Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 02

Lecture – 09

Univariate Data - Central Tendency and Variability

Hello friends, welcome to the course Multivariate Procedure with R. So, now you can recall that in the last couple of lectures, we had discussed some basic concepts and basic operation in the R software. Now, I believe that you have a fair background about the type of syntax, commands, what we are going to use in the R software. Now, I would leave the remaining part to you that you try to make yourself better in implementing the R software in different types of data analysis. And now from today onwards, I will try to move towards statistics. So, whenever we are talking about the statistics, so, the statistical tools are developed for univariate data, bivariate data and so, on multivariate data.

What is the difference? In univariate data, usually we have observation on only one random variable. Suppose I say height. Height is my random variable. So, we try to collect the observation only on height of the persons. And suppose we try to collect the height of 20 persons, then we say that we have a sample of size 20.

When we try to take the data or collect the data on two variables, then this is bivariate. Bivariate means two random variables. For example, if I have 20 persons and we try to collect the data on their height and weight, each of the 20 persons have given their height and weight. So, for every person, there are two observations, one on height and say another on weight. And we have here two random variables, one is height and another is weight on which we are trying to collect the data.

So, this is bivariate. When we have more than two variables, then can be tri-variate or multivariate. For example, if I try to collect the data of 20% on say height, weight and age, then there are three variables and we can say that this is a tri-variate distribution or tri-variate case. We have three random variables. And when we have more, then we usually call them as multivariate.

When we are trying to handle the tools in univariate, bivariate, tri-variate, multivariate, you have to understand that the tools are developed from the mathematical side. The rules of mathematics and statistics have to be applied to develop a tool. But here in this lecture, we are more concentrated on the, we are trying to emphasize more on the application. But theory is also important, So, I will try to give you the brief concept behind the tools and then application. So, although our objective is to learn about the multivariate procedures, but we would like to understand first like univariate case, bivariate case, etc.

The procedures which I was telling, they are developed for univariate, bivariate, tri-variate and say multivariate, they differ on different aspects. For example, if I say mean, suppose I have only one random variable, then it will have certain mean. But if I have two random variables, then I will have two means. For example, if I have univariate case, suppose I want to collect the data on height, then I have only the mean of the height. But when I am trying to collect the data on height and weight, then I will have mean of height and mean of weight, that means the arithmetic mean of the observation collected on height and arithmetic mean of observation collected on weight.

And similarly, if I try to increase the order, tri-variate, multivariate, similarly this number of means will also increase. On the other hand, in case if I take say another parameter, say this variability or variance, So, when I have only one variate, univariate case, I will have only variance. And if I have two variables, then I have the variances of both the variables as well as their interaction. Suppose if I have the data on only height, then we will have the variance of the data on height. But when we have height and weight, bivariate data, then we will have the variance of height as well as variance of weight, but their interaction, their covariance will also come into picture.

So, we will also have covariance between height and weight. Similarly, when we try to go for tri-variate or say higher dimension or say multivariate case, then we will have the variance of each of the variable and then we will have their covariance also. Covariance between first, second variable, first and third variable, first and fourth variable, second and third variable and So, on. So, that is why we need to understand these multivariate procedures, but the knowledge and information about the univariate procedures, bivariate procedure is a must to understand. The reason being it is easy for me as an instructor and it is easy for you as a student to understand the univariate thing.

Possibly bivariate is also not difficult, but when we try to go into the higher order or say multivariate things, then it is difficult to visualize. We all agree that if I can see anything from our eyes, then it is easier for us to understand. Well other things we can think, we can imagine. I am not saying they are difficult, but they are not So, easy also. For example, if I want to explain you about aircraft, if I show you the aircraft you will understand it very easily, but if I try to create a picture about the aircraft, it may not be

easy for you to understand that what I am trying to say. Whether the wings are going to be like this or like this.

So, that is why when we are trying to develop and understand the procedure for the multivariate statistics, it is important for us to first understand the basic simple tools in the univariate and bivariate. And it will be easy for me if I can explain you the univariate concept and then I try to extend it to a multivariate concept. Above all, beside all these things, when we are trying to deal with the different types of statistical procedures which are used in the multivariate case, then some concepts are needed like as mean, variance, covariance, correlation coefficient, covariance matrix, etc. So, as I do not have the background of all the participants who are attending this course, So, I feel that it is important for me to cover the basic minimum things in this lecture and devote some time over the tools which I am going to use further. So, that is why in this lecture, I am going to concentrate on the basic fundamental tools of a univariate statistics.

And as a matter of fact, whenever the data comes to us, there are certain characteristics which we would like to understand for before we take the decision that how we want to move forward. For example, I would like to know about their central tendency, variability, etc. Well, these are very simple concept and I am sure that you understand you might not have seen them mathematically and how to operate them in R, but these are very simple concept in which we all are always interested. For example, as a student, you always ask your teacher what are the average marks of the class. Why? By looking at the average marks, you get to know that where do you stand in the class.

In case if the average marks out of 100, they are 50 and suppose you have got 40 marks, then you say that okay, you are close to the average, but in case the average is 90% and the person has got only 40%, that shows that there is a huge difference in the performance of the candidate with respect to the remaining class. And similarly, it is the concept of the variability also. For example, if I ask you that how much time will I take in going from my house to the railway station, somebody says okay, it will be say 15 to 20 minutes and somebody says it will be 5 minutes to 45 minutes. So, you understand what I am going to say. Definitely, you are going to believe on the first sentence more because you say that it is more precise.

As soon as you use the word precise, I say in Statistics, this is about variability. So, now we understand these concepts, but in order to implement them in the multivariate procedure, we need to formally understand it. And above all, whenever we are looking at the multivariate data, then we try to analyze it by looking into a univariate direction or we try to divide the high dimension into smaller pieces So, that we can understand the behaviour of the data. Keeping all this in view, in this lecture, I am going to talk about descriptive statistics in a univariate case. And I am going to address here different types

of tools which are used in measuring the central tendency of the data and variability in the data.

And I am not going into detail, but I want to show you what are these things and how they are computed in the R software So, that it will become easier for you to understand the topics in the multivariate procedure. So, let us begin this lecture and try to understand this basic concept and their implementation in the R software. So, now in this lecture, we are going to talk about the univariate data and we are going to concentrate on the measures of central tendency and variability. So, before I go into any other direction, this is obvious that whenever we get any data, we would like to get the first-hand information. Before we go for any multivariate procedure, we would like to understand the nature of the data.

Based on that, we can take a proper call. So, usually, we try to understand what is the central tendency of the data, what is the variation in the data, what type of relationship are existing in the data and we want to study them. For example, in order to understand the central tendency of the data, we try to use different statistical measures like arithmetic mean or commonly called as mean, median, mode, geometric mean, harmonic mean, etc. In order to understand the variation in the data, we try to use different statistical tool like variance, standard deviation, standard error, mean deviation, etc. And similarly, to understand the nature of the relationship existing in data among different variables, we try to use the concept of covariance, correlation coefficient, etc.

And definitely, when we are trying to understand these tools, we are looking forward for the analytical tools. Analytical tools mean those who are based on some mathematical formula, numerical values, etc. There are some graphical tools also like different types of graphics, two-dimension graphics, three-dimension graphics, different types of plots, etc. They also reveal very important information hidden inside the data. So, as an instructor, my advice to you all will be that whenever you try to get the data, try to use the graphical as well as analytical tools both together.

And if everything is fine, in the sense that you are trying to use the correct statistical tool over the correct data, then whatever information you are getting from the analytical tool and whatever information you are going to get from the graphical tools, both are going to match. They will have their own advantages and limitations, that is a separate aspect, but then they are going to match. So, now let me try to take up the topic of central tendency and variation in this lecture and about relationship study. I will try to address these topics in the next lecture. So, first we try to understand how do we measure the central tendency of the data.

What is this? Suppose you get a data and suppose if you plot it, right, and suppose the data comes out to be here like this. So, what do you see here? I can see that here data is

concentrated around this point and if there are two sets of data, one is here like this and another here is like this, then I would say the data is concentrated at this point and as this point. So, we are simply trying to understand the central tendency of the data. For example, means every day we ask what is going to be the day temperature or night temperature and we simply say okay during day the temperature will be 35 degree and during night the temperature will be around 15 degrees. What do you say? Do you think that the entire day the temperature is going to remain 35 degrees and during whole night the temperature is going to remain around 15 degrees? Certainly no.

But if you try to interpret it in different way that if you try to take the temperature over different time point during the day and different time point during the night, then their arithmetic mean or their average will be around 35 degrees centigrade or say 15 degrees centigrade. So, this central tendency of the data gives us the first-hand information about the behavior of the data around the central value. And from this we try to infer different type of information. Now in order to understand this central tendency we have to quantify it. And there are different ways, different statistical measure by which we can quantify it like arithmetic mean, median, mode, etc.

So, I am going to address here some of the commonly used statistical tool to measure the central tendency of the data. So, this central tendency of the data gives us an idea about the data is clustered around what value. Suppose we have got the data which is represented here as $x_1, x_2, \ldots x_n$. What does this mean? I will try to explain you here. Suppose, I have data on height and suppose I have here 3 persons.

Let me call here say person number 1, person number here 2 and person number here 3. And I try to measure their heights. Suppose their heights comes out 150 centimetre, 160 centimetre and say 170 centimetres. So, this is going to indicate here as x_1, second one is x_2, and third one is x_3,. So, and here the random variable here is X which is your height.

So, this x_1, is indicating the numerical value 150, x_2 is indicating the numerical value 160 and x_3 is indicating the numerical value 170. And, suppose I try to compile all these values in a data vector like as your X equal to 150, 160 and 170. So, this data vector is given here as like here X. So, now I try to compile or combine all this information on the numerical value inside this data vector X. Now you want to find out the arithmetic mean.

The arithmetic means of $x_1, x_2, \ldots x_n$ this is defined as say $\bar{x} = \sum_{i=1}^{n} x_i$. It is arithmetic mean but commonly it is called as mean. This is how people try to address it. The next question is if you want to compute the arithmetic mean in R software how are we going to do it. So, the command here is m e a n mean and inside the parenthesis you have to write x that data vector.

Okay. Similarly, we know that there are some other types of means like as geometric mean, harmonic mean. Well, they are arithmetic mean, geometric mean, harmonic mean

they are used under different types of condition depending on the objective of the study. I am not going into that detail but I want to show you that how you can compute them. So, we know that the geometric mean is defined here as say $\overline{x_{gm}}$ here like this which is where we are trying to write down here. This product means the way it is written here is means try to take all the observation $(x_1 * x_2 * ... * x_n)^{1/n}$.

As for finding all the arithmetic mean there is a built-in function mean but for geometric mean there is no built-in function but now you can see that it is not very difficult to write this statement. If you try to see what is this thing? This is the product of all the values. So, we know that this can be obtained by the built-in function p r o d and this is here what 1 /n. What is here n? This is the total number of observations in the data vector. So, essentially this is the length of data vector.

So, there is a command in R which is called here a length l e n g t h and if you write here length(x) then it will try to give you the total number of elements in x. So, I can use this function here or use this syntax here to find out the value of 1/n. So, 1/n can be written as 1/ ln(x). So, this is what I am trying to write down here.

So, this will compute the geometric mean. And similarly for the harmonic mean this is defined here like this. And if you try to see here this can be written here as $\frac{1}{\frac{1}{n}\sum_{i=1}^{n} 1/x_i}$. So, do not you think that the quantity in the denominator looks like as if we are trying to find out the arithmetic mean of $1/x_1$, $1/x_2$, $1/x_n$? Yes. So, this I can write down simply here for example here mean of 1/x. What is this mean upon mean by 1/x? 1/x is trying to give us the inverse of all the values in the data vector x.

That we had discussed earlier in the vector of data vector operations. And then you are simply trying to find out the arithmetic mean. So, if you try to see just by writing here 1 upon mean of 1/x you can very easily compute the harmonic mean. So, the concept which I want to convey here is not really to compute geometric mean or harmonic mean but also that just by writing small statements even if there is no built-in function but still these quantities can be computed very easily. Another popular measure of central tendency of the data is median.

And this is the value So, that the total number of observations above it is equal to the total number of observations below it. Well, that is the definition but what are the situation in which it is going to be used? Suppose if I take here 3 values 1, 2 and 3. Now if I try to find out their arithmetic mean, this mean will be1+2+3 that is 6/3 which is 2. And if I try to find out its median usually it is the middle most value which is when the observations are in the increasing order So, it is going to be 2, no issues.

But now I try to add here 1 value suppose here 100. What will happen to mean? This will be 106 divided by 4 which is equal to here 26.5. Where is about this median? This will

become second and third observation divided by 2 which is equal to here 2.5. What do you understand by this example? I want to explain you here under what type of situation this median can be used.

When you see that the observations are quite homogeneous that means there is not much difference among the different values inside the data vector then arithmetic mean and median they are pretty close to each other. And suppose I try to add here a big value. This value is quite different than the remaining values. Now this arithmetic mean becomes here 26.5, whereas the median becomes only here 2.5. So, if you try to see what is the moral of the story?

The moral of the story is if you get an extreme observation, if you get any observation which is far much away from the remaining observations then arithmetic mean is going to be heavily impacted by this extreme observation whereas medium this is very less impacted by the presence of this extreme value. Suppose if you try to add this value 100 then arithmetic mean goes from 2 to 26.5 whereas median goes from 2 to 2.5 only. So, there are many situations where you would like to use the median to be on the safe side if you feel that there are some observations which are extreme observation and which are not very useful in the sense that they are not indicating that they can be the part of the process.

In such cases we want to find out the median and this median is defined here as say MEDIAN inside the parenthesis X and if you try to do it in the R software it will give you the value of the median. Now let me try to take here a very simple example to explain you. Suppose I try to collect some marks out of 100 about some student and I try to score their marks in this data vector here marks. Now suppose if I want to find out here the arithmetic mean then I simply have to write down here mean of here marks inside the parenthesis and you will get here the value 63.2.

And similarly, if I try to find out the value of the geometric mean then I simply have to write down the same command where I have to write X equal to marks. So, the command will become prod of marks hat 1 upon length of marks and this value comes out to be here 59.61099. And similarly, if I want to find out the harmonic mean this is 1 upon mean of 1 upon marks it is coming out to be 55.78 and the median of marks is coming out to be here 63.

So, you can see here that it is not difficult to compute them in the R software and now let me try to show you these things in the R software also. You see So, this is here the data vector marks and if I try to find out here the mean of this marks this will come out to be here like this you can see here like this. And suppose if I want to find out here the median of this marks this is here like this. And similarly, when you try to find out here the arithmetic mean that is 1 upon mean of 1 upon marks you can see here this is here like

this. And in order to find out the geometric mean let me try to copy this command to avoid any mistake it is here 59.61.

So, now the question comes here that how are you going to use this value. Well, those who are familiar with the descriptive statistics they understand that by looking at the value of mean and median you can decide about the nature of the distribution of the values. If mean and median are very close then it indicates that possibly the distribution of the data it is nearly symmetrical. For example, if I say symmetrical around the mean value or the median value. But if distribution of the data is very skewed say more on one side this way or this way then this indicates that the data distribution is skewed.

And depending on these options you have to choose the appropriate statistical tool which you want to use. So, now this is all about the central tendency of the data. Now we come to another aspect about variability. So, what is variability? If I try to take here suppose two scatter diagrams. Suppose one data set I am trying to put it into say this green colour and then I am trying to put it into here red colour.

You can see here. Now these are my two data set what do you understand from here. What's about central tendency of this data set? You can see here the central tendency is somewhere here and this is true for data set I as well as data set II. But then can you really see that both the data sets are the same? Certainly not. They are different. Why they are different? Suppose the distance between the observed data and the central value this is much more when the where the data is indicated in the red colour and it is smaller in the data which is indicated in green colour.

So, now how to capture this peculiar characteristic of the data? This is actually indicated and measured by the variability. This is actually the spread and scatteredness of for data around any point but preferably we try to use it around mean value. So, let me try to give you here one very simple example to show you that why this variability concept is needed in the data. Suppose, I have two data sets one value here is 360, 370, 380 and its arithmetic mean is coming out to be here 370. Similarly, we have one more data set whose values are 10, 100, 1000 and its arithmetic mean is coming out to be 370.

Now do you think that both the data sets are the same? You can see that here 360, 370, 380 they are very close whereas 10, 100, 1000 they are quite different quite far away. So, the question is this when you are getting the data either in the multivariate case or in a big data case you cannot observe these individual values but you have to differentiate and you have to take out this characteristic out of the data using the statistical tool. So, our objective is this how should we differentiate between the two data sets. So, now we consider the concept of variability. What is variability? In case if you try to see if I have got here two data sets one is here like this and suppose another data set is here like this whose scatter diagram if I try to plot looks like this.

Both have got the same mean value. But if you try to see the scatteredness of data around this mean this is different for green color which are contained in this circle and for red color dots which are contained in this circle. So, this gives us the idea of scatteredness around this mean value and this is the concept behind the variability. This is actually variability and we want to capture the variability in the data set. So, in order to know the variability in the data set we have got different types of measures and among them variance is one measure which is quite popular. I would like to have your attention here that in order to measure the variance what we try to do that we try to consider every observation say xi and we try to take its deviation from the arithmetic mean and now this difference can be positive or this difference can be negative.

So, in order to get rid of the sign of this deviation we square it. So, now this is always positive. Now there are more than one data set So, we try to take the average of all such squared deviations. So, this is basically the definition of variance. Now this variance has two forms. One here is like $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and another here is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

So, you can see the basic difference say between the two forms of the variance is the divisor. One case the divisor is n and say another case the divisor is n minus 1. The reason for these two types of or these two forms of variance comes from the statistical inference. form of the variance this is an unbiased estimator of population variance whereas, this is a variance form with divisor 1 upon n this is a biased estimator of population variance. Well, we are not going into the aspects of unbiasedness or biasedness of estimator because that relates to the statistical inference and we are not going into that aspect.

But because of this property in the R software this variance is computed by this expression which has a divisor 1 upon n minus 1. Well in case if you try to see if the sample size is quite actually large then the value of say here 1 over n or say 1 over n minus 1, they are nearly the same for large sample. Right So, practically it does not make any difference whether the divisor is n or n minus 1 the difference between the two forms of the variance will be very small. But in R software we have to understand what R is computing. So, the R is computing this variance with divisor n minus 1 and the R command to compute the variance is VAR(x) right and one thing we have to observe that this is the sample based right means all the values in this expression $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ they can be computed from the sample values.

When we try to take the positive square root of this variance this is called as standard deviation. Well, there is another terminology which is standard error. So, ideally what we are doing here it is actually we are trying to find out the standard error. Standard error is always a function of the sample values whereas standard deviation contains the population values also means population parameters which are basically unknown to us.

But anyway, in practice say from the user point of view people usually do not differentiate between the use of the words standard error and standard deviation.

So, we are not going into that detail but our basic objective is that we want to compute it on the basis of given set of data. So, now if you want to find out the square root of this variance So, the command here is sqrt and inside the parenthesis you have to simply write VARX. So, this is how you can find out the variance right okay. Now we try to consider the other variant of this variance.

Now suppose you want to compute the variance with divisor 1 over n right. Now in order to do it you have to understand what R is computing. The R is computing this quantity $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Now in case if you want to find this quantity then what I can do that I can multiply and divide it by here n minus 1 and then this will become here like this $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Now what you can do this n -1 and n will come here (n-1)/n and $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Now this quantity is going to be computed in R by variance of x command and this can be multiplied by here n -1/n like this right.

So, if you try to see this is what I am trying to do here what is here n, n is the number of observations in the data vector x. So, this is here n and now I can write down here n minus 1 upon n into variance of x and this is how you can compute the variance with divisor 1 over n. The next concept to find out the variance is the range right. Range is the difference between the maximum and minimum values in the data set and in order to compute it in the R software you have to use the built-in function max and min. For example I can if I have stored the data into data vector x then I have to write down here max of x minus min of x.

I want to have your attention here that you have to be careful that in R software there is also a command range but if you try to give here the range of x then it will give you the values which are the minimum values and the maximum values. It will give you only two values but whereas here we are interested in the statistical range. So, many times we call it as a statistical range right. Similarly, we have one more concept interquartile range.

That means you try to divide the data distribution into four parts. The first part is called as Q1, second part is Q2, third part is Q3 and fourth part is Q4 and it is like that Q1, Q2, Q3, Q4 they are the four quartiles which are trying to divide the total frequency of the data into four equal parts. So, anyway, So, the values at these points where I am trying to highlight, they try to give you the value of the first quartile, second quartile, third quartile and fourth quartile. So, interquartile range is defined as the difference of third quartile of the data minus first quartile of the data and it also give you a sort of measure of the variation right. So, this can be obtained in the R software directly by using the command IQR(x) but you have to be careful here IQR they are in the upper-case alphabets right.

And similarly, if you want to find out the quartile deviation that is the half of the interquartile range that is third quartile minus first quartile divided by 2 and this can be obtained directly by IQR(x) divided by 2.

And is another popular measure of variation is mean deviation or this is also called as absolute deviation. For example, if you try to find out the deviation of Xi minus X bar but it is dependent on the sign this can be positive or negative. So, earlier in the case of variance we had taken the square but now I am asking you to take the absolute value and then take its arithmetic mean right. So, in order to find out the absolute deviation or mean deviation around arithmetic mean there is no built-in formula but it can be written very easily.

For example, there are two ways either you can use the function sum or directly the mean. For example, if you try to see here $X_i - \bar{X}$ this can be written as X minus mean of X and you want to take its absolute value. So, this will be ABS function and now you are trying to find out its mean. So, this will be your mean of absolute value of X minus mean X or the second option is that you simply try to instead of finding out the mean first you try to find out the sum which will be here sum of absolute value of X minus mean X divided by N which is the length of X right. So, this mean deviation is a least will give you the least value when it is measured around the median whereas variance give you the least value when it is measured around the arithmetic mean right. Now let me try to show you these things over a data set here like this and So, this is the same data set what we have considered in the case of measure of central tendency that the marks of a student out of 100 they are stored in a data vector here marks and then we try to find out its variance which is here VAR of marks which is here like this when you want to find out the standard deviation or standard error this is the square root of the variance of marks which is here like this and similarly if you want to find out the inter quartile range you simply have to write down IQR marks which is here 33 and then you have to find out the quartile deviation which is IQR marks divided by 2 and if you want to find out here the mean deviation which is here the same statement what you have developed earlier it will come out to be 17.41333 right.

So, this is how you can go here like this but before I try to move to the last slide let me try to show you these things over the R console right. So, let me try to So, this marks are already here So, if you try to see here this is the variance of marks like this and if you find out its standard deviation or standard error you have to simply find out the square root of the variance which is here like this inter quartile range if you see here of marks it will come out to be here like this and if you want to find out the quartile deviation it is a half of the inter quartile range which is here like this and if you want to find out this the absolute deviation it is here like this and now I would like to find out here the range also right. So, let me try to clear the screen and let me show you here this marks. Now if you try to find out here max of marks minus min of marks this will give us the statistical

range but if you try to find out here the range of marks this will give you here the minimum and maximum value you can see here 29 is the minimum marks and 96 is the maximum marks. So, you have to be careful when you are trying to use it whatever you want you have to write the correct program.

Now if you try to see in the beginning in this slide I have given you this two data set where the arithmetic means are coming out to be the same and my question was how to differentiate between the two data sets. So, now I have got the answer that find out their variances that is what I have done here right. If you try to see the mean is for the data set 1 is 370 but when you want to find out the variance of this data set it is coming out to be 100 and for the second data set the mean is 370 same as the first data set but its variance is coming out to be 299700 it is very high. So, it gives us an idea which is much much bigger than the standard deviation or variance of the first data set. So, that is why our conclusion is very simple the difference between first data set and second data set is that the variance of the first data set is very very low although the arithmetic values, arithmetic mean or the mean values of both the data sets are the same.

So, now we come to an end to this lecture and you can see here we have given we have explained or we have understood very basic concepts mean and variance and but believe me these are the two concept which are which acts as like as a foundation in the descriptive analysis or say exploratory data analysis. Okay here I have given you very brief and short introduction to this concept of mean and variance but there are ways there are several statistical properties which are used and if you know those properties you can use them better. So, one option is that you can look into the books and find out this thing and another option is that I have two more discourses on NPTEL one is descriptive statistics with R software and another is on SWAYAM that is it will get the data analysis in R software. If you want to have more information about this procedure and this concept you can have a look otherwise what I have told you here this is enough for understanding the basic concept which we are going to use in the further lectures. So, my request will be that you try to take some data sets and try to operate these commands over there and try to understand what they are trying to indicate.

Finding all the numerical value is not difficult that your R software is going to do but our job as a statistician or as an analysis or a data analysis that becomes more important and that begins when we have these numerical values before us for which we have to give the right interpretation. Just like you have seen that sometime some people look at the horoscope and tries to tell the future well horoscope is the same but it depends on the qualification of that person that how well worse he is in understanding those values to give the proper interpretation. So, this exploratory data analysis and these tools are something like the same thing that by looking at those values you try to make very important conclusion about the data set.

And the most important part is that you are not looking at the data sets. The data sets are hidden. They are So, huge that you cannot understand anything if you try to look into the data set. The beauty is that you try to look at these numerical values of mean, standard, error, and then you try to reveal different hidden intrinsic properties inside the data and that is the success of using the statistical tool. So, you try to take some data set where knowingly you try to introduce some variation and try to compute these different values like mean, median, variance, mean deviation, etc. and try to understand what they are trying to do.

You need to understand their language, what they are trying to interpret. The more you understand, the better you understand. That will help you in becoming an expert in data analysis. So, I would request you that you try to take some data set either from the book or you try to create your own data sets and expose them to this software and try to understand what they are trying to tell us, try to understand their language. You try to practice it and I will see you in the next lecture. Till then, goodbye.