Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 03

Lecture – 12

Measuring Central Tendency with Missing Data

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we had understood that how are we going to handle the missing value in the R software. That means if there is some data set which has got missing values because of any reason. Now how R is going to handle them? Then how are we going to identify the places or that where we have missing data? How are we going to identify whether my data set has missing values or not? Then how are we going to identify that what are the places where we have missing values? How are we going to create a vector after removing the missing values? Now the next question comes that how are you going to implement this concept in different types of functions related to particular statistics? That in case if there is some missing data then how are you going to implement the statistical procedures with the missing data? And our basic objective is now here is that which will be remain the same in this lecture and in the next lecture and in all the basic concept when we are trying to handle the missing data in the way I explained that we want to remove the data and then we want to compute our statistical procedures with the remaining available data. Please note that I am not handling here the topic of missing data models where we try to estimate the missing values and we try to impute them inside the data vector So, that the data vector looks complete. So, please keep this in mind and if you want to learn how to handle the missing data there are different functions available in the R software but definitely, I am not going to consider them in this course.

So, now in this lecture we are going to concentrate how are we going to measure the central tendency of the data when the data vector has some missing values. For example, suppose I want to know that how much a new medicine is going to control the body temperature that means if somebody has got a high temperature and suppose if the new medicine is given to the patient for how much time the body temperature will remain in control after taking the medicine. Now suppose I take a sample of suppose 5 patients and

I decide that I will give them a medicine today and tomorrow they will come and report me and this process I will continue for say 5 days. Now suppose the one patient comes on first day takes the medicine comes on second day takes the medicine, but does not come on third day.

Now I do not know whether the what was the recording of the body temperature or the duration in which the body temperature remained in control for the third day. Now the patient comes again on fourth day and fifth day. Now I want to compute suppose the arithmetic mean on the basis of given set of data that means first day second day fourth day and fifth day. So, how are we going to compute the mean and similarly how are we going to compute all other measures of central tendency that we had discussed in the earlier lecture. So, let us begin this lecture and we try to understand that how are we going to measure the central tendency of the data when we have missing data values.

Okay So, now you recall the arithmetic mean of the observations $X_1, X_2, \ldots X_n$ is defined as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and we had understood that mean(X) that is the R command to compute the value of the arithmetic mean of the data in the data vector X. All this data $X_1, X_2, \ldots X_n$ this has been stored in a data vector X and just to give you an example what I am going to do here that I have created here a small data set in which the time taken by 20 participants in a race is recorded and it is in seconds. So, what I will do that this is my complete data and after that I will try to remove some data and I will replace them by NA and I will try to create the data with missing values and I will try to execute different operation over the same data sets. So, now we know that if you store all these data values in a data vector time and if you want to know the arithmetic mean of this of the values in this data vector then I can find it by mean of time and it comes out to be here 56. Well you are aware of this thing So, you can see this is here the screenshot.

Now if there are some missing values in the data then I have to modify this command mean and we try to write down here mean X and we try to add NA dot rm is equal to true this true is a logical variable right. So, if I try to add this command here NA dot rm is equal to true So, this NA means this is trying to indicate the missing values and this here rm this is trying to indicate here the remove right. So, when I am trying to say here that NA dot rm is equal to here true then it is trying to say here that please try to remove the missing values or missing values have to be removed the answer here is yes and this is indicated by true and if there is false that means the missing value need not to be removed and this is actually here the basic default operation right. So, now this command will give us the arithmetic mean of the values after removing the missing data that is NA values from the data vector. So, now let me try to consider here the same example in which we had recorded the time taken by 20 participants in a race in seconds.

So, now you can see here I have made the first two observation here as say NA well I have made it here underlined just to have your attention otherwise you do not have to make it underlined. Now I try to store these values in a data vector and I indicated by time dot NA So, time was my original data vector and dot NA is simply indicating that there are some NA values yes there is no compulsion to put here a doubt that you please do not misunderstand that you have to give the data vector name with a dot. I am just taking it for the sake of convenience right otherwise this can be time 1 time 2 also or whatever you want. So, now you can see here this is my here data vector and it is entered in the R software also. Now if you try to find out here the mean of this data vector time dot NA it will come out to be here NA because it is trying to find the arithmetic mean of NA plus NA plus all those values which are here like 45, 83, 74 like 45 plus 83 plus 74 and … divided by 20.

So, this will come out to be here NA but now what you are trying to do you are trying to remove this NA values and you want to compute here 45 plus 83 plus 74 up to here divided by 80. So, for that I try to add here a command NA dot RM is equal to 2 and this will give me here the value 58.5 and in case if you try to use here the command NA dot RM is equal to false you can see here this is here the NA which is here just like here NA and this is actually the default mean option right. Now I would like to have your attention once again that if you try to see when you had the completed data set then the arithmetic mean was 56 and now when you have removed two values and the mean is changed and it is coming out to be 58.5.

So, that is indicating that when there is a missing data in the values then the values which could have been based on the complete data set will change and that is why it is important to understand how are we going to handle this missing data right. So, this is here the screenshot but before going forward let me try to show you these commands on the R console also right. So, I already have taken here this data I have entered you can see this is your time and this is your here time dot NA right. If you try to find out here mean of your time this will come out to be here 56 and if you try to find out the mean of your time dot NA this will come out to be here NA. Now if you try to use here the command mean time dot NA and NA dot RM is equal to here true then you can see here it is coming out to be here mistake error why because you are trying to put here additional this bracket.

So, this is the concept of here error I explained you in the last lecture also that means once there is error then the command will not move forward but if you try to correct it then it will move forward and this value will come out to be 58.5. On the other hand if you are getting a message like warning then this value will come but then you will have to look into your command that how are you that where is the mistake where is the slip right. And if you try to use here this command that NA dot RM is equal to false you can see here it will come out to be here NA which is equal to here finding out the mean of

those values which are having the missing values also right. So, this is how we go forward this thing right.

So, this is where I am trying to explain you that what is happening that when you are trying to find out the mean of all the complete values 20 then it is computed here as say 1 upon 20 summation i goes from 1 to 20 x i and this value here is 56 but when you are trying to compute the mean in the missing data vector those So, those two values they are removed and then we have here 20 minus 2 18 observation and the arithmetic mean of the remaining complete observation is obtained here as x bar is equal to 1 upon 18 summation i goes from 1 to 18 x i and this value is here 58.5. So, now this type of changes will happen in each and every function. The reason why I am trying to show you here this difference that will in the case of this arithmetic mean I have shown here after that we are going to consider different types of function in this lecture and in the next lecture. So, this type of changes are going to happen in each and every case but your objective should be that you should actually know what is happening.

So, that will help you in better understanding otherwise R will do in the way it has been programmed right. Now in case if you recall the in the last in the earlier lecture we also considered the concept of median to find out the central tendency of the data. So, now if you have all the data available that means there are no missing values then you can use here the command median x that is m e d i a n inside the parenthesis data vector x which all the values no missing values here right no missing values. But in case if your data vector has some missing value then the same command can be used as median x but now you have to add here the option na.rm is equal to true and after that if you go to the help menu there will be many other options. Well, I am not going into those details but I request you to please look into the help menu and try to understand what those different procedures are trying to explain you.

As I said in the beginning that our software is not a black box it has all the options the only thing is this you have to customize them according to your need and requirement. So, if you try to use this command then all the missing values which are indicated by na they will be removed and the median of the remaining observations will be obtained right. So, if you try to see here the means I have taken here the same example which I took earlier in which the some time it taken by some people to complete a race is now recorded but I have changed the data name So, that you do not get confused between the two. So, if you try to store all these values in our data vector, I have given it a name here minutes minutes and then if I try to find out the median of this data set minutes it will come out to be here 26 right it is here like this.

Now in case if you try to remove here first two data points suppose if I remove here and I try to create here another data vector minutes.na then if you try to compute here median of minutes.na and right here the command na.rm is equal to true then it will come out to

be here 26 right. So, this is how it is going to be computed right but let me try to show you first these examples on the R software. So, let me try to first take this data set minutes you can see here minutes and then if you try to find out here minutes of here median of here minutes it will come out to be here like this and now if you try to take here the data sets minutesna and if try to find out here the median of this minutesna you will see here it is NA but if you try to add here the command na.rm is equal to true you can also write here capital T also it will come out to be here 26. Now in this case actually if you try to see both the values are coming out to be 26 because it is a robust estimator So, you should not think that the values are not going to change in the median. Median is a robust estimator that is why a small change in the data values may or may not alter the value of the median but anyway means I am trying to convey that this 26 and 26 are here same in this particular case but if you take another data set they may change right.

So, do not think that these missing values are not going to impact this median right. Okay So, now we come to an end to this lecture and you have seen that okay I have taken here the two measures of central tendency which are here mean and median in which I have explained you that how are you going to compute them when we have missing data in the data vector. Now similarly if you want to execute another type of this measures of central tendency like geometric mean harmonic mean etc they can also be executed on the similar lines but definitely I am not going into those details here but I leave it up to you because the concept I have given you and now I expect that you should also move along with me So, that you can handle them very easily. So, my request will be you try to take the same data set here time dot na or median dot na and try to compute this geometric mean harmonic mean etc and try to see yourself are you able to do it or not. So, in the next lecture I will try to now consider the measures of variations and I will try to show you that how you can handle the different measures of variation when the data vector has missing values.

So, you try to practice it and I will see you in the next lecture till then goodbye.