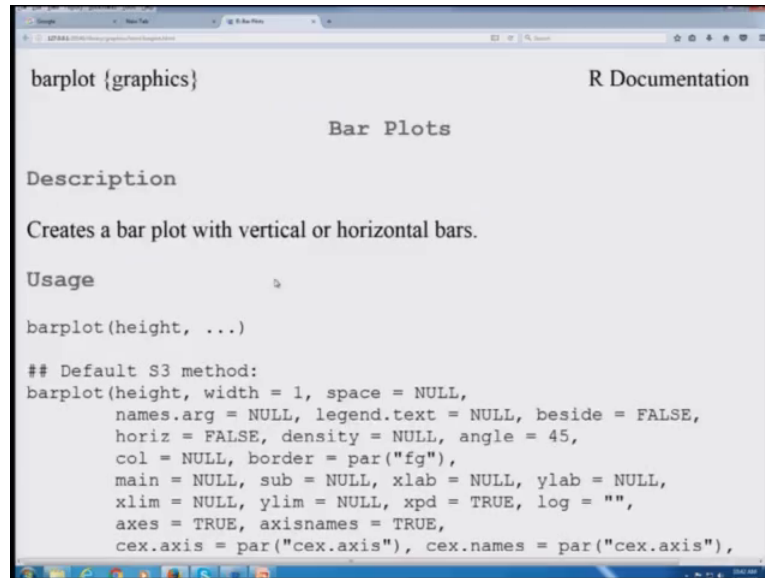
You can see here I have given a screenshot. Now in order to show you that how to obtain the help on the bar plot, let me go to the R console. And here I type help then bar plot. And you can see here this comes out to with the R server. And here you can see here this all this details are there.

(Refer Slide Time: 06:05)

```
> help("barplot")  
starting httpd help server ... done  
> |
```

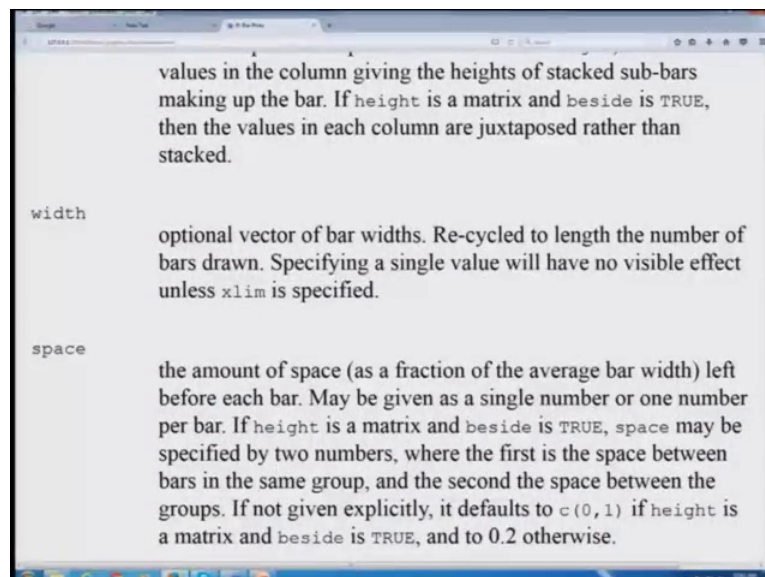
So, I am not going to discuss these details, but I will request you to go through with details. And here I have given the options which are available on the site. Now, we try to take some example and we try to create the bar plots.

(Refer Slide Time: 06:16)



```
barplot {graphics} R Documentation  
  
Bar Plots  
  
Description  
Creates a bar plot with vertical or horizontal bars.  
  
Usage  
barplot(height, ...)  
  
## Default S3 method:  
barplot(height, width = 1, space = NULL,  
names.arg = NULL, legend.text = NULL, beside = FALSE,  
horiz = FALSE, density = NULL, angle = 45,  
col = NULL, border = par("fg"),  
main = NULL, sub = NULL, xlab = NULL, ylab = NULL,  
xlim = NULL, ylim = NULL, xpd = TRUE, log = "",  
axes = TRUE, axisnames = TRUE,  
cex.axis = par("cex.axis"), cex.names = par("cex.axis"),
```

(Refer Slide Time: 06:27)



```
values in the column giving the heights of stacked sub-bars  
making up the bar. If height is a matrix and beside is TRUE,  
then the values in each column are juxtaposed rather than  
stacked.  
  
width  
optional vector of bar widths. Re-cycled to length the number of  
bars drawn. Specifying a single value will have no visible effect  
unless xlim is specified.  
  
space  
the amount of space (as a fraction of the average bar width) left  
before each bar. May be given as a single number or one number  
per bar. If height is a matrix and beside is TRUE, space may be  
specified by two numbers, where the first is the space between  
bars in the same group, and the second the space between the  
groups. If not given explicitly, it defaults to c(0,1) if height is  
a matrix and beside is TRUE, and to 0.2 otherwise.
```

So, here you see we have considered the same example that we considered in the last lecture. And I also requested you to keep in mind 2 examples the data on pizza delivery.

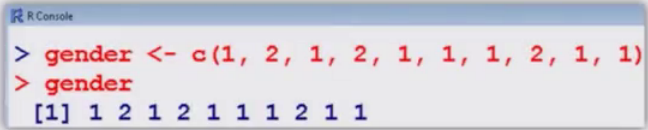
(Refer Slide Time: 06:39)

Example:

Code the 10 persons by using, say 1 for male (M) and 2 for female (F).

M, F, M, F, M, M, M, F, M, M
1, 2, 1, 2, 1, 1, 1, 2, 1, 1

```
> gender <- c(1, 2, 1, 2, 1, 1, 1, 2, 1, 1)
> gender
[1] 1 2 1 2 1 1 1 2 1 1
```



6

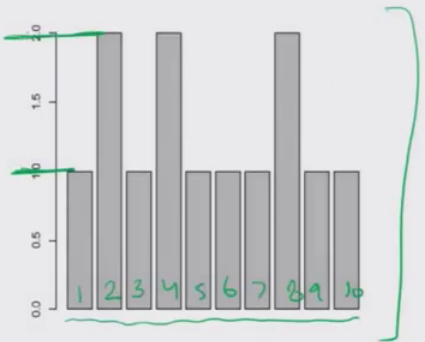
And this the data on 10 persons who are categorized into 2 categories male and female and their categories are denoted by one and 2 one for male and 2 for female and this data was stored in a variable gender right.

(Refer Slide Time: 07:18)

Bar plots:
Example

```
> barplot(gender)
```

Do you want this?



7

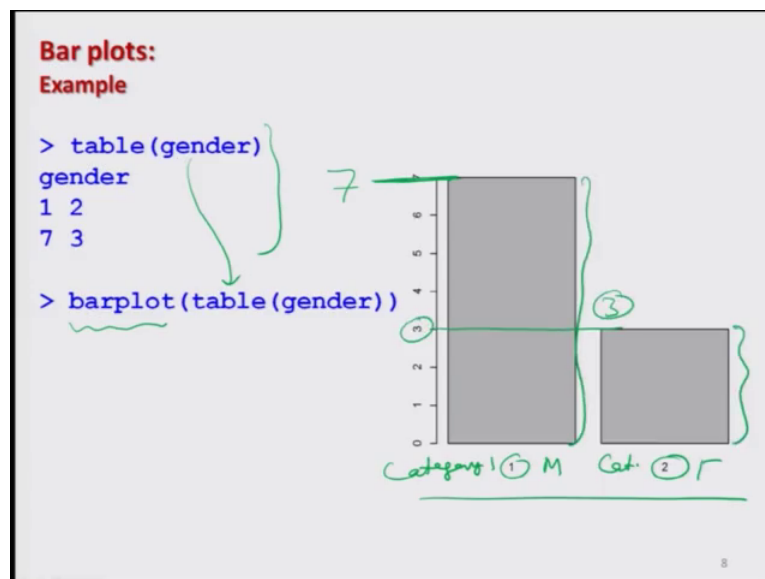
So, now I try to create the bar plot. And as usual I simply try to write down bar plot inside the variable gender. And I get here this thing. Do you really want to have this type of plot we discuss that there are 2 categories? There are 2 categories, one is for male and

another is for here female. There are 7 male members and there are 3 female members. So, I really there should be only 2 bars what is really happening over here.

If you try to see here we have made here one mistake, I have used here the variable directly I have not using the table function that is why this column is coming and actually is not a problem this is trying to give you the bar plot for each of the observation. This is for observation number one, this is for observation number 2 3 4 5 6 7 8 9 10. So, you can see here you had got here 10 observations. And every observation has value 1 and 2.

For example: if you trying to concentrate on the first 2 observations, first observation is one and second observation is 2. And the same is denoted here this is here one the height is proportional to the value one and this height is proportional to the second observation 2 right. So, this is a mistake being careful.

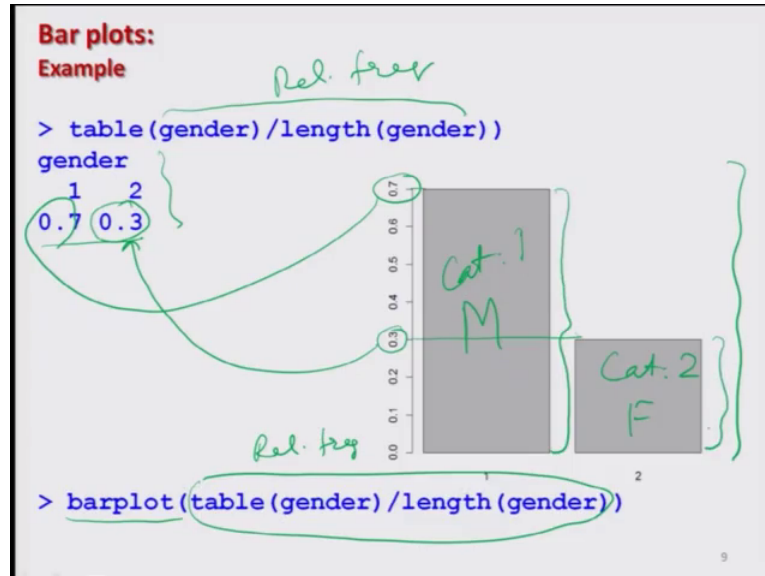
(Refer Slide Time: 08:44)



Now, I obtain the table and then I try to create the bar plot. So, if you try to see you over here I have obtain: first that table of this variable gender and then I am trying to create here the bar plot of this variable table gender, and I get here this type of graphic. So, you can see here that here this is denoting the category one, which is per year male and this is the category 2 per year females. And you have seen that there are 7 male and 3 female. So, this number here is 7 and this number here, you can see this is here 3. So, their heights are proportional to their absolute frequencies.

And similarly if you want to plot the same bar plot with respect to the relative frequency you can see here.

(Refer Slide Time: 09:39)



Here we are trying to find out the relative frequency and I am using here the bar plot function with the relative frequency.

So, I obtain here this graphic. Now suppose I want to construct this bar plot with respect to the relative frequencies. So, what I try to do here I am trying to compute the relative frequencies. And they come out to be like this that we had obtained earlier, 0.7 0.3, and then I use the bar plot function with the relative frequency. So, this is over here relative frequency and we get here this type of here graphic. So, you can see here this value here is 0.7 which is this value and this value here is 0.3, which is here this value and the height of the first category this is my category number one which is here male and this is my here category number 2 which is here female, you can see here this is 0.3 and this is 0.7.

And in both the graphics this is simply trying to indicate that the height of this bar is more than this bar. So, that is really indicating that the number of male persons are higher than the number of female person in the given data set. So, now, I (Refer Time: 11:05) try to do the same exercise with the R software.

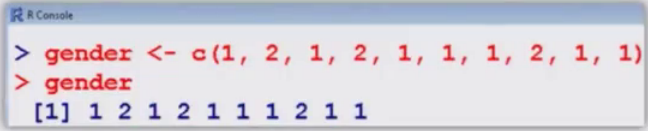
(Refer Slide Time: 11:10)

Example:

Code the 10 persons by using, say 1 for male (M) and 2 for female (F).

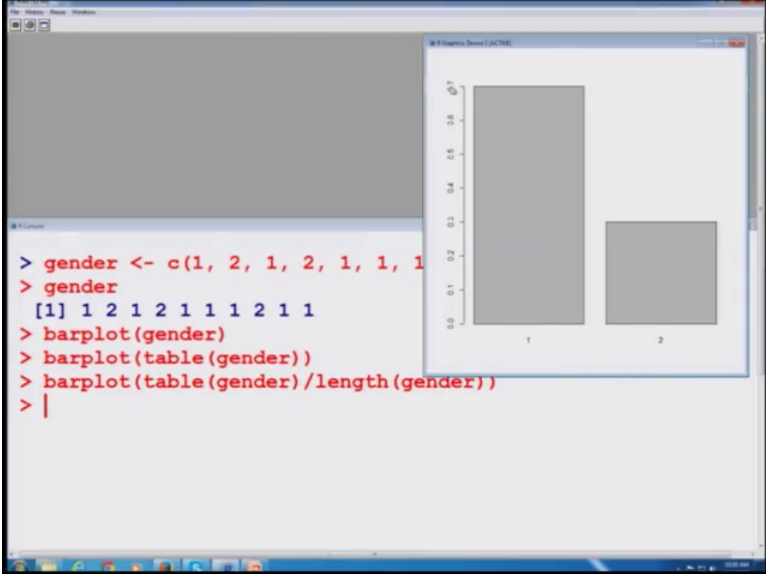
M, F, M, F, M, M, M, F, M, M
1, 2, 1, 2, 1, 1, 1, 2, 1, 1

```
> gender <- c(1, 2, 1, 2, 1, 1, 1, 2, 1, 1)
> gender
[1] 1 2 1 2 1 1 1 2 1 1
```



6

(Refer Slide Time: 11:20)



```
> gender <- c(1, 2, 1, 2, 1, 1, 1, 2, 1, 1)
> gender
[1] 1 2 1 2 1 1 1 2 1 1
> barplot(gender)
> barplot(table(gender))
> barplot(table(gender)/length(gender))
> |
```

Gender	Count	Proportion
1	7	0.7
2	3	0.3

So, you can see here I try to copy this here gender and this is our data set gender. And as usual I try to make here a mistake by simply writing gender variable with bar plot, you can see here I get this thing right. Now, I try to go with here. And now here you can see when we are trying to move forward, do not you think that it is more easy to use the R studio possibly you have forgotten that we had discuss about the R s studio software in the initial lectures.

So, now we are going to a place where you can see that here I am trying to write down the command and this graph is coming over a different window. And then I have to switch over this, but if you are using the R studio software than these things are given directly on the same window we will try to show you later on. So, you can see here when I try to do here bar plot we with respect to absolute frequencies this comes out to be like this. And when I try to create the bar plot with respect to a relative frequencies, this comes out to be here like this. You can see here this value here is 0.7 where my cursor is moving now.

So, now you know that how to create these things. And same thing I would now try to create with my another data set direction that we had used in the earlier lecture.

(Refer Slide Time: 12:59)

Example

'pizza_delivery.csv' contains the simulated data on pizza home delivery.

- There are three branches (East, West, Central) of the restaurant.
- The pizza delivery is centrally managed over phone and delivered by one of the five drivers.
- The data set captures the number of pizzas ordered and the final bill

```
> setwd("C:/Rcourse")  
> pizza <- read.csv('pizza_delivery.csv')
```

10

This data set was derived from the data set pizza delivery that we discuss in the earlier lecture. And we had extracted information on one variable called as direction.

(Refer Slide Time: 13:11)

Example

Consider data from Pizza. Take first 100 values from Direction and code Directions as

- ❖ East: 1
- ❖ West: 2
- ❖ Centre: 3

```
direction <-c(1,1,2,1,2,3,2,2,3,3,3,1,2,3,2,2,3,1,
1,3,3,1,2,1,3,3,3,2,2,2,2,1,2,2,1,1,1,3,2,2,1,2,3,2
,2,1,2,3,3,2,1,2,2,3,1,1,2,1,2,3,2,3,2,2,3,1,2,3,3,
3,2,1,1,1,2,1,1,2,1,2,3,3,1,2,3,3,2,1,2,3,2,1,3,2,2
,2,2,3,2,2)
```

11

And if you recall that there were 3 direction east west and center which we are coded as say 1 2 and 3; and we had obtain this type of data that is stored in a vector say here direction. Now I would like to create this bar plots with this direction vector.

(Refer Slide Time: 13:31)

Bar plots:
Example

```
> barplot(direction)
```

Do you want this?

12

So, you can see here when I am using here directly here the direction variable without using the table function I am getting this type of graphic, you can see here these are actually hundred values here one second category is here 2.

And third category here is 3 for the east west and center directions, but this is the graph that we do not want. So, we try to do the correction and we obtain the correct graphics you can see here. Now I am using here a function. So, the Intel index become bar plot of table direction which is then absolute frequencies. So, there are 3 types of categories 1 2 and 3 and you can see here.

(Refer Slide Time: 14:17)



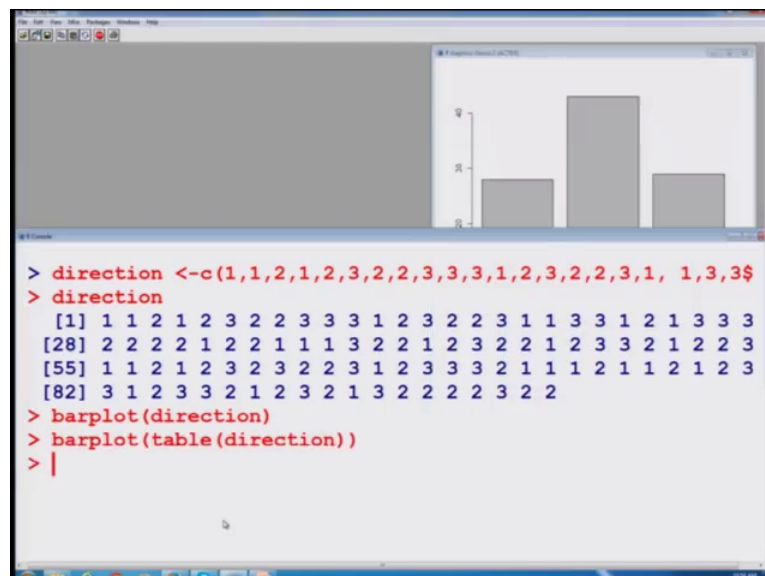
Now, it is giving us this category one category 2 and category 3. So, that is simply indicating you can see here it is saying that the category 2 is highest. Then there is a small difference with the category 3 and category 1. So, I can order their number of orders for the pizza they are the highest in the category 2 followed by then category 3.

(Refer Slide Time: 14:48)



And minimum in category number one. So, if I try to do the same thing with the relative frequency I can write down the command for the relative frequency here inside the bar plot arguments and means I am getting a similar information, but now all this values are in fraction because they are corresponding to the proportion between 0 and 1 right. So, now, let us try to do the same thing on the R console also. So, first I try to copy here the data.

(Refer Slide Time: 15:17)



So, this is my here data. You can see here. And then I try to make here a bar plot directly with the data. So, you can see here that is giving you this type of picture which I had pasted on my slide. There are hundred categories over here, but this is the thing which we do not want we want to use here the table function.

So, I try to use here this command and when I try to plot the bar plot with this command, I get this type of bar plot based on the absolute frequencies this is category one this is category 2 and this is category 3. So, there are 3 categories and 3 bars and similarly if you want to have it here with respect to the relative frequency. Then I can use the relative frequency inside the arguments of bar plot function and you can see here we get this thing this curve over here and here the values are now here, 0.4 instead of 40 right. So, this is how we try to create the bar diagram.

Next important diagram is pie diagram. So, what is the pie diagram or a pie chart? This is also use to visualize the absolute and relative frequencies, but in this case they are not the bars, but they are the sections of a circle right.

So, a pie chart is essentially a circle which is partition into different segments and every segment represents a particular category. For example, in case if I have 2 categories as in the example of male and female. So, I will have a 2 partition 2 segments inside a circle and similarly if I have some more data, where I have got 5 categories then there will be 5 partitions and they are indicated by different colors and different types of values.

(Refer Slide Time: 17:40)

Pie diagram:
Pie charts visualize the absolute and relative frequencies.

A pie chart is a circle partitioned into segments where each of the segments represents a category.

The size of each segment depends upon the relative frequency and is determined by the angle $(\text{frequency} \times 360^\circ)$.

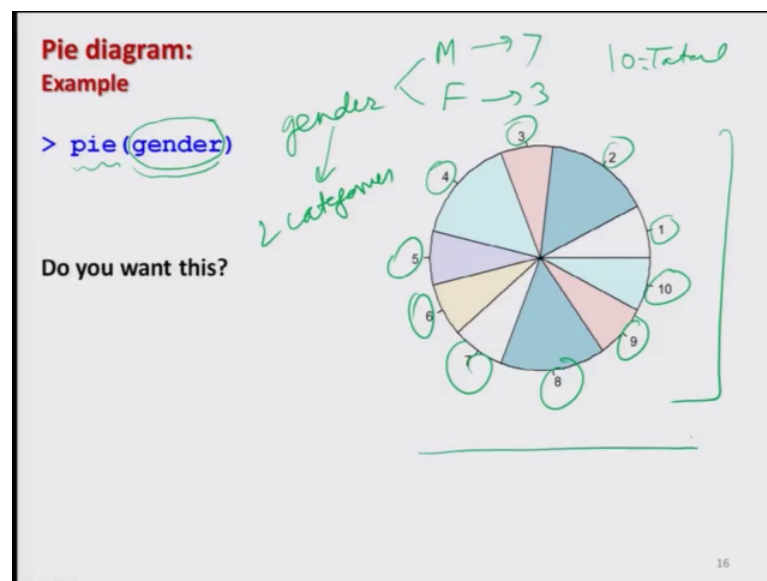
```
pie(x, labels = names(x), ...)
```

data *help("pie")*

15

So, the question is how this size of segment is computed. The size of the segment is computed by this formula frequency into 360 degree right. And the syntax to construct a pie chart is `pie` and inside the arguments you can write here the data. And then you can give here different types of options are there to get labels and other thing I will suggest to use the help with the option here. And then try to look about a different options for example, it is possible to change the color if it is possible to give different types of labels all those things are there, but here my objective is to retain the elementary level of this introduction, ok.

(Refer Slide Time: 18:52)

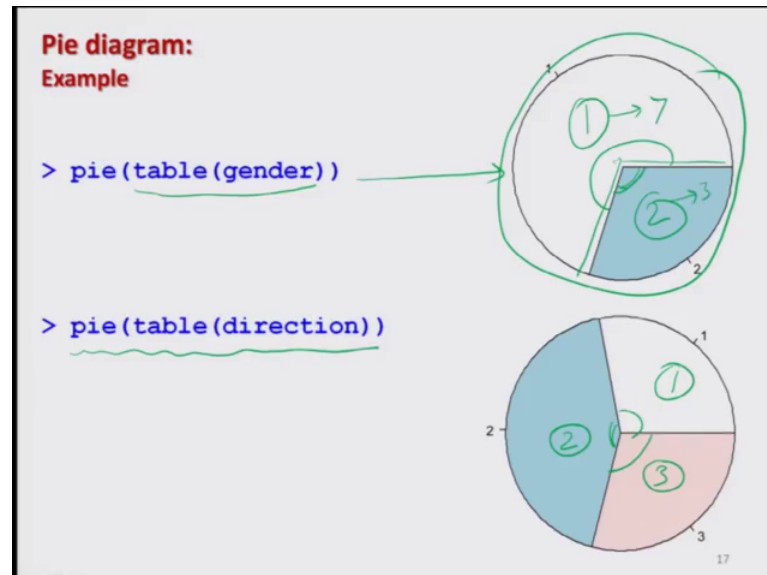


So, now I try to create some pie diagram using the same data set. So, if you remember we had created a variable `gender`. Which has 2 categories male and female, and there were 7 males and 3 females and total were 10, 10 in total right. And now if I try to create a pie diagram I simply have to give here `pie`, and I give here inside the argument `gender` and you can see this this type of outcome.

I am going to get what is this do you really want this there are so many categories how many 1 2 3 4 5 6 7 8 9 and 10, but how it is possible there are only 2 categories in your data male and female. So, because I have use here directly the `gender`, I have not use here the table `gender` that is the common mistake which people make so, but this is not a mistake R is doing whatever you ask to do it you get the you get a and divided the data into 10 different categories. So now, we try to correct our mistake and we try to construct

the pie diagram using the say absolute frequency, and this can also be done with the relative frequency that will not make any difference.

(Refer Slide Time: 20:04)

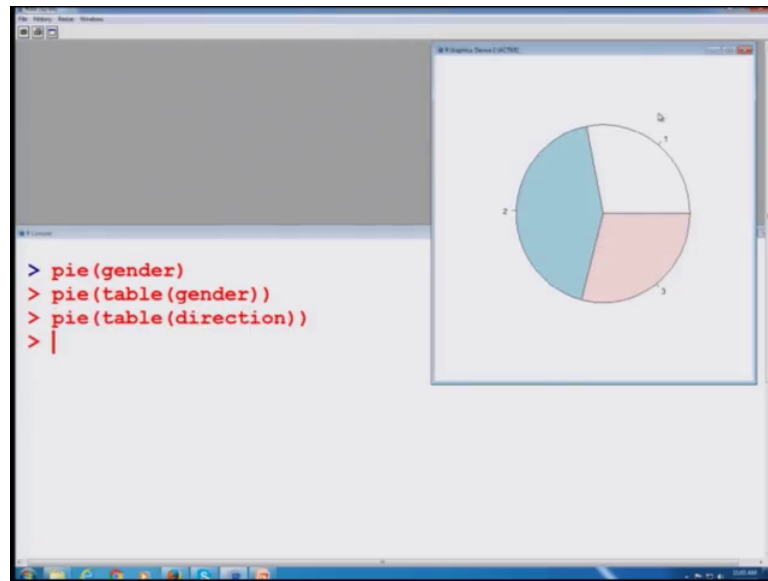


Because, the angle will be remain the same whether it is 0.7 or 7 because angle is computed with respect to the total frequency.

So, when I try to do it I get here this type of data. So, you can see here that here is a circle and the circle is divided into 2 categories, one category is this and 2 category is given here in blue color. And similarly when I try to use this pie function to create the pie chart for another data set direction, you can see here there are 3 categories one 2 and 3. So, one is given by here white color 2 by blue color and 3 by another color and you can see here that this angles here, they are proportional to the frequency you can see here in the first case this angle is smaller than the second angle because the number of values in the category one is 7 and the number of values in the category 2 is 3 right.

So, this is how we try to create the pie chart, but before that we try to do it over the R console ourselves right.

(Refer Slide Time: 21:26)



So, first I try to use here only the gender, and I try to show you what mistake we can make. You can see here that is giving you 10 categories. Now I try to correct my mistake and I try to say construct the pie chart, where the absolute frequency. So, you can see we are getting here the same thing that we had seen in the slide. And similarly if I try to use it over another variable here direction, you can see here we get here this type of chart. So, you can see here we can get this pie chart without any problem.

(Refer Slide Time: 22:17)

Histogram:

Histogram is based on the idea to categorize the data into different groups and plot the bars for each category with height.

The area of the bars (= height X width) is proportional to the relative frequency.

So the widths of the bars need not necessarily to be the same

```
hist(x) # show absolute frequencies
hist(x, freq=F) # show relative frequencies
```

See `help("hist")` for more details

18

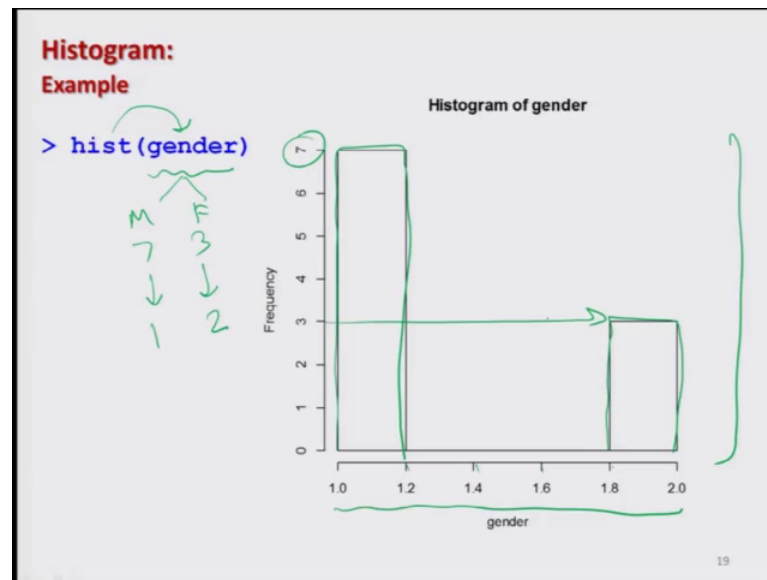
And that is not difficult to obtain. Now we come to our next chart that is histogram now histogram is another popular graphics. And the histogram is based on the idea where we try to categorize the data into different groups. And for each of the group we plot a bar. And the discrimination between a bar chart and a histogram; that means, the bar of a bar chart and the bar of a histograms comes here. In bar chart the height of the bar is proportional to the absolute frequency, but in a histogram the height of the bar is proportional to the area of the bar, that is what you have to keep in mind. And area of the bar is here height into width usually we have seen the histogram, where the width of the bar is kept the same.

And that is why we only try to look on the height of the bars, but that is not actually the correct reason this width can also be different. So, in case if the width increases then the height of the bar will also change that will actually decrease. So, that is the point which we have to keep in mind that the area of the bar is the parameter which determines the bar of a histogram. And then there is no condition also that the width of the bar has to be same they can be different also depending on the situation in order to create a histogram in say R we have a command `hist`. And inside the argument we try to give here the data and this function will create a histogram using the absolute frequencies.

And suppose you want to create the histogram with respect to the relative frequencies then you have to write like this. Command `hist` inside the argument write the data and you have to write down here frequencies equal to false this here `f` is the logical value which is say here false. So obviously, in the first option where I am not writing anything actually I am essentially writing frequencies equal to true, that is the default value right and if you want to have more details on the `hist` option please try to look into the help menu that is not difficult for you now to see right.

Now I try to consider my earlier data set and I try to construct histogram over them.

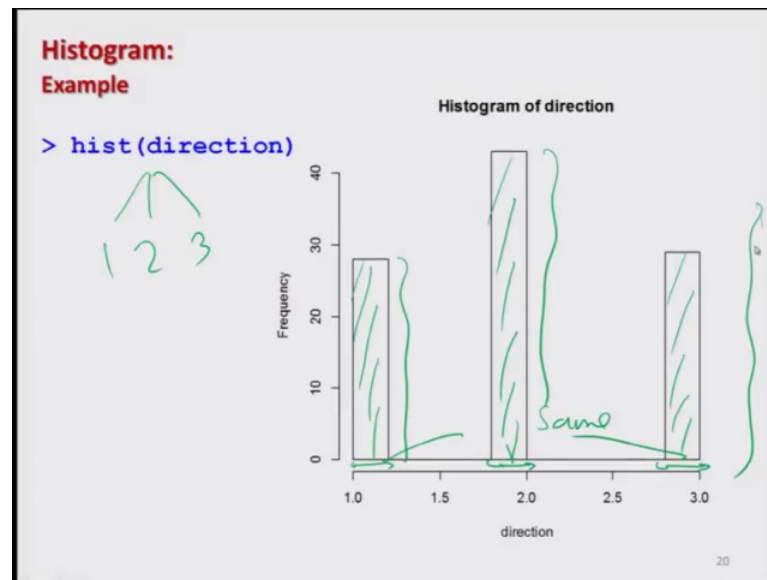
(Refer Slide Time: 25:09)



So, first of all I try to take the same data on gender which has 2 categories male and female there was 7 males and 3 females and when I try to write down hist gender, you can see here I get this type of graphics and now you see how to interpret it. You can see here we had given the values for male as one and for female as 2. So, that is why it is trying to divide the value between one and 2. In some equal partition because the software does not understand the difference between a continuous and a discrete variable, right. So, it is trying to say here this is my here category number one, where there are 7 values and there is another value here 2 which is here at 3.

So, you can see here in this case you need not to compute the absolute frequencies or relative frequencies yourself, but histogram will compute it automatically right.

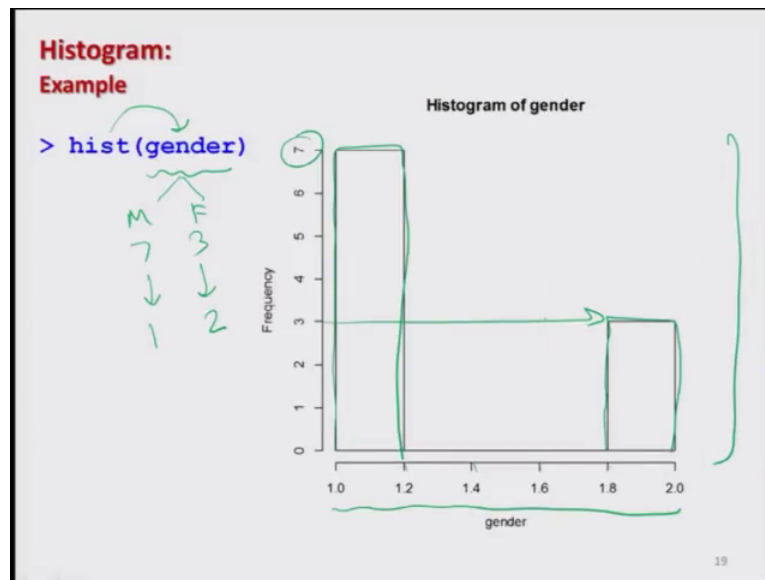
(Refer Slide Time: 26:07)



Similarly, when you try to go for direction data there were 3 values 1 2 and 3 there were 3 categories 1 2 and 3. So, if I try to use this thing you get here this type of data. And you can see here that this area, this area, and this area, that is proportional to the frequency, but since here the width are going to be here same, you can see this is the width they are going to be here same.

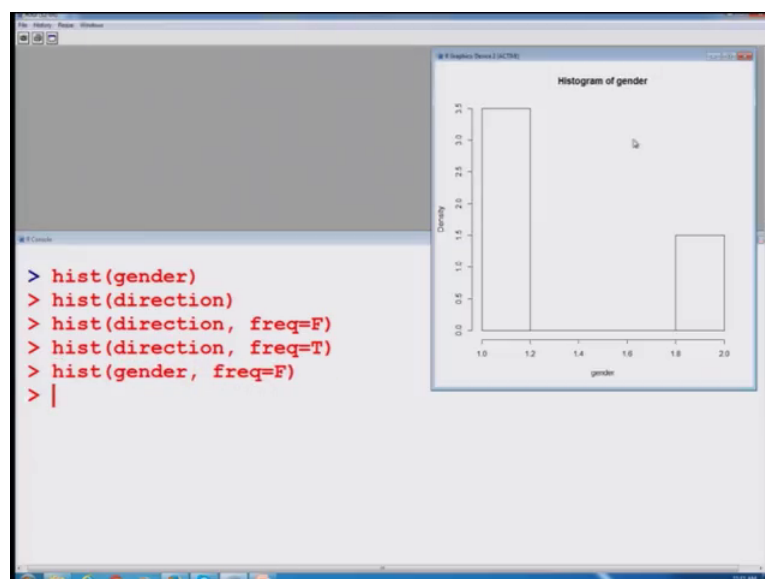
So, that is why I can take a conclusion based on the height only right, but there is an option to change the width also, but I would like you to explore yourself by looking into the various arguments and various options which are given in the histogram. So, now, let us try to create this histogram over the R console.

(Refer Slide Time: 26:57)



So, I will try to make it here histogram of gender and histogram of say here. So now, (Refer Time: 27:19) try to create this histogram. You can see here I will use here that same data set hist gender. So, you can see here I am getting this thing. And similarly you can plot the histogram with the data direction. And you can see here that it is coming here like this right. And here you can see here all this value where I am moving my cursor this values here 40.

(Refer Slide Time: 27:15)



So, that is based on the absolute frequencies right, if you want to use here the relative frequencies I have to give here frequencies is equal to false. And if you try to look over here this value is going to be change initially this is 40. And as soon as I enter here this becomes here say here something else because now, that is based on the relative frequency and in the same command if you try to make it here true you will again get here the 40 value you can see here right. Similarly, if you want to have the histogram with relative frequencies for the data gender you can just create here gender, you can see here that is the data with relative frequency.

So, now we conclude our lecture here. And we have discuss bar plots pie chart and histogram. And we have restricted our discussion to an elementary level now there are different types of options for example, if you want to change the color in a pie chart or if you want to make a particular type of label you want to give a name to different categories simply in say histogram.

Or in bar plot you want to write something on the x axis something on the y axis as per your requirement, these things can also be done and these are very simple thing. You simply have to go to the entire command of say bar plot pie or say histogram, and then you have to simply give this option inside the argument. And you have to give the required values over there. And you will get the graphics in the way you want.

So, we stop here. And in the next lecture we will come up with some more details on statistical function, till then goodbye.