Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 04

Lecture – 15

Boxplots and Grouped Boxplots

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we considered the command summary and, in that command, we wanted to compare more than one dataset with respect to different type of characteristic like as mean, minimum value, maximum value, median etc. Now when you have more than two datasets, yes summary command is going to be very useful but do you think if such values can be expressed in a graphical way, why graphical way? We understand nowadays that if I want to express my emotions and if I am very happy, I do not write a very long sentence to express my happiness but I simply send an emoji like smiling face. And if I simply want to laugh at some comment or something, some statement then once again I try to send emoji of a laughing face and that conveys more than what we can convey from our language. So, graphical ways are also important. So, now gradually I will also be entertaining the different types of graphics.

So, first I will try to take some popular graphics and then I will try to extend them to a multivariate setup but as I said because of the limitation on my slides I am not able to take 20 or 30 datasets but the way I am going to explain you is that if something is happening with two or three datasets that can be extended to any number of datasets. So, that is how I am trying to handle this multivariate situation. So, now in this lecture we are going to consider the aspect of box plots. These box plots are simply the representation of the outcome of the summary command in a graphical way.

I am not saying at all that the graphicals are graphics are only useful, no. This will give you a fair idea but if you want to get more accurate idea then you have to go to the analytical tools. So, if you think suppose you have the outcome of summary command and also you have the outcome of box plots, the graphical presentation of the summary command. Do not you think that it is going to be very helpful? So, with this objective let

us try to understand the concept of box plot in this lecture. Okay so let me begin our lecture now.

So, we are going to talk about here the box plots and group box plots. So, we have understood in the last lecture that if I try to use the summary command over a data vector x then it gives us the information upon the minimum value, maximum value, mean value, first quartile, second quartile and third quartile. So, essentially it is trying to give us the spread of the data. But in the sense for example if I say if this is the frequency curve of a data then here is somewhere the minimum value, here is somewhere the maximum value, then here is somewhere here is the mean value. If the data is symmetric then the mean and median may be close and then here is the first quartile, second quartile, third quartile, fourth quartile.

So, if you get this much of information you are practically getting sufficient information to understand the behaviour of the datasets or you can get a fair idea about the hidden tendencies which are inside the dataset. So, just for the quick revision in the last lecture we had considered this example where the time of 20 participants in completing a race are measured in second and they are reported in the time vector and then we had obtained the outcome of the summary command over the time. We have minimum value, first quartile, median, mean, third quartile and maximum. And then we had created another dataset where there are two values which are very high. So, in the same dataset I had made these two values to be very high.

And after that by creating a data frame of time and time 1, time is the original dataset and time 1 is the dataset where I have made first two values to be very high. I tried to compute here the summary and we had got this type of information and we had discussed how we are going to conclude. Now the same command, same outcome can be presented graphically using the concept of box plot. So, box plot is a graph which summarizes the distribution of a variable by using the median, quartile, minimum and maximum value. And it is very helpful when we are trying to compare different datasets.

For example, the box plot looks here like this, this type of picture. So, if you try to look into the bottom value, this line, this is indicating here a minimum. And then there will be here this type of here box. And then there is going to be a central line here and then there will be a top line here and then these values are going to give you different types of information. So, I will try to explain you one by one.

So, basically, we have these are the values which are important where I am trying to highlight it. Now let me try to take it here one by one. So, as I said the lowest value that is indicating the minimum value. And then this line, this is here the value of the first quartile. Then we have the line in the middle which is indicating the second quartile, that is the median.

And then we have here this line which is indicating the value of the third quartile. And above of them here is this line which is indicating the maximum value. So, well these lines which I am now going to highlight in yellow colour, they are actually not giving any information but yeah. But so basically, we have to look vertically but this picture is created in the format of a box. So, that is why this is called as a box plot.

So, if you want to draw this box plot in R software then the command here is box plot. And inside the parenthesis you have to give the variable. And then if you try to go to the help menu of this box plot in the R software then there are several options which are available here. So, let me try to give you here some idea how you can do it. So, if I try to take here the same data set here time which I have taken earlier.

If I try to make here the box plot of here time it will look like this. So, this is now here the minimum value which you can see here minimum. And this is here the first quartile or Q1. This is here the median. Second quartile which is somewhere here between close to 55. Then we have here this Q3 somewhere here and then this is here the maximum value here. So, if you try to see whatever are the values suppose if you try to see minimum is here close to 30 and the first quartile is close to 40 that you can compare from here this value. The minimum value is 32 and the first quartile is 41. Median is 56.5 and mean is 56 and third quartile is 68.

So, that you can see from here also that yeah second quartile here is something close to 56. It may be difficult to find out exact value by inspection it is 55 or 56 but from the numerical value you will get the exact value and similarly if you try to see this third quartile is close to 68 but at least you can see that it is very close to 70 and similarly here is the maximum value. So, this is how we get a graphical view of the outcome of the summary command. Now if I try to take the second data set where I had made the first two observation to be quite high for example you can see here first two observations are increased from 32 to 320 and 35 to 350 respectively. Now if you try to create the box plot of the same data what happen you can see here this is here like this.

The maximum values are lying here. These are the extreme values which are lying here and your box plot is somewhat here. You can see from this box you are getting the first idea that if there is any value which is very much different which is very much away from rest of the data box plot can identify it very easily. Just by looking at the box plot I can see that here there are two values which are around 320, 350 and this is captured by this box plot and remaining values you can see here they are between say here somewhere here like this between 30 and say 80. So, now if you think that these box plots are created and if they are created side by side then they can give you an idea of the relative behaviour of the data set and this box plots have different types of advantages.

For example, if you come here and if you try to see here suppose let me clear the screen so that I can explain you easily. If you try to see here this distance these are minimum and maximum value and this is simply here the statistical range and if you try to see here this difference between this and this this is simply giving you an idea about the quartile deviation or say interquartile range. If you try to see this distance then possibly it gives you the idea about the symmetricity. If the frequency distribution is symmetric then ideally this distance and this distance should be the same that half of the values below the median value and half of the values are above median value. So, this type of different types of conclusions you can draw from such graphics very easily.

For example, if I try to plot both the graphics inside the same one although here, I have not plotted it jointly but I have just copied and pasted both the graphics side by side just to give you a fair idea because you have to be careful you cannot compare them because the scales on the Y axis they are different. Here there are 30 to 80 whereas here there are 50 to 350. But still if you try to make such graphics on the same scales of Y then possibly it is very easier to compare them. So, now how to do it? So, in order to create such box plots of different data sets together we can use the concept of data frame and as such box plots will be called as group box plots. As the name suggests that this is a group of box plots.

So, in order to create group of box plots that means more than one box plots of different data sets but on the same scale. So, in order to do it first I need to create the data or I have to convert my data sets into the framework of a data frame. So, I know that if I have three data vectors X, Y, Z then using the command data.frame then inside the parenthesis X, Y, Z I can create here a data frame. Now using this data frame, you can create the box plot where essentially you have to give the you have to replace the data vector X by the data frame.

So, if I write down here box plot same command and inside the parenthesis you write the data frame. Then you will see the group box plot of X, Y and Z. So, just to give you an idea here so if I try to take here the same data set time and time one where these two values 32 and 35 they are changed to 320 and 355 all other values are remaining the same and if I try to create here a data frame first such that data frame of time and time one and then if I try to use this name here I have given it here a name data box plot and if I say here box plot inside the parenthesis data box plot you can see here now I can compute the group box plot where the scale on the Y axis is the same for all the data sets and now you can compare them very easily. For example, if you want to find out that out of these two data sets does any data sets have extreme observation or say outliers and you do not want to look into the data set. For example, here I can show you very easily because we have a small data set and this data sets have been artificially created so I know that 32 and 35 are extended or converted into 320 and 350 but if you think that you have a data set of

thousands of observations or millions of observations can you really look at them visually possibly not.

In such cases such a statistical procedure like as box plot helps and in case if I try to compare these two box plot these two points here they are indicating that based on certain concepts to find out the outliers in the data in statistics they have been plotted here and if you try to see the R software or the concept of box plot has not considered them as the part of the data. If so this maximum value would have been gone to this place. So, that is the beauty of statistics that if you try to use the correct tool at the correct place in the right type of data then you get the correct information. So, by looking at this group box plot you can see very clearly that these two points are possibly the outliers. Now that is your decision that how you want to handle these outliers whether you want to remove from the data or you want to use some robust statistics whether they are the part of the experiment or they are not the part of the experiment but they are coming but they are entering into the data without any problem because of any mistake etc.

These different types of conclusions can be taken by you. So, if you try to see here this is the screenshot I am taking a time and time 1 and this is here the box plot of the data set and then the data frame creation and then the box plot. So, let me try to create this data set here and then try to show you how you can create this box plot. So, let me try to copy the numerator part so that I can save some time. So, if you try to see this data of say time and time 1 this is already because we have just used it and now if I try to say here box plot of time you can see here it is here.

Now in case if I try to create here the box plot of here time 1 you can see here this how this picture will change. So, I make a time 1 you can see here now the graphic is changed but now if I try to create here the data frame of time and time 1 then if I want to create here now the box plot of this data frame you can see here this is here now like this is here the two box plots and these two points here where I am trying to move my cursor these are the outliers. So, you can see here that it is not difficult and you can imagine that if you have more than two data sets and if you want to compare them possibly these types of box plots will be simultaneously here and then you can compare them very easily. So, this is one way by which we try to create the graphics for the multivariate data. We try to consider this as a univariate data and we try to create one box plot for each of the data and we try to compare them together.

Now let me try to give you here one more example to illustrate the utility of this box plot. Suppose the marks of 10 students in two different examinations are obtained and we want to compare them using the box plots. So, these are the marks 1 and marks 2 of these marks and now I want to create here a data frame. So, I try to create a data frame here by writing the command data.frame inside the parenthesis marks1, marks2 and it is stored in data marks.

So, now if you try to see this is the screenshot now if you try to create here a grouped box plot so you can see here this is here marks1, this is here marks2. Now you can compare here very easily that the minimum marks in the set of data of marks2 is much much lower than the data set of marks1. The maximum marks obtained in both the data sets that is nearly the same and the distribution of the marks in the two data set is like that that more number of students in the class have got higher marks in the data set of marks2 than in the data set of marks1. It is also seen by looking at this distance that in the marks2 most of the students they are almost equally spreaded with respect to quartiles whereas in the case of the marks1 this distance is different. So, I can say that the data in the marks1 is more skewed whereas the data in the marks2 is more or less symmetric around the median.

So, this type of different type of information you can create here and let me try to show you these things on the R console also so that you can understand them very easily. So, let me try to create here the two data set. Say here marks1 and now here marks2 and then I have here data marks and then I have to create here box plot of data marks. You can see here it is here like this. So, now we come to an end to this lecture and you can see here as promised in the last lecture we have now demonstrated the graphical way of presenting the information in the summary command.

And you can see here that when you are trying to consider the multivariate data then this summary command or this box plot, they are going to be very helpful. You try to create all the box plots side by side and it will give you a broad spectrum of the properties of the data in different data sets. Well, I have taken here very basic formation of this box plots. It is possible to change the color, it is possible to write something on the axis, something inside the box etc. and various types of formation or various types of parameter control are available in the creation of box plot.

Well, my objective is here to give you the basic fundamentals. Now if you want to make your graphic more beautiful you can look into the help command and try to experiment with those options so that you can make these graphics more beautiful and you can incorporate different types of information inside the plots so that they become more informative. So, now my request once again is that you try to create data set and then try to introduce different types of issues inside those data sets. For example, suppose you have got a data sets of 100 values. Now try to make 10 values very high then you try to make 20 values very high and try to see the box plot of the 3 data set, the usual one and then these 2 additional data sets where you have increased the number of extreme values.

Try to see how the bars of the box plot fluctuate. Similarly, you can do that try to reduce some values and then try to make them away from the center and then try to see what is happening in the box plots. So, when you try to look at the box plot and then you can compare that the problem or the disturbance which you created inside the data sets that

has been captured in the box plot in this particular way and the data sets looks like this. This process is similar to as I took the example of the medical doctor in the last lecture similarly I can say that if you try to see doctors look at the different types of graphic either that is ECG or EEG or different types of pictures of X-ray and from those X-rays they try to decide what is happening inside the body. They do not just open the body and try to see what is happening inside.

So, similar is the training what you have to get that by looking at different types of graphics you can judge what is happening inside the data. The more you practice your judgment will become better and you will be able to diagnose the problem more accurately and more correctly just like an experienced doctor. So, my request to you once again is that you please try to take different data set practice and try to improve your skills as data scientist. So, you try to practice and I will see you in the next lecture till then goodbye. Thank you.