Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 04

Lecture – 17

Pie Diagram, Histogram and Multiple Histogram

Hello friends, welcome to the course multivariate procedure with R. You can recall that in the last lecture we had discussed the bar plots and its different variants when they can be extended to multiple or say multivariate case. Now the bar plots are used when we have a discrete data or a categorical data means the data can be divided into different categories. For example, if I can say we have a sample of students which are male and female, so I can have two categories. So, the bar plots can be created where each bar is going to represent the absolute or the relative frequency of the total number of male or female students. And then we also discussed that how you can modify, how you can update, how you can make the bar plots more informative.

And the moral of the story was I had given you some of the options and more or less the similar option or the same options will continue to other graphics also. So, now in this lecture we are going to do two types of graphics. One is here pie chart. Pie chart I am taking just to give you a confidence because bar chart, pie chart they are the graphics which you study right from the beginning in your career as a student.

So, I just want to make you confident that okay these things can be done and then I will be going to histogram. The main difference between say bar chart or say pie chart and histogram is that bar chart or its variant, pie chart or its variants they are used when we have the categorical data whereas histogram is used when we have a continuous data. What is continuous data? Suppose if I say height of the students, so height of the students can be 155 centimeter, 156 centimeter or even 155.1 centimeter and so on. Well, there is a very proper and say good concept of continuous random variable that we will be discussing later on that how to define a continuous variable in a mathematical way.

But in practice I can always say that when the variable is trying to take all possible values right then it can be considered as a continuous variable that can be height of the students, weight of the crop, blood pressure, temperature, etc. So, these are continuous

variable. So, this is what you have to remember when there is a categorical data you can plot bar plot pie diagram, when you have the continuous data then you have to plot histogram. So, now let us try to begin this lecture and we try to see these concept and how to create them in the R software right okay. So, what is the pie diagram? This pie charts visualizes the absolute and relative frequencies and it is just like a bar plot also.

The only difference between bar plot and pie chart is that bar plots in the bar plots the data is represented as a bar and in the pie chart it is represented as the different sections of a circle right. So, a pie chart is a circle partitioned into segments and where each of the segments represent a category. Just like a bar represent a category here the segment represents a category and the size of each segment depends on the relative frequency right. It is similar to the height of the bar depends on the frequency or relative or absolute frequency. Similarly, the size of the segment like this one, this is one segment, this is another it depends on the relative frequency and the size of each segment is determined by the angle that means this angle or this angle is going to determine the size and it is obtained as relative frequency into 360 degrees right.

Now in case if you want to create a pie diagram then the R command is p i e and inside the parenthesis you have to write here the data, pie(data), then you have labels, then you have different types of option just like what you had seen in the case of bar chart also, bar plot also there were many options. For example, you can see here labels, radius, mean, color, clockwise, etc. So, this label is going to give you the idea about the description with the slices of the circle and about this radius, radius indicates the radius of the circle of the pie chart and it takes the value between minus 1 and plus 1. Mean just like others it will give you that title of the chart, c o l it is going to give the colors with the different segments or color to the pie charts and if you say here clockwise so clockwise is used to indicate if the slices are drawn clockwise or anticlockwise and it is a logical variable that means it takes the value true or false. So, these are some different commands which are specific only to the pie diagram that is why I am trying to illustrate them here but I request you that you please look into the help menu and try to understand the remaining options.

So, now let me try to take the same example which I took in the case of bar plot that there are three sales persons in a shop which are denoted by 1, 2 and 3 and now we try to take the data of first handed customers entering into the shop and we try to record which sales person has helped them. So, this number 1, 1, 2, 1 etc it is written here and this data has been stored here as a sales person. Now if you try to create here this pie diagram so first as I said in the case of bar diagram also it is going to represent different categories so for that first we have to create the table that means we have to tabulate the data. So, that first we try to create here that the frequency table so table sales per will give you here this data that means in the sales number 1 has helped 28%, sales person number 2 has helped 43% and sales person number 3 has helped 29 customers. So, now if you try to create the

pie diagram using the pie and say table sales per command this will here look like this right.

So, this you can see here sales man 1, this is here sales man 2 and this is a sales man 3 and this is the angle here which is trying to determine the size of the segment and for example the size of the segment for the sales person number 1 is like this the size of the segment for the sales person number 2 it is here like this and so on. So, this is how you can create the pie chart and if you want to make it more informative you can add different types of information to this pie diagram for example if you try to say here labels is equal to say this see SP1 SP2 SP3 means sales person 1, sales person 2, sales person 3 you can see here it is here like this right and if you want to give here a title that sales persons attending the customers so you can use it under main and within double code sales person attending customer and it will appear here and if you want to make it here the colors so you can give the colors whatever you want suppose I am giving here C O L is equal to green red and blue so now they are appearing in the same order green red and here blue you can see right. So, you can see here it is not very difficult to do this the creation of part of this pie chart. So, now let me try to show you this thing in the R console so you can see here this is here like this and now if you want to create here the pie chart you will see here it is coming out of here like this same sheet same graphic which I shown you on the slide and if you want to make it here more informative by adding more information it will come here like this right. So, this is how we try to create these graphics in the R software right.

Now let me try to give you the here the idea of histogram. I am sure that you all are actually aware about these graphics but as I said my idea is that I want to give you confidence that you can create this thing and I want to extend this thing to a multiple case also right and then how to use them in the R software and how you can make them more informative. So, histogram is based on the idea to categorize the data into different groups and plot the bars for each category with height. Now what is the difference between then bar plot and histogram both are going to give you here the bars for example bar plot will give like this and histogram will give like this. They will be adjacent bar because it is created that the data has to be continuous.

Now if you remember what I said what I told you that the height of the bars in the bar plot that is proportional to the frequency or absolute frequency whereas in the case of histogram the area of the bars like as if I say here this is the area of this bar, this is the area of this bar and so on this proportional to the frequency or the relative frequency. So, if you try to see the area of the bars depend on the height and the width of the bars. What does this mean? This is here the width of the bar and this is here the height of the bar. So, this is what we mean that the area of the bar is proportional to the frequency or relative frequency. So, this means what? If I try to create here a histogram like this and histogram like this it may be possible that the area of this bar and area of this bar that is the area of

bar number 1 and area of bar number 2 they may be same but by looking at them it is not so convenient to draw the proper statistical inference.

So, that is why for the sake of understanding and for the sake of better illustration usually as an experimenter we always try to create the bars of the equal width so that it is easier for us to draw inference from there. But the width of the bars need not necessarily to be the same. In case if you want to create the histogram in the R software then the command here is h i s t and inside the parenthesis you have to give the data and this will create the histogram with respect to the absolute frequencies. Now in case if you want to create the histogram with the relative frequencies then you have to use here an option frequency is equal to f which is written here as say freq all in lowercase is equal to false. So, if you use this option then it will not use the absolute frequency but it will use the relative frequencies to create the bars.

And then in case if you want to make it more informative means my suggestion is that please look into the help menu you will get many options there so that you can control different parameters of the histogram. Some of them are like as mean you know that mean is for the title of the chart, col color for cutting the colors of the bars, xlab is for the description on the x axis, xlim and ylim they are trying to specify the range of values on x and y axis respectively. But I will suggest you please try to look into the help and try to understand what are the different parameters and you try to make these graphics more informative. Now let me try to give you here an example so that I can illustrate that it is not difficult at all to create the histogram. So, now I have collected the data of 50% and their heights in centimeters are recorded as it follows and this data has been stored here in a data vector height.

Now if I try to create here the histogram by using the command hist height it will here look like this. So, this height is coming by default and this title histogram of height that is also coming by default. On the y axis it is trying to create the, it is trying to use the absolute frequency. So, you can see here now these bars the width of the bars this is here the same. But their heights are varying and heights are varying according to the absolute frequency and you can see here this is the point here 125.

So, it is trying to say that there are 5 values which are in the interval of 120 to 125. Similarly, if you try to see here this is the bar, this is here the bar let me choose different color, this is here the bar, this is here the bar and these are the points here 145, 155. Right, so it is trying to say that the absolute frequency or the number of persons having the height between 145 and 150 centimeter as well as 150 centimeter and 155 centimeter this is here the same and there are 5 persons who have got such heights. So, this is how we try to take interpretation from the histogram right. Now in case if you want to make it here more informative for example if you want to give here the title here as a height of

person then we are going to use the same command main, main is equal to heights of persons.

And if you want to give it the color green then you have to simply use here col is equal to green right. Here I am just using green in the double quote because there is a single color. If there are more than one color then I have to use the different colors combined with the C command. Now in case if you want to give here the label on the X axis then I use here the command xlab is equal to heights in the inside the double quote and if you want to give on the Y axis the title number of persons then I have to use here ylab and it is here number of persons right. So, this is how we try to do it.

Now let me try to show you this graphics first in the R software so that you gain more confidence. So, let me try to copy this data on the height here and let me try to close the earlier diagram. So, you can see here this is here the data on the height. Now if I say here hist of height you can see here it will here come like this right. Now in case if you want to make it here more beautiful you want to give here some information.

So, I can use the same command here and you can see here it will give you the histogram in green color and this is the same histogram which I have just illustrated here in the slides. So, now you can be confident that how you can create the histogram. Now we gradually try to move to a multivariate case. Now in case if you try to think about histogram these histograms are trying to indicate the distribution of the frequencies or distribution of the values which are partitioned into different intervals. So, one histogram is trying to take care of one dataset.

Now if you want to compare more than one dataset that can be 2, that can be 3 or any number. Then do not you think that if you can create the histogram of different variables in the same chart then it will be easier for you. For example, if I say the total the marks of the students in class 10 in say 20 colleges right the marks are suppose going to be between 0 and 100. So, this is a continuous variable and now you have got here 20 colleges. So, now you are going to have the dataset 1, dataset 2 up to here dataset say 20 for different 20 colleges and here there are the marks of the students in that college.

Now if you want to see the distribution of the marks then one option is that you try to create these 20 histograms on different sheets. But if you can make here these 20 histograms on the same sheet with the same scale of x and here y don't you think that it is going to be really helpful. So, this is called as multiple histograms. The multiple histograms are created on the same graph. So, what we have to do we have to use the same command hist but I have to add here an option add is equal to true a double d that is all-in lower-case alphabet is equal to true and this helps in comparison.

The point where you have to be careful and you have to pay your attention is that when you are trying to set the minimum and maximum values of the entire datasets to set the

range on the x axis right. Because you see well, I have taken here this example in which there are marks which are given out of 100. So, the marks are going to be between 0 to 100 but suppose one college is giving the marks between say between 0 and 20 somebody is giving trying to mark between 0 and 40 and another college is giving say between 0 and 100. So, in this case you have to be careful that when you are trying to choose the range on the x axis to plot the multiple histograms then you have to choose 0 and 100 whatever is the minimum value of all the data set and the maximum value of all the data set with respect to the range on the x axis right. So, now I try to take here the similar example.

So, the height and weights of 50% they are recorded here as follows. These are here the heights and these are here the weights of 50%. Now I want to create the histogram of height and weight on the same graph right. So, I try to store this data in a data vector here height and then I try to store the data here weight in this data vector right. Now we try to use here the command hist and first we try to plot the data of height and I want to make it in blue color and range on the x axis which I want to give.

This is here from 20 to 180 right. So, you can see here this is here the data on here see here height. This is blue in color and the range of here this 20 to 180 this is here somewhere it is here 20 and somewhere it is 180. And the title this multiple histogram this is appearing here mean is equal to multiple histograms and then this here x is appearing by using the command here xlab is equal to x. So, that is giving you the title on the x axis. Now you want to also add here the histogram for the weight.

So, I try to use here the same command hist of weight data vector and color is equal to green in double quotes and then I try to use here the command add is equal to true. So, now once you execute this command the second command then this graph will appear here right. And you can see here this is green in color and this is denoting the distribution frequency distribution of the data on weights right. And first this blue histogram will appear when you try to execute the first command that means this one and once you then you try to compute or execute the second command then this histogram will appear. I will try to show you it on the R console so that it becomes here more clear right.

So, let me try to keep or copy here the data on height and weight. So, let me try to clear it. So, this is here height and this is now here weight. So, you can see here this is height and this is here weight right. And yeah, if you try to make it the histogram of your height this will here look like this and if you want to make here the histogram of your weight this will look like here this.

But now I want to make both the histograms on the same sheet. So, I try to you can see here first I am trying to copy here only the first command right. So, that you can see how I am doing it. So, now let me remove this thing and now if I say here this command you

can see here only the first histogram on height is created. And now if you try to see here I try to use here the second command.

The command is the same only I have to add here means add is equal to true right. And suppose if I want to add here one more graphic in the sheet right. Then what I have to do I have to take the data I have to change here the name of the variable and then I have to just add here add is equal to true that is all. Suppose another data is suppose here age. If I see here histogram h i s t of age, color is equal to suppose yellow or red and but the main important point is that you try to add here the command add is equal to true.

So, this will try to create the histogram in the same dataset right. So, now you can see if you want to compare these two graphics this is pretty simple right and it can be done very easily right. Suppose if I try to use this graphic, I can show you here that how you can do it I can increase the size of this graphic. And now if you try to see here it is not difficult to use this because one concept which I would like to give you here and later on I will try to explain it when we are trying to consider the concept of probability density function that these histograms are trying to give us the idea about the distribution of the frequency or distribution of the data. So, if you try to take here the mid values of each of this bar like this here like this and you try to create here a smooth curve.

Smooth curve means you try to keep your pen and then start joining all the points without lifting the pen. So, if I try to do it here it will look like this, somewhere here like this. It is a smooth curve so now means you can also ask that okay I should make here like this but anyway this is only a smooth curve. And similarly, if I try to say here it will be here like this, like this. So, this type of curve and if I make here a smoother curve possibly, I can make it in this way for example I can make it here like this one also, this thing also.

Yeah, it is not very nice, let me try to try again like this one, right. So, this program you have to keep in mind and when I will be using or introducing the concept of probability density function at that time, I will request you to recall it. And these are here the bins, this is here bin, this is here bin. So, all this, this is here a bin. And if you try to think that if I try to make the width of the bin to be very small and then if I try to join all these midpoints of the bin then ultimately it will look like a smooth curve.

So, that will be representing the probability density function. So, that is the relationship but anyway I will try to take it in a different lecture. But anyway, if you try to compare here these two curves over here you can have a fair idea that okay that the number of, when the frequency of height and weight up to 2 it is nearly the same, right. But the number of people who are having a frequency here 5 here like this. So, yeah, so different types of conclusions can be taken here and it depends on your capability and the questions which you are going to answer, right.

So, now we come to an end to this lecture and now you can see here we had considered here this concept of histogram and we have tried to extend it to a multivariate case. Multivariate means more than one, more than one variable. So, now you can see this is one way where you can extend the concept of univariate histogram to a multiple histogram. The idea is that you want to consider more than one histogram at the same time so that you can get a correct outcome.

So, this is one way out. I have taken here an example of only two data set but surely you can take multiple data sets and you can give them different types of labels, titles, colours to make them more informative. But again, it depends on your aptitude and your experience that how you can make this these histograms more informative and how you can present them so that you get a good outcome. And it is also possible to combine different graphics in the same graph. For example, if you want to combine the pie chart, histogram as well as these multiple histograms or say some other graphic in the same plot, then we have another command here mfrow. Well, I am not going to discuss here in this lecture but I am telling you that it is possible by using the parameters mfrow, and then you can combine different types of graphics in the same means one single graphic.

So, that is some time needed when you try to prepare the reports and for that you want some graphics in a particular way. So, those graphics can be exactly arranged as you try to arrange the numbers in the matrix. There can be a 4 by 4 graphic, there can be a 2 by 3 graphics whatever you want. So, those things are possible but they are not actually multivariate cases because they are trying to combine different graphics together but here what we are trying to do the same graphics which is on different variables of the same type that we are trying to plot in the same scale also. So, that is why I am calling them in sort of multivariate in nature.

As we are doing it here different multivariate procedures, so this is about the graphical procedures right. But in these cases, you can see that the limitation is that you can handle one variable at a time in the subplot or say bars subplot. You can have the division of the observation within the one category. But now suppose you have more than one variable and you want to express them in a single chart, how are you going to do it right.

So, that is another issue. But at this moment I would request you that you try to take different datasets and try to create this type of graphics and try to practice how you can retrieve the important information hidden inside the data through these graphics. And I will try to explain you those graphics where you can consider more than two variables and you can create beautiful graphics in a compact way in the further lectures. So, you try to practice it and I will see you in the next lecture till then good bye.