Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 03

Lecture – 13

Measuring Variation with Missing Data

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we had considered the different measures of central tendency when we have missing data. Now continuing on the same lines in this lecture we are going to consider different measures of variation. So, the way you have seen that we had modify our median commands to handle the data vector which has got a missing value. Similar concept is going to be implemented in different types of measures of variation because may be range, this may be interquartile range, this may be variance, standard deviation or say absolute mean deviation right. So, exactly on the same concept the way we had developed the measures of central tendency in the last lecture on the same lines I will continue with the different measures of variation in this lecture.

So, let us try to begin our lecture right. So, now I am going to consider different types of measures what we had discussed in an earlier lecture on measures of variation and my objective is very simple that earlier I had explained you the basic concept. Now I will simply explain you that how those concepts and commands are going to be modified for the missing data vector. How to handle the missing data vectors that we already have understood? So, now in this lecture basically I have to explain you how are you going to modify your basic commands and functions to handle the missing values.

So, let me try to take different types of measures one by one and I will try to explain you. So, let me try to first take the concept of statistical range. So, now you know that the statistical range is defined as the difference between the maximum and minimum values. So, now if you have a data vector X, so in order to find out the range or the sets of range we use the command here max(X) minus min(X) right. And I would like to have your attention that if you are trying to use the command here range R A and G then it is also is value the statistical function in the R software, but it will give you only the minimum value and maximum value in the form of an interval.

But whereas we are interested in finding out the difference of the maximum in and minimum value. So, that is why you have to go like this. Now suppose if there is a data vector which got missing values, suppose it is denoted by X and A which has got some values which are here capital and capital N A. Then the max and min command they are going to be modified here like this. So, you can see here max command is now for the data vector which has missing value X and A and then I have to use here the command N A dot R m is equal to true right.

And when I am talking about the command min which is finding out the minimum among the values in the data vector, then the data vector is and then na.rm = TRUE inside the parenthesis. So, you can see here basically we have learnt that how to modify the minimum and maximum command when there are missing values and in turn this will give us the value of the statistical range. And as I said the command this range R A N G E all-in lower-case alphabets, this returns a vector containing the minimum and maximum values all the given values inside the data vector. So, I will try to take here the same example that we considered in the last lecture where we had recorded the time taken by 20 participants in completing a race in seconds and right and this data vector is stored in now here the data vector time and yeah you get out on I will try to remove these two values here and I will try to create a data vector with missing values. So, now you can see here if you try to find out the statistical range by the complete for the complete data set it can be obtained by max(time) – min(time) which is coming out to be 52.

Now as I said if you somehow use the command here range then range of time will give you the value here 32 and 84. So, you can see here 32 here is the minimum value and 84 here is the maximum value. So, it is indicating you only the two values, but this is what we do not want right. So, be careful right. This is the screenshot of the same command here right.

Now if I try to modify this test set and if I make two values to be N A the first two values are here not available. So, now this is my here new data vector which I have given the name time dot N A right. Now if you try to find out here the statistical range of maximum of time dot N A minus minimum of time dot N A it will come out to be here N A right and that was expected, but on the other hand if you try to find out here the maximum of time dot N A with N A dot R M is equal to true minus minimum of time dot N A with N A dot R M is equal to true. So, basically you have to add here this command this N A dot R M is equal to true and it will give you here the value 49 right. So, you can see here this is the screenshot here and now let me try to show you these commands on the R console also.

So, I can just create here two data vectors see here time and then I will try to create here this data vector here see here time dot N A right. So, this is your here time and this is your here time dot N A right. So, if you try to find out here the max of time minus min of time you can see here this is coming out to be 52, but if you try to find out here the maximum of time dot N A minus here this minimum of time dot N A right then this is going to be here like this you can see N A. But on the other hand if you try to add here the command here time dot N A dot R M is equal to true and here also time dot N A with N A dot R M is equal to true then it is coming out to be here 49. So, you can see here this is how you can find out here the statistical range right.

So, now we come back to our slides and try to understand the further things. Now you can recall that earlier we also had discussed a concept of interquartile range which was the different between the 75th and 25th percentiles or equivalently I can say the third and first quartiles. So, this was defined as I Q R is equal to third quartile Q 3 minus first quartile Q 1 and this covers basically the central 50 percent of the distribution of the observation. For example, if you have a something data set which can be represented like this. So, this is your here first quartile this is your here and the second quartile and these two are the values in the.

So, Q 2 is indicating this value and Q 2 is indicating the data set in this observation. So, this a quartile interquartile range is indicating the spread of the data in the second and third quartile right. So, this is because every quartile is trying to divide the total frequency of the data into four equal parts. So, every quartile contains 25 percent of the observations and so this 2 2 and 2 3 combined together they contain 50 percent of the observation right. And as we had discussed that any data set which has a higher value of interquartile range is said to have higher variability and obviously for the decision making a lower value of interquartile range is preferable.

For example, if you have two data sets and suppose their interquartile ranges are coming out to be I R 1 and say here I R 2 right. If this interquartile range of first data set is greater than the data set 2 then this I R 1 means the data set in for which we have measured the interquartile range is said to have higher variability or more variability than the data in the second data set right. So, now we had also understood that the R command to compute the interquartile range is IQR(X), I Q R all in uppercase alphabet and X is the data vector. Now, in case if X data vector has missing values and suppose the new data vector is indicated to say here X say X and A then the R command of I Q R is modified like this the command will remain the same I Q R and the data vector will now change which has missing values and now we are going to use here the option N A dot R M is equal to true right. So, and similarly if you try to recall the definition about the quartile deviation it was simply the half difference between the 75 and 25 percentile or equivalently the half difference between the third and first quartile.

So, essentially it is only the half of the interquartile range and quartile deviation is defined only here as say in a half of the interquartile range that is I Q R divided by 2 which is equivalent to Q3 minus Q1 divided by 2. And in order to make the decision making the we had understood that any data set having a higher value of quartile deviation is supposed to having more variability higher variability. So, this is the same concept that we considered earlier. Now, if you recall that earlier for the complete case data set that means, there are no missing values the we had considered the command I Q R of data vector X divided by 2. Now, you know the way to be modified it is very simple you already have understood how are you going to modify your I Q R command where you are trying to write I Q R in the upper case alphabets X N A and you are adding here the option N A dot R M is equal to true and it has to be divided by 2.

So, this way you can very easily compute or modify the expression for the quartile deviation from complete case to missing data case right. So, for example, if I try to take here the same example of this here time then the inter quartile range of time comes out to be a 27 and if you try to find out the inter quartile range of time it comes out to be 27 divided by 2 this is 13.5. Now in case if you try to take here the missing values right. So, I try to take the same data set as I did earlier and the first two values are converted into N A and now this is my data set here time dot N A all the data values are the same except the first two value they are now N A and in that case if you try to find out there the value of here I Q R time dot N A it gives you here an error that it cannot compute it.

But if you try to make it here I Q R time dot N A and N A dot R M is equal to true then it gives you here the value 25.25 and if you try to find out the quartile deviation which is the half of I Q R which is here I Q R time dot N A N A dot R M is equal to 2 divided by 2 which is here half of 25.25 which is 12.625. And yeah so let me try to first show you these things in the R console.

So, you can see here we already have our data set. So, basically I will show you here that you can see here this is your here data set time and this is your here data set time dot N A. Now, if you try to find out here inter quartile range of here time it will be here like this and the quartile deviation is I Q R time divided by 2 like this. Now, in case if you try to find out the I Q R of time dot N A it will come out to be error it is not computing it right. So, it is not even giving you the value N A which was happening in the case of mean and median.

So, that is what I was trying to tell you that N A dot R M does not work always, but then you have to see that how it is going to work. So, the idea which I am going to give you is that if you want to compute any particular quartile or percentile with missing values although I am not covering here, but you please look into the help menu and try to see how the missing values have to be handled right. But anyway in our case if I want to find out the inter quartile range with this missing value then I simply have to give here the value the option N A dot R M is equal to true or T and it gives me the value 25.25 and similarly if I want to find out the quartile deviation which is half of the inter quartile range it is 12.625 right.

So, now I hope this is clear and now I come to another concept of absolute mean deviation which we had considered in the earlier lecture. So, we have thus the data vector here x and if you recall that what we had done in order to find out the this absolute mean deviation what can we call that if I have the deviations xi minus x bar that every observations deviation from its central value right. Like as for example, if I say this is my here data set it is the scatter diagram somewhere it is here the mean value and now, I am trying to suppose this is my ended data point say xi. So, I am trying to find out here x i minus x bar, but now xi minus x bar can be positive can be negative. So, one option is this I can take its absolute value or I can square it.

So, when we are trying to take the absolute value then based on that we try to define the absolute mean deviation and when we are trying to take the square then we try to define the variance concept right. So, firstly let me try to handle the absolute mean deviation with complete case right. So, we know that how we are trying to define this absolute mean deviation this is something like here 1 upon n summation i goes from 1 to here n absolute value of here x i minus a right a is some value around which we want to measure the variation. So, a can be here sample mean, arithmetic mean say here x bar, a can be here median also or it can be any arbitrary number. But here I would like to inform you that when we are trying to measure this mean deviation this is going to be minimum when a is sample median.

That means, instead of computing this $\frac{1}{n}\sum_{i=1}^{n}|x_i - a|$ around any arbitrary point if you try to choose a to be the median of the observation then this is going to give you the minimum value. On the other hand, yes just for you the sake of your information if I try to take here the squared value like here $\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2$ then this is going to be minimum when a is equal to sample mean x bar. So, that is why in the case of variance we always try to take the x bar. Now, in the last lecture when we consider or in the earlier lecture when we consider the concept of absolute mean deviation then we had measured it around mean. That is your choice from where you want to measure it.

But as I told you that this is going to provide the minimum value when it is measured around median. So, just to give you this idea also in this lecture I am considering the mean deviation around median. But you know that this expression is now very simple $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x_{medıan}}|$ which where x bar median is the median of $x_1, x_2, \dots\ x_n$ right x bar median is the median of say $x_1, x_2, \dots\ x_n$. So, this I am trying to deviate from my earlier lecture, but I wanted to give you this concept also that is why I have taken it here differently. Now, there is no built-in function for finding out this mean deviation, but you know that it can be written very easily as we did in the last time.

We are simply trying to find out here the arithmetic mean of the absolute values of x i minus x bar median. So, x bar median can be computed by here the command here median and those absolute value can be found by using the function abs it will give you absolute values. And now if you try to find out the 1 upon n summation of these absolute value then you have to use here the function mean. So, by writing this command you can very easily find out the absolute mean deviation around median for the complete case. Now in case if you have some missing value then now you have to see that how are you going to modify it.

You can say if I try to illustrate it here you can see here where you have to change the symbols and notation or where are you going to modify it. This x is going to change because it will have now n a values. The way you are trying to compute this median this is going to be changed because now it has missing values. The mean the way you are trying to compute this will also be changed because now it has missing value. So, the way you have learned how to change the mean how to change the median and then you also have this absolute function which also need to be modified because you have here n a value.

So, now all this concept which you have learned in the past will help you in writing the modified commands when the data vector has missing values and you want to find out the absolute mean deviation. So, if the data vector is now given as say here x n a our usual symbol notation which has missing value. So, now x becomes here x n a. Now this median if you try to see I am using here this command here n a dot R m equal to true. Now you have used here the absolute value of this thing absolute value is absolute value because now you are trying to remove the n a values.

So, it will remain as such, but now you are trying to find out here this mean. So, you know that in order to find out the mean with the missing values you have to give here the option n a dot R m is equal to true and so now this expression will give you the absolute mean deviation around median when we have missing values right. So, let me try to try to explain this concept on the same data set if I have a data set on here time. So, you can see here this is my data set I try to find out here the median and then this here like this 50.5 and if I try to find out the absolute mean deviation around median this comes out to be here 14.5.

Now on the other hand yeah this is the screenshot of the same operation I will try to show you on the R console also right. Now on the other hand if you want to do it with the missing data also you can do it easily right. So, let me try to show you this things on the R console it is a better option for me. So, now if you try to see here I already have the data of here this one. So, let me try to write down first here this the complete data case.

So, now I have to just modify using you see here this is your here time and this is your here time.na. If you try to use here this thing this is x because now I had just copied and pasted here the command x, but if you try to see you had taken x to be earlier as here like this. So, I will try to modify it. So, just be careful when you are trying to copy and paste the commands you have to choose here the right command here time and then here see here time. Now if you see here you will you are getting here the value here 14.5, but on the other hand if you try to make it here time.na the data vector which has missing value you will see here it will give you here the n a right.

But on the other hand if you try to write down the command here for here missing values like this for example, if you want me to write this here command I can write down here you can see here control c and then I try to modify here this command let me try to put here time.na you can see here how I am doing it right. Let us say time dot n a and then this is na.rm is equal to true. And now if you try to come copy and paste this command over here this will come out to be here like this right. So, now this is indicating here the absolute mean deviation when we the missing value this plus sign because now this command is coming on the next line.

So, do not get worry do not get confused right right ok. So, now let me come to our next topic where I would like to give you this idea. So, now the next command here is variance which is the popular command to find out the variation in the data. So, we we know that when we have the complete data set then the variance is computed by here the function v a r. And we know that v a r is computing this variance as $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

And this is this is what actually this v a r command in r is computing. Whereas you can recall that I had discussed earlier that we have one more a command for the variance which has a divisor 1 over n. And if you want to compute it then you can compute it by this command that we have that we had discussed when we were trying to discuss the concept of variance that I am trying to write down here this here as n minus 1 upon n minus 1 $\frac{n-1}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$. And now I will write down here $\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. And now this quantity can be computed by variance and now this is how you can write down this command here where n is the length of data vector x.

And I have explained you that why we have these two formulae of command, but in r you have to remember that it is computing the variance with a divisor n minus 1. And in case if you want to find out the standard error or say standard deviation then you can you simply have to take the square root of the variance command. If you want to have a divisor 1 upon n minus 1 then use this square root of VARX and if you want to have divisor 1 upon n then you simply have to find out the square root of the variance which of that expression which you have just obtained here like this one right. Now in case if the data vector x has some missing values suppose and this data vector is indicated by here now see here x N A. So, in that case the variance command is modified here as VAR data vector x N A and then N A dot R m is equal to the same approach what we have used multiple times.

And if you want to find out the standard deviation or standard error with the data vector which has missing value then you simply have to take the square root of this variance which I can obtain by here this value of variance command square root of variance of x N A with N A dot R m is equal to true right. So, let me try to take here the example to explain you. I try to consider here the same data time. So, you can see here this is your time and if I try to find out here the variance of time it will give me this value which has here the divisor n minus 1 and if I try to find out the standard error or standard deviation this value here is 16.83355 which is obtained after taking the function or using the function square root over the variance of time.

Now, in case if you want to have the variance which has divisor n then you can you know that this is how you can modify this command over here and based on that you can also find out the standard error or standard deviation right. So, these are the screenshot of the same operation they are straight forward. So, I will not go into these details, but I, but you can just do it. Now, my thing is this now I am trying to operate this thing when we have the missing data. So, now, I try to take the same data set which has got here with this here N A and N A which are the first two observation and my data vector here is time dot N A right yeah.

So, now, we try to compute the variance and standard deviation or standard error with the missing data vector. So, now, I have here this command this data vector here time dot N A and now I am using here the command N A dot R m is equal to true. So, which is giving me here this value and if I try to find out its standard deviation or standard error this is here like this. So, if you try to see this value here is 250 and, but when you have the complete the data set it was this value was 269.2.

So, this is going to change and that is why it is important to have it right. Now, I would like to have your attention. In case if you want to find out the variance with divisor n where you have a missing data then in case if you try to modify your R command then there are two components one is here the length of the data vector and another is the variance of data vector how are you going to do it. So, variance of data vector that can be done as we have done now earlier, but for the length this N A dot R m will not work, but now you have to use the command here N A dot omit and the data vector. If you recall N A dot omit we had discussed when we explained the use of or when we learned how to handle the missing data values.

So, if you in place of here N in case if you use this command here length N A dot omit and inside the parenthesis data vector then your this expression will be modified and if you try to simply replace by this thing and you modified here the expression this variance will come out to be here like this and it is standard deviation or standard error that is just by finding out the square root of this that you can find out very easily right. So, this is what I want to show you here that if you try to use here the length of this time dot N A with N A dot r m is equal to true then it is giving you here error it will not work, but in case if you try to use here length with N A dot omit then that is 20 minus 2 there are 18 observation and this is here 18 observation. So, this is how we are going to work on it. Now let me try to show you these things on the on the R software.

So, let me try to copy this functions over there. So, that I do not make any mistake in the in copying it right ok. So, if you try to see in the R console now let me try to clear the screen you have here the data vector time and then you have here the data vector time dot N A. Now in case if you want to find out the variance of time you can very easily find out, but if you want to find out the variance of time dot N A you will see here it is coming out to be here N A. Now similarly if you want to find out the square root of the variance of time you can find out very easily you can see here, but if you want to find out the variance of time dot N A you will see here it will give you the error. So, now if you want to find out the variance of this missing data vector then you have to write down the command variance time dot N A and then you have to write N A dot R m is equal to true and now you can see here it is giving you the value.

And similarly if you want to find out the standard deviation or standard error you can find out the square root of its value and it is coming out over here like this. Now if you want to have this variance with the divisor N you can see here I have just copied the command which I have pasted here and it is coming out over here like this. And similarly if you want to find out for the square root of its then you will obtain the standard deviation or the standard error. On the other hand if you want to see here what is the length of the data vector here time it will come out to be a 20, but if you want to find out here length of data vector here time dot N A it is coming out to be 20 because there are 20 values. And if you try to use here this if you try to use here this N A dot R m is equal to here true then it is not working it is giving you error.

So, that is why we have to use the command here N A dot omit. Ok, so now we come to an end to this lecture. Now you can see that in this lecture and the other lecture we have considered that how to handle the missing values and how they are implemented in measuring the central tendency and variation. Well I have taken these two concepts because you are familiar with these concepts. But now as we try to take more concepts the problem will remain the same means every function will have two option to handle with the complete data and to handle with the missing data. I will try to handle here all the cases where we have the complete data and I will not be handling any case further where we are going to handle the missing data.

Why? Because now I have given you the basic concept and now you have understood that just by making small changes or modification in the basic command you can obtain them for missing data also. Now where you have to use N A dot R M where you have to use N A dot omit this is your job to find out from the help. Right, so now I will say that you try to take other different commands which you know in R and try to see how they can be modified for a missing data vector. This will give you more confident and you will be more confident when we are trying to handle more commands in the further lecture where I will not be handling the missing data vector. So, you try to practice it and I will see you in the next lecture till then goodbye. Thank you!