Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 04

Lecture – 14

Coefficient of Variation and Summary

Hello friends, welcome to the course Multivariate Procedure with R. So, now you can recall that up to now we have understood different types of measures of central tendency and variation and how to implement them when you have complete data as well as some missing data. Now from this lecture we would like to start moving towards more than two cases. In case if you want to compare more than two data sets there can be different characteristics by which we would like to measure. For example, there are two data sets and suppose they have got different means and different variances, different type of central tendency, different type of variation. In practice you would like to consider actually both the aspects.

Mean and variance or say variability both are important aspects of any data set and suppose if you want to compare such two data sets considering their mean and variance how to do it. For example, if I say suppose the salaries in India and US, many times you read in the newspaper that somebody has got such a very high salary in US which is very low in comparison to an Indian employee and people get very excited that person in US is getting much much higher salary for the same job for which an Indian employee is getting much lower salary. Do you think that are they really comparable? Yes, I agree that in India we are getting the salary in Indian rupees and in US I will be getting the salary in terms of US dollars and approximately 85 rupees is equivalent to 1 US dollar. So, if you try to multiply the US dollar in by 85 then the salary becomes huge and people get really excited.

People actually try to think that as if they are getting the same salary in India. In India depending on the different cities the cost of living varies a lot. It may be possible that in India in an income of rupees 85000 a month possibly in a given city I can have a comfortable life which is equivalent to 1000 dollars, but do you think that in 1000 dollar can I have the same standard of living in US? It is very difficult to say yes, because house rent whatever I am going to give here in Indian rupees that cannot be simply converted into the house rent to be given in US in US dollars. So, these types of different question comes and we have different types of data sets which vary with respect to different characteristics. So, now the first question is that how to compare these different type of data sets with respect to central tendency and variation jointly number one.

And then whenever we are trying to analyze any data set there are different types of characteristics there can mean, median, different types of quartiles, minimum value, maximum value all are important. They all are jointly trying to characterize the data and just by comparing one particular characteristic may not give you the correct outcome. So, the question is how should we compare such data sets or how should we present the data in such a way such that we can easily compare them. So, keeping this aspect in mind I am going to introduce in this lecture the coefficient of variation and summary command for the data sets. Although I will be giving you here the example of two data set, but without any problem without any loss of generality they can be extended to more than two data sets.

So, with this objective let me begin this lecture right. So, now if you try to see the two topics which I am going to cover here are the coefficient of variation and summary command. So, now what is coefficient of variation? The coefficient of variation which is popularly called as CV this measures the variability of a data set without reference to the scale or units of the data. So, that is why when you are trying to compare the salaries in US and salaries in India the main hurdle is coming because of the units and their conversion issues. So, now somehow if you can make the comparison unit free then possibly it will give you a correct outcome.

So, that can be achieved by the use of coefficient of variation and it is useful in comparing the results from two different surveys or test in which the values are collected on different scales right different unit different scales all are possible. Now let me try to consider here an example for example, if I say suppose I have got two data sets and I try to find out their sample mean. Suppose the first data set has sample mean $\overline{x_1}$ and second data set has a sample mean $\overline{x_2}$ and suppose the standard error of first data 1 is S1 and standard error of second data set is S2. Now how to compare these two data sets up to now we have considered either the sample means or the variance or standard deviation standard errors. So, now we can use here the concept of coefficient of variation.

The sample base coefficient of variation measures the variation which uses both automatic mean and standard deviation and it is defined here as like this ratio of standard deviation and X bar. So, CV is equal to S/$\bar{x}$ where S is your here standard error or equivalently in common language I will say standard deviation of the data. The positive square root of variance and X bar here is the arithmetic mean. Now you have to understand that standard error or standard deviation is always greater than 0. So, now if this here S and X bar they have got different types of signs then possibly this may be misleading.

So, we assume that C V is well defined when sample mean is also positive right and in order to compare them we have a rule that we try to compute the C V of different data sets and the data which has a higher value of coefficient of variation is said to be more variable than the other. So, now if you have here suppose three data sets in which are varying or differing with respect to mean variance or both etcetera. So, I can compare here the coefficient of variation as CV1, CV2 and CV3 and then I can say that what is ever is the maximum value among this CV1, CV2, CV3 the corresponding data set has the highest variability and So, on right.

So, let me try to illustrate this with a very small example. Suppose there are two experimenters who measure the heights of the same group of children in meters and 20 meters and suppose the first experiment finds the sample mean to be 1.5 meters and the second experimenter finds the mean as 150 centimeter and similarly the standard deviation or standard error of the first data is 0.3 and second data is 30, but you can see here S1 is 0.3 meters whereas, S2 is 30 centimeter.

Now how to compare it if you go by the absolute value then the second cricketer set has got a higher mean as well as higher variability, but is it correct. So, in order to understand it we try to find out the value of coefficient of variation and for the first data set this comes out to be 0.2 and for the second data set it is coming out to be 0.2. So, that means both the answers are the same. So, now you can say here that the both the data sets are the same and there is no difference with respect to mean and variance right. So, that is how the concept of coefficient of variation helps us right.

So, this coefficient of variation helps in comparing data sets on two completely different measurements right. These variables may be measured on different scales, but their dimensionless coefficient of variable dimensionless coefficient of variation helps us in making the comparison of the variation of these variables or data sets. For example, if you try to say the rents of houses in a metro city as well as in a village they vary a lot, but then the income in the village may be lower than the income in the metro city. So, the rents of houses for example, in India in Mumbai they are much much lower in comparison to the rent of houses in London because the rent in London is measured in pounds and the rents of houses is measured in India in Indian rupees. So, now how to compare, but in such cases this coefficient of variation helps us and this can be extended to more than two data sets without any problem and the data set having a higher value of coefficient of variation is said to have more variability and obviously any data set which has got the lower variance is always preferable right.

Suppose if you have two data sets and we find out their coefficient of variation is CV 1 and CV 2. So, if CV 1 is greater than CV 2 then we say that the data from which we have obtained the CV 1 is having more variability or say less concentration equivalently around mean than the data in the second data set from where we have obtained the CV right. So, the concept of this CV helps when there are various weight data sets, multiple data sets and in case if I try to make a systematic ordering of all the values of CVs then it helps in identifying the data sets with varying probability right. You try to obtain the coefficient of variation of all the data sets and try to arrange them in ascending or descending order and then you can decide that which of the data sets are going to have a similar variability or which are the data set which are quite different than other data sets. So, in case if you want to compute this coefficient of variation in the R software it is not a difficult thing.

You know how to compute variance, you know how to compute the square root of variance and you know how to compute the mean of the data set. So, if you have got a data vector X then the coefficient of variation can be obtained just by square root of variance of X divided by mean of X and yeah if X has some missing values, then you know how to compute the variance and mean with the missing data values. So, suppose the data vector having missing values is indicated by xna then in that case this expression will be converted into a square root of variance of xna by using the term this option N A dot R M is equal to true and finding out the mean of this X and A data vector using the option N A dot R M is equal to true right. And many times, when you have got the group data also can also this type of definition can be defined for those things also I although I have not considered here but whatever we have considered here in mean, variance, median etcetera etcetera they can be defined for ungrouped data sets right. So, let me try to show you here the example.

So, here once again I am going to take the same data set which I have used earlier that we have the data of 20 participants who participated in a race and their time taken to complete the race is recorded in seconds which is here and this data is recorded in a data vector time. So, if you want to find out the coefficient of variation of this data time then it can be found as square root of variance of time divided by mean of time which is coming out to be here like this right. And in case if I try to consider here the missing data vector and if I assume that the first two data values in this data sets they are NA and suppose this data is stored in another data vector time dot NA here like this these two are the missing values then in that case we can use the expression for the missing data and we introduce the option na.rm in the computation of variance as well as mean over here and then we try to compute the coefficient of variation which comes out to be here like this. So, it is not difficult and this is the screenshot, but you let me try to show you it on here on the R software. So, if you just to save the time I already have stored this time and time dot NA data in my R.

In case if I try to find out the coefficient of variation So, that will be square root of variance of time and divided by here mean of time. You can see here this value is coming here like this and if you try to do here with time dot NA then you can see that what will happen it will give us the value NA because we are not removing the missing data. Now in case if you try to use here the expression for the missing data you can see here like this now it will come out to be here like. So, it is not difficult at all. So, now let me come back to our slides and we try to proceed further.

Now before trying to understand the another statistical function for handling different types of data sets. Let me first try to give you here a quick review of the concept of data frames. I am assuming that you are well aware of the concept of data frame in R software because we are now going to use this concept of data frame in this lecture and in the further lecture. So, I thought that let me try to give you here a quick idea that what is data frame and how do we construct it. So, we know that whenever we have more than one values in a data set then we try to use different types of command to combine them.

For example, we have used here the command c to write c(1, 2, 3) this type of data vector then we have a command cbind then we have a vectors and matrix function which also try to combine the values in different ways, different formats and they follow different types of mathematical rule. Similar to that there is another option in R software which is called as data frame. The data frame is something like you can say spreadsheet right because for example if you can recall if you try to construct a spreadsheet in say MS Excel software then how it looks like these are your here columns, these are your here rows and there are different ways in which you can assign the data in here in different types of here cell right. And the option here is this either you try to put here number 2 3 4 or you try to write down here the characters apple, banana, cake etcetera. In a spreadsheet you can have all types of data with you.

So, similar is the data frame where we can combine the variables of equal length with each row in the data frame containing observation on the same unit right. What does this mean? In a spreadsheet one can see here suppose these are the name of persons person 1, person 2 and so, on. So, this row is giving the information on the details of person number 1 and similarly the second row tries to give the details of the person in the person number 2 right. So, similar is the statement which I am trying to make here right. So, the advantage is that one can make the changes to the data without affecting the original data.

For example, suppose we have here different rows and different columns and we want to extract a small data set consisting of some selected rows or selected columns then it is possible to extract subset of the data from the spreadsheet. For example, all the files from this MS Excel software which we commonly call as excel file they are the spreadsheets right. And the advantage is that one can combine the numerical values, character string as well as factors in the data frame. On the other hand, if you try to look into the commands the C bind and matrix, they cannot be used to combine different types of data sets and this data frames are special type of objects in the R software which are especially designed to handle the data sets right. So, as I said the data frame format is similar to a spreadsheet where columns contain the variable and observations are contained in rows okay.

So, now my objective here is not to give you the complete details and information about the data frame but we need to create a data frame So, that we can arrange the data in a required format which is needed for different type of statistical functions. So, that is why I would like to show you here that how we can create the data frames. In order to create the data frames we have a command here DATA.FRAME all in small case and this data.frame function is used to create the data frame by adding different column vectors to a data frame.

So, let me try to give you here an example. Suppose I try to take here 3 variables out of them one is number and say another 2 are alphabets of different types. For example if I take here X, X is my here numbers 1, 2 up to here 16, Y here is the first upper case alphabets like A, B, C etc. and Z here is the lower case alphabets A, B, C up to here first 16 alphabets. So, you can see here this is here X 1 to 16 number, Y here is character from A to P upper case alphabets and Z here is small a to small p as a lower case alphabets.

Now I want to combine these 3 variables or say the data in these 3 variables which are considered as a column in a single spreadsheet. So, for that I have a command here data.frame. Now I have got 3 variables, 3 data vectors you can see which I have defined. I simply write to them here separated by comma X comma Y comma Z.

And now you can see here how it looks like. The first column here is the values of here X which were here you can see. Now secondly if you try to see the Y, Y is here upper-case alphabets from A to P which are your here and then the third column here third column here is Z which is the lower-case alphabet from A to P that you can see. And this is here the screenshot how it will look like when we try to execute on the R software. Now after this I am not going to give you the details of this data frame because there are various types of operations which can be done with the data frame.

How you can extract a particular data set, how you can manipulate them, different types of things. But my objective here is to illustrate here that how I can create a data frame. If you try to see here my X here is like this, Y here is like this and Z here is like this. Now if I try to create here a data frame, data.frame(X, Y, Z) you will see it is coming here like this.

And if you try to interchange the ordering of these values suppose if I make it here X, Z comma here Y then you will see here that the variables are arranged in the way we have defined them. And even if you try to make here some changes for example if I try to take here X, Y, Z and if I want to write down here once again here X if I try to add it here you see without any problem it is coming here. This X.1 is the name which has given automatically to X because X is already here but then anyway there is an option to change the name of the variables also without any problem.

So, this is how data frame is helping us. So, now we come to our command that we want to understand in this lecture. In our software we have a command to summarize the information about our data set which is called the summary right. Cortial, mean, minimum value, maximum value etc. all these things they can be just obtained together in a single command which is a summary command, summary all-in lower-case alphabets.

So, suppose if I have a data vector here X So, I can write here summary inside the parenthesis X and then it will give us the information about minimum value, maximum value, mean value, first quartile, second quartile which is actually median and the third quartile. So, you get basically a good spectrum of the statistical information of different data sets. So, let me try to explain you with this example. So, I try to take here the same data set in which there are 20 participants who participate in a race and their time taken to complete the race in seconds it is stored here in a data vector time right. Now if I try to use the command here summary time.

So, now if you try to see the outcome, the outcome here is the first value is giving us the minimum right. Then second value is the value of the first quartile which is here 41. Next value is the median.

The median of this data is 56.5. Then here is the mean. Mean of this data set is 56. The value of the third quartile here is 68 and similarly the maximum value of this data set is 84. Now the next question comes here how this information is going to help us. You see whenever you are trying to get different data structure there can be different properties mean variance etc you can compute directly.

But if you want to see for example the skewed nature of the data whether the data is more concentrated on left hand side or right-hand side such type of information can be obtained by comparing the first quartile median mean and third quartile. Well, I am not going to explain this idea over here but if you wish you can look into the course on descriptive statistics with R software and you will get the complete idea how to take such decisions. What I am trying to tell you here that this is the summary command which is giving us six pieces of information and when such pieces are combined together in different ways, they give us a very good information about the data set. Now suppose if I make the first two observation to be too high suppose it is here 32 and 35 but suppose I make it here 320 and 350.

Now I want to compare both data set. This data set which is here in time and this data set where I have made two values to be very high which I have stored in a data vector time1. Now means if you try to take here the time1 and more than two variables you will get summary commands for each of them and if you want to compare them then you have to write down all the results side by side and then you have to take a proper conclusion. But in case if you can combine all such variables in the format of a data frame then when you try to operate the summary command over a data frame then all the results which are in the outcome of summary command they can be obtained in the same frame at the same place So, that it becomes easier for us to compare.

For example, now you can see here that in the first and second cricketer sets the main difference is that two values are very high. So, now if I try to create a data frame by using the command data.frame of time and time1 and then if I try to operate here a command summary then what do you get? You can see here this is here the first variable time and second variable here is time1 and this is here the minimum value of both the data sets. Similarly, if you try to see here the value of the first quartiles of first and second data sets it is here like this. The value of the second quartiles they are given here like this. The value of the arithmetic mean of time and time1 which is here like this. Similarly, the value of the third quartiles of data set 1 and data set 2 they are here given here like this.

And similarly, the last row is about the maximum values of this first and second data sets. So, now you can see here that all these values are written here. For example, if I try to clean it then if I want to compare the minimum values of the first and second data sets I can compare them very easily. And if I want to compare their skewness how the data is spread just by comparing the median and third quartile of first data and second data sets, I can get a very good information very easily. And similarly, you can suppose there are some more data set time2, time3 and so, on.

All these values will be written here and you can numerically compare them without any problem. Right. So, if you try to see this is here the screenshot of this one and yeah, all this information if that is presented in a graphical way this will also be good but that we will try to see in the next lecture but here we try to first see what is happening with the time and time1. So, if you try to see time data is already here now, I try to enter here time1.

So, this is your head time1. So, if you try to see here summary of your time it is here like this and summary of your time1 this is here like this. But in case if you want to compare both of them suppose if I say if I want to make here a data frame say d let me define data.frame and then inside the parenthesis time and time1. Right. So, if you try to see here this is my here d this is data frame and now if you try to find out here the summary of data frame d then it is here like this.

So, you can see here it is very easy to compare mean versus mean first quartile versus first quartile and so, on. Right. So, now let me come to an end to this lecture and you can see here that we have device here of way out by which we can compare different characteristics in a single view using these commands. Be it coefficient of variation or be it is summary command this will help you in comparing more than one data sets and more than one parameter at the same time. So, if you try to see the way we have progress that first we introduce the concept of mean then we introduce the concept of variance then we introduce the coefficient of coefficient of variation and then we introduce the concept of summary which is trying to take care of more aspects of the data set.

So, if you try to see gradually now, we are moving towards from univariate to multivariate and when we are talking of the multivariate once again there can be two aspects one is the analytical aspect and another is the graphical aspect. For example, whatever we have done up to now this is all analytical option. Analytical option means numerically you are trying to see the data. Another option is visualizing them for example using some different types of graphics, plots etc. So, now you have seen here how summary command is trying to compare the values together numerical values together but can you express it in a graphical way also.

We all understand that graphics are very easy to understand. A smiley can or an emoji can express the emotions much easily than the numerical values. So, now gradually we will try to consider more aspects but my request to you once again is that you please try to create some data set yourself try to take a small value and try to create different types of situation inside the data sets. For example, you try to take the values between say 20 and 30. Now you try to make one value to be very high say 200.

Then you try to see what happens to this summary command compare them. Now you try to make two values very high or you also make one value two values different values to be very low also. Try to see how the variation of this in the values of different types of quartiles that occurs. And this will give you an idea that how numerically you can inspect the data and can draw different types of statistical inference. It is just like the medical report by looking at different types of the values in a blood test report etcetera the doctors get this idea that what is happening inside the body. By looking at the blood report there are many parameters there are some numerical values are given there is some range normal range etcetera.

But when the doctor tries to look at it they can reveal that what is happening inside our body. Sometime univariate that means single value of some characteristic may not be helpful but they try to correlate whether the okay this value is high but what about this value is it low. And then they try to take the decision in a multivariate way. So, that is the quality which I would like you to develop and this quality can be developed only when you try to play with the data set. So, initially you try to create disturbances in the data set try to compute these values and try to see how these disturbances are captured by these values.

Once you get confident now you can go into the reverse direction looking at the numerical values you can judge what possible disturbances are occurring inside the data set. And that is how you will become a good data scientist. So, you try to experiment these things and I will see you in the next lecture till then good bye.