Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 05

Lecture – 20

Random Variables: Continuous and Discrete

Hello friends, welcome to the course Multivariate Procedure with R. Now, from this lecture and in the next couple of lectures, we are going to talk about the Random Variables and Probability Distributions. And we are going to begin with a univariate case and then gradually I will extend it to a multivariate case. Now, the next question comes why do we need a random variable? The basic job of statistics is to analyse the data. Now, how the data is collected? On what? It depends on the objective. Suppose the objective is to know the age of the persons staying in a colony, then what are you going to do? You are going to collect the observations on age of those people who are staying in that colony.

You are not going to collect the observation on the weight of those people who are staying in the same colony. Also, you are not going to collect the age of those people who are staying in some other area, some other colony. So, if you try to see in this very simple example when the objective is to find out the average age of the people who are staying in a colony, you are going to collect the observations on the age of people who are staying in that colony. So, now what is age? Age is a variable on which you are going to collect the observation.

Similarly, if you want to know the effect of a medicine in controlling the blood pressure, would you like to do? You will try to give the dose of the medicine to a group of patients and then you will try to record the data on blood pressure. You are not going to collect the data on the body temperature of those patients. But in case if your objective is that how the medicine is going to control the body temperature or fever of the people, then what are you going to do? Then you are going to collect the data on the body temperature. So, you see based on the objective of the study you always try to define a variable on which you are going to collect the data. Now, when you are trying to collect

the data on certain variable that is always going to be some random variation which is beyond our control.

Suppose if I ask you that what is the content of nitrogen in the soil in a field. Do you agree with me that the content of the nitrogen of the soil in the entire field is going to be the same or similarly if I say what is the level of moisture in the soil in that field. Do you think that the level of moisture at every point in the field is going to be the same? There will be a small variation. So, that is why there is some uncertainty involved in the observation. And that is why in practice or practically we say that we have got a random variable on which we are going to collect the observation.

Now, different questions come up. How should you define the random variable? How should you measure this uncertainty? And the and when the observations are going to have uncertainty their behaviour will depend on the type of process what they have. So, how to characterize the behaviour of the process through a mathematical function? Why do you need a mathematical function? We understand and we believe that any result which is based on the foundations of mathematics is acceptable and it is well accepted at every place. It is just like if I declare in my class that all the students who are wearing a blue shirt, they will be getting a grade A and all the students who are wearing a say this white shirt they will be getting a grade B and all those who are getting say this grade C. Those who want to have grade C they have to wear suppose a black color shirt and so on.

Do you think that this type of decision is going to be acceptable? Certainly not. But if I say ok, I will take your exam and I will give you some questions for which I have taught you in the whole semester and if you get a more than 90 percent marks, I will give you grade A. In case if you say get a between 75 and 90 percent mark that I will give you grade B and so on. Do you think that this decision is going to be acceptable? Answer is yes. Now, what is the difference between the two ways of making decision? In the second case there is an involvement of mathematics.

So, that is why whenever we are trying to find anything from the data that has to be in some mathematical format and that is why whatever we are observing that has to be a numerical value. Whatever is the uncertainty involved that has to be defined in such a way which is mathematically compatible. So, now all these different aspects we are going to discuss in this lecture and in a couple of lectures in the future. Now, in this lecture we are going to talk about the random variables. So, basically, I can tell you we have a very well-defined definition of random variables which is based on strong foundations of mathematics.

But definitely the objective in the course is not to go into the details of mathematical statistics. We have a topic major theory in which we try to define the random variables in a proper way using the proper mathematical definition. But definitely here I am going to

make it in a very simple way so that you can understand it from the application point of view. Now, in case if I try to broadly classify you can have two types of observations. For example, if you roll a dice then there can be possibilities of number like 1, 2, 3, 4, 5 and 6.

But if I ask you to measure the heights of the students that can be 155 centimeters, that can be 155.1 centimeter, that can be 150.5.1, 1 centimeter and so on. So, you can see that in the case of height the values can be taken in a defined interval. Whereas, in the rolling of a dice the values are going to be taken are only the integer values. And beside this thing when you try to roll a dice then there is always the probability of occurrence of the event. And in case if you try to define your event then the probability also probability is defined and if the event changes, then the probability of event also changes. So, event what you are going to observe is something what you want to translate it to a mathematical function. Because the basic requirement in statistics is that we need the mathematical data or we need the values in mathematics on which we can do different types of mathematical operations.

So, now, we try to understand these things one by one and in this lecture let us try to understand the basic concept of continuous and discrete random variables. So, let us begin our lecture. So, now, the first question comes here why do you need a random variable. And this random variable can be univariate, bivariate or say multivariate. Univariate means there is only one variable, bivariate means there are two variables and multivariate means there are say this more than one variable.

Bivariate is also a particular case of multivariate. So, we know that some concepts are required to draw the statistical conclusion from a sample of data about the population of interest. Why if you try to see what is the basic objective of statistics? Statistics is always trying to observe a small sample of its data and based on that it always try to take some statistical conclusions which are valid for the entire population right. For example, in case if you want to know that what is the initial salary of students who are graduating in the statistics subject from a university. Now, what you can do? That you can draw there a sample.

Suppose I try to draw a sample of say 100 students who are going to complete their degrees in the statistics program and based on that we will try to find out that what is the average salary of these 100 students, what we are going to use it to define or to give the information for the whole population. That mean the expected salary of all the students who are graduating or completing their degree with the statistics subject. Similar is the case when we are trying to deal with the clinical trial. Medicine is exposed or given to a group of people, but whatever are the conclusions they are going to be valid over the entire population right. Suppose there is some new medicine new drug that is given to a

sample of selected patient and suppose some patients show improvement and some patient does not show the improvement.

So, and our objective is that we want to understand what is the what are the different effects or consequences of this medicine when it is given to the people in this country or in this world. So, that means, the population of patients. So, now in order to understand this thing in order to implement this concept that how to implement the conclusion based on a small sample to a whole population, we need some basic statistical concept and among them the random variables are going to lay the foundation of the theoretical concepts which are required to make such statistical inferences and cut the details. We would like to go for sampling, we would like to go for estimation, we would like to go for test of hypothesis which are the different components of the statistical inference and we will see the beginning of all this procedure is the random variable. Now, let me try to give you an idea or an example through which I can explain you that what is an experiment, what are the outcomes and how a random variable is associated and how it is transformed to a statistical function.

So, let me take here a very simple example of coin. Here we toss the coin and whenever we toss a coin the outcomes are either head or tail right. So, we know that the sample space in such a case which we are trying to denote by Ω that is a standard symbol for the sample space, it is a collection of all such points which I am trying to indicate by ω. We have omega where this ω can be either head or tail right. So, I can see now here that ok the outcome of a tossing of a coin is either head or tail.

So, just by writing s t or head or tail, do you think that is it really going to help us? It is very difficult to handle the elements in that case if they are not in numbers. So, now we need to do something so that we can convert this outcome of head and tail into some numbers which is in turn going to help us in doing different types of statistical calculations. So, now we suppose we define this capital X to be a function such that when ω is observed as head that is ω, then the value of the X which is indicated by here X_ω takes 1 and if the outcome is tail then X_ω takes value 0. So, now if you try to see on one side you have here head and tail and on other side you have here the numericals some numbers and in that sense as soon as we have numbers, we are in a position to implement our mathematical functions. So, you can see here that now X is a real valued function which is defined on the sample space Ω, but what is happening? The Ω has head and tail now it is bringing us to some other space where these are going to be indicated by here 2 numbers like this 1 for head and 0 for tail.

So, this type of transformation is needed and this type of transformation is being done by this function here X right. So, this X is called as random variable why this is called as a random variable because when you toss a coin then unless and until the experiment is completed there you do not know whether you are going to get head or tail. So, there is

some randomness involved. So, now, this head is denoted by 1 and tail by 0. So, this sample space Ω has now 2 points 0 and 1 right.

So, you can see here earlier this Ω was here H and T, but now Ω is transformed to a new set of values here this takes value 0 and 1. So, now, in any random experiment we are always interested in the value of some numerical quantity which is determined by the result. We are not really interested in all the details of the experiment I am not really interested that how you have tossed the coin whether it is 5 feet high or 7 feet high or in the 20-degree centigrade or say 100-degree centigrades this is not my concern what I want I just want please give me the outcome head tail, tail head etcetera. And these quantities of interest we are determined by the result of the experiment and these are associated with these random variables right. Now, the because this these values are determined by the outcome of an experiment and we are trying to associate here a random variable whose outcome is not known with 100 percent confidence that if you toss a coin unless and until it comes from the ground you do not know whether it is going to be head or tail both the possibilities are there.

So, suppose we assume or we say that ok there are 50 percent chances a head will come and 50 percent chances the tail will come. So, in that case we can associate the probability with the outcomes of the random variable. So, you can see here that x is taking values here 1 and 0 1 for head. So, if I say that x equal to 1 that means, head is observed and the probability of observing x equal to 1 that is the probability of observing here head this is 1 upon 2 0.5. And similarly, the probability of observing tail that is the random variable x takes value 0 is also here 1 by 2. So, now, we can view x as a random variable which collects the possible outcomes of the random experiment and capture the uncertainty associated with them right. So, this is how we are going to introduce the random variable. So, now, if I try to translate whatever I have explained you. So, now, in short, I can say here that let Ω represent the sample space of the random experiment that what are the possible values it can be observed and let this capital R be the set of real numbers right.

Now, I can define or I can say that a random variable is a function X which assigns to each element a ω belonging to Ω, one and only one number. For example, X_ω is equal to small x small x belongs to R that is X is trying to map Ω to R. So, that you can see here Ω in the earlier example had two points H and T and R had here two points 1 and 0, H is being transported to 1 and T being transported to 0. So, H and T are mapped to 1 and 0 respectively right. So, before I go further with the details of this random variable and other aspect let me try to explain you here what is the convention in statistics that how do we indicate the random variable and its values.

You know that when you are trying to define a random variable then you are going to get the observation on them. For example, suppose if I say that I want to have the observation

on the heights of students. So, these heights are going to be measured for different student in say suppose the computer. So, I can define the heights of the student as a random variable capital X and suppose now I obtain the height of say student number say here 1 and suppose it comes out to be suppose here 100 centimeters. So, this is indicated by here small x and this is the first observation.

So, I can say here this is small x_1 and similarly if I say that height of the student number 2 is suppose 180 centimeters then this can be denoted by here small x_2. So, now what I am trying to do I am trying to indicate the height of the student as capital X in capital alphabets and the associate values in lower case alphabets and this numbers are trying to indicate the observation number this is the first observation second observation and so on. Similarly, in case if I try to suppose I decide that ok let me try to also collect the weights of the students. So, now I can use here another variable here say here capital Y to indicate the weights of the student and suppose we are going to measure them in kilogram. Suppose the weight of the student number 1 comes out to be suppose here 50 kilograms.

So, this will be denoted by here small y_1 and suppose weight of the second student comes out to be 60 kilograms then it is going to be indicated by small y_2. So, if you try to see what I have done I am trying to indicate the random variables by the upper-case alphabets and their values by the lower case alphabets. So, actually this is a convention in statistics to indicate or denote the random values by capital letter alphabets upper case alphabets and their values by small letters say this lower-case alphabets. So, if I try to take the random variable to be here capital X then their values will be small x right. So, that is and if I try to write down here say suppose if I write down here let $x_1, x_2, \ldots, x_{100}$ be the heights of students.

What does this mean? That means, we have got 100 numerical values or we have got the heights of the 100 students which are some number which are some numerical values right. Now, when we are talking about the random variables then we have 2 types of random variable here. One is called as discrete random variable r v means random variable and other is continuous random variable. What are these things? So, let me try to first give you a brief idea and then I will try to give some more examples of these type of random variables. So, when you are trying to consider a random variable which is taking the values only at some points then it is called in simple word as discrete.

For example, if I say the number of heads obtained when you are crossing a coin 100 times. So, you cross a coin 100 times try to count the number of heads once again try to cross a coin 100 times count the number of heads. So, this number is always going to be an integer and it can take all possible values between 1 2 3 4 up to 100. What if I say there are 25.5 heads that is not there. So, all the values they are concentrated only at some point. So, the random variable whose set of possible values can be written as a finite sequence x 1 x 2 x n or as an infinite sequence x 1 x 2 dot dot dot they are said to be

discrete. So, the sample space of such a random variable that is said to be discrete if it consists of a finite or countable set of outcomes right. For example, in the example which I just explained you that you try to cross a coin for 100 times and then try to count the number of heads. In such a case the random variable it is a random variable whose set of possible values is the set of non-negative integers 1 to 100 and this is a discrete random variable.

So, now we consider an example suppose there is a customer here has a phone and which has 30 external lines that means, there are something like 30 telephone connections and different customers are calling to the customer here. So, at a particular time that some of the lines may be used and others may be free. So, at a given point of time suppose 5 customers are calling then 5 of the lines you will get occupied if there are 10 customers which are calling then that 10 or 9 will be occupied and suppose if more than 30 customers are calling at the same time then the all the 30 lines will be used and the customer will be asked to wait right. So, now at a given point of time how many telephone lines are being used that is not known to us.

So, this can be indicated by a random variable X. So, now, if you can see here X can take different between different values say from 1 to 30 that means, only 1 line is being used or 2 lines are used or 30 lines are used. Now, suppose if I say that at a particular time there are 5 telephone lines which are being used that means, the value of the capital X here is 5. So, this is going to be indicated by a small x. I can say here if 5 lines are used then the small x will be equal to 5 and this is how we are going to interpret it ok. Now, I have given you a fair idea about the discrete random variable.

Now, I come to continuous random variable. Suppose the length or temperature fluctuation or calibration cutting to wear, bearing wear etcetera they have to be measured right and they are going to be measured in some units. For example, the temperature fluctuation that can be in degree centigrade or degree Fahrenheit whereas, those cutting to wear that means, when you are trying to use an equipment, it gets distorted. So, how much it has been distorted after being used that can be measured in some units like how that how much material is reduced in grams or volume etcetera. But anyway, you can see that whenever you are trying to measure such observations then what is going to happen? Yeah, first of all there will be small variations in the measurements and these variations may be due to different reasons. For example, just now I had given you an example that if you want to measure the moisture level in the soil in a field and the level of moisture content in different parts of the field it is not going to be same, but there will be small variation.

So, now in such a situation if you try to represent by capital X the quantity which you are going to measure then you cannot say that the random variable is going to take the values only at some finite values. For example, as in the case of discrete we say that if there are

5 telephone lines are going to be used then the value of the random variable is 5 and if 8 lines are going to be used then the value of the random variable is 8. But in the case of moisture level in the soil or say temperature this can be 30 degree centigrade, 30.1 degree centigrade, 30.11 degree centigrade, 30.111 degree centigrade and so on.

So, there will be some random variation and these values are going to be defined as if the random variable is going to assume the values in the form of an interval of real numbers right. So, if I try to consider that for any equipment you are going to measure the length, temperature, fluctuation, calibration etcetera etcetera. This type of model will give us an idea about the precision in the length measurements, but these values are going to be measured on some continuous scale. There will always exist values between two values that is what I mean to say.

So, now, I can say that there exists random variable that take on a continuum of possible values and such variables are known as continuous random variables. And another example I can give you is the for measuring the lifetime of a bulb that can be 1 day, 1 day, 1 hour, 1 day, 1 hour, 1 second etcetera. So, in case if you want to make an assumption on the bulbs lifetime then we can assume that the lifetime of the bulb is lying in some interval say a and b where a and b both are positive right. So, in such a case we have a concept of say this continuous random variable and in both the cases we have considered the univariate case that way there are only there is only one variable. Now, we are going to consider a situation where the process is going to be affected by two random variables.

So, we call them as a bivariate random variable. So, these two random variables they can be independent they may be dependent. What does this mean? The two random variables are going to be independent if the occurrence or non-occurrence of one variable does not affect the occurrence or say non-occurrence of the second variable. And they are dependent that if the occurrence in the first variable affects the outcome in the second variable right. Well, we will understand that how to define such a independence through the random variable in the forthcoming lectures. And both these random variables they can be continuous means both are continuous both are discrete or a combination of them that one of them is continuous and another is discrete.

So, this is the general setup of a bivariate random variable right. For example, in case if I say I want to measure the height and weight of a person. So, this can be indicated by saying that that X_1 indicates the height and X_2 indicate the weight of a person. And what we know that height and weight they are the positive value. So, I can say that the person's height and weight are assumed to take on any value in the interval a_1, b_1 and a_2, b_2 respectively right.

So, height lies between a_1, b_1 and weight lies between a_2, b_2. For example, height can be between suppose say 50 centimeters to suppose 170 centimeters. And the weight can be suppose here 10 kilograms, suppose here 60 kilograms right. So, these are some numerical values. So, now the question is this in statistics how do we express such a bivariate random variable.

So, it is indicated by a 2 cross 1 vector. Now, this is a mathematical vector, vector like vectors and matrices it is not a data vector of the R software. So, we try to indicate here as say here it is a vector of random variable X_1 and X_2 and we write this here as say x underscore is a 2 cross 1 vector or because I can write down here say $X_1\ X_2^T$. And when we are trying to have 2 random variables then some other properties also come into existence. Now, if you try to see I started with a univariate random variable, I extend it to 2 random variable and similarly I can extend to a general setup where we have more than 2 random variables.

And this setup will be called as multivariate random variables. So, suppose there are more than 2 random variables and once again all the random variables they can be independent, they can be dependent and we have a statistical measure by which we can actually judge. And it is possible that all the random variables they can be continuous, they can be discrete or they can be a combination that some are continuous and some are discrete. So, this is the setup of a multivariate random variable. So, say one simple example can be if I try to extend the example of height and weight that suppose I want to measure the different parameters of a human being. So, suppose I want to measure the person's height, weight, age, blood sugar and blood pressure.

So, there are going to be height can be measured by X_1, weight can be measured by X_2, age can be measured by X_3, blood sugar can be measured by X_4 and blood pressure can be measured by X_5. And suppose these all these variables will have some numerical values which are going to be assumed that the height which is indicated by X_1 it is lying between a_1, b_1, second variable X_2 is lying between a_2, b_2, third variable is lying between a_3, b_3, fourth variable is lying between a_4, b_4 and fifth variable is lying between a_5, b_5. And for example, I can assume depending on the situation suppose I assume that all a_i, b_i s they are greater than 0 right. Although I am taking here in example where all the variables, I am assuming to be continuous, but they can be discrete also or it is possible that suppose X_1, X_2, X_3 are continuous and X_4 and X_5 are discrete that is also possible. But here my objective is to give you the idea of what is a multivariate random variable.

So, now all these 5 variables they can be expressed in a vector which is of order 5 by 1 and the way I had expressed X_1 and X_2 similarly I can extend it to X_1, X_2, X_3, X_4, X_5 or this is a transpose of x 1, x 2, x 3, x 4, x 5. Now in case if you want to take the observations suppose if I consider here a bi unit pattern. So, suppose if I say here person number here 1 and say person number here 2 and suppose the height of the person

number 1 comes out to be 150 centimeters and the weight of the person number 1 this comes out to be suppose here 50 kilograms. And for the person number 2 suppose the height comes out to be 160 centimeters and the weight is suppose 60 kilograms. So, now these are the observations which are going to be indicated by X_1 underscore like this and the second observation is going to be like this X_1 underscore right.

Similarly, if you try to consider here a multivariate setup. So, now you will have suppose I have the data of here person number 1 and person number 2. For the person number 1 you will have here a vector of 5 by 1 high weight age blood sugar blood pressure. And similarly for the person number 2 high weight age blood sugar and blood pressure and this will be indicated by here x underscore 1 and the data of the second person will be indicated by x underscore 2. So, that is indicating that these are the observations on the case here 5 variables right. So, now I can now express in general that suppose we have got here p random variables which are indicated by capital X_1 capital X_2 up to here capital X_p.

So, all these variables can be defined in a in the setup of a p x 1 random vector. And this vector can be indicated by here as x underscore and which will be here a p cross 1 vector consisting of value $X_1, X_2, \ldots X_p$ and this can also be written as say here $(X_1, X_2, \ldots X_p)^T$ transpose. Now, in case if you want to get here the first value then it will be like as a first value on X_1 first value on X_2 first value on here to here X_p like I can say $x_{11}, x_{12} \ldots x_{1p}$. And similarly, you can have here the second value third value and so on. So, in such a case the space of X is a random vector will consist of set of n tuples that means n sets of observation if I see here, we are getting here $X_1, \ldots X_p$ like this ok.

So, now we come to an end to this lecture and you can see here that was a pretty elementary lecture, but my objective was to expose you with the concept of a random variables because you will see that in all the statistical tool our first sentence is going to be let X be a random variable that can be a linearity random variable that can be a bivariate random variable or that can be a multivariate random variable. It depends on the requirement and objective of the study that what we really want to do. And depending on the setup that depending on the choice of random variables our statistical tools are going are also going to differ. In case if my random variable is continuous then the way I am going to operate or find out different types of tools for example, mean or even variance that is going to be a different way in comparison to when my random variable is discrete. And suppose if we have a combination then we have to see that how are we going to compute different statistical quantities.

Now depending on the nature of random variable we have different types of characterization of the probability function and based on that all other tools they also change. So, now but at this moment my request to you will be try to have a look in some

books and try to read more about this random variable. Once again, I will say the way I have to explain you the random variable concept in this lecture this is a very elementary one. But my objective was to expose you so that I can take you forward to a multivariate setup. In case if my objective is to give you a course on probability theory or measure theory possibly, I will give you all those basic definitions here.

And believe me we have very concrete solid definitions which are based on hardcore mathematics for random variables and different types of concepts that we are going to use further. But once again I will say that I am not going into the mathematical details well I have limited time here. But I will request you that you try to now start reading a book on any this probability theory or the probability distribution mathematical statistics whatever book you want there are books which are varying from lower level of mathematics to a hardcore mathematics. This is the way by which we both can go together and the learning and in the same time the learning process can make you learn better you will understand more concept. So, you try to read from the book try to understand this concept and I will see you in the next lecture till then goodbye. Thank you.