Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 05

Lecture – 21

Random Variables: Probability Functions

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last lecture we had discussed the concept of random variables. So, we had discussed that there are two types of random variable discrete and continuous and the random variables can be one more than one also. So, based on that we had a concept of univariate random variable, bivariate random variable as well as multivariate random variable. My idea was only to make you familiar and comfortable with the concept of random variables. So, that whenever I am going to use this word it looks comfortable and known familiar word.

Now, I come to another topic. In case if you try to see what this random variable is going to do. This random variable is going to represent the data generating process. Because you are trying to observe the data on random variable.

For example, if I say height then you are trying to get some observation on height of the students and if I say this x is temperature random variable is temperature then you are going to collect the data on temperature and these values have some random variation. So, if you try to see the way data is being generated either on say temperature or any other random variable there is always a process behind the random variable and the data is generated according to that process. Definitely in this world there are so many types of experiments they can be conducted or there are so many process which are going on in the nature and the way they are occurring they are entirely different. The way it rains it is entirely different the way it the way the sun shines. The way the yield of a crop is produced it is very different than the way some machines are produced in a factory.

So, every process may have one or more than one random variable depending on your nature, but behind them there is a process. So, now the question is how to get an idea about that process and we have now understood that unless and until we try to express anything without using the mathematics people are not going to believe on us. So, that is

why now it is very important that if I try to define a random variable then associated with it there is a process. Suppose if I take an example of health suppose x indicates height then height will have a different process by which it works. For example, heights of a human being the increases right from the birth till certain age say age of 21-22, but after that means either they become stabilized or the growth is very little, but the blood pressure blood pressure of the same human being will have a different process that how it varies.

So, now my question to you is that depending on the choice of random variable how are you going to mathematically present this process. Now, it depends whether your random variable is continuous, discrete, univariate, bivariate, multivariate and based on that we need to have a mathematical function which can define the process. Based on that we have a concept of probability density function, probability mass functions and based on that we have different types of concepts in the context of univariate, random variable, bivariate, random variable. And yeah all these concepts actually in statistics they are taught in detail, but here I am not going to explain this concept in that much detail, but my idea is to give you some information a reasonable knowledge so that it is not difficult for you to understand the topics of the multivariate procedure that we are going to discuss later on. So, these topics random variable their density function etcetera they are going to help when we are going to discuss the multivariate procedure that is why I am trying to create a foundation in the beginning.

And once again I would say I am covering here only limited number of topics which are required to understand the topics which I am planning to cover in the forthcoming lectures only. If you want to have more information on these topics my advice to you all is that please try to look into the books on statistics, probability theory, mathematical statistics etcetera and you will get solid foundation and solid knowledge from those books. So, in this lecture let us try to discuss about different types of probability functions and we try to understand them quickly. So, let us begin our lecture now. So, now in this lecture we are going to talk about the different types of probability functions right.

So, first I try to define the concept of probability density function which is shortly expressed as pdf, p here is for probability, d here for density, f here is for function. Now this probability function for a continuous random variable and for the discrete random variable they are different. I mean they are trying to represent a similar concept, but mathematically they are expressed in a in different ways. In the case of continuous random variable this is called as probability density function pdf and in the case of say here discrete random variable this is called here as a probability mass function or say here pmf that is a popular name. So, now let me try to give you the idea of this pdf and pmf one by one.

So, now I am going to consider first the case of continuous random variable and in the in this case we have a probability density function. As the name suggests probability density function or probability functions that means how the probability of the values of of random variables are going to behave. For example, if you recall in the last lecture we had taken an example of a coin crossing where the outcomes were head or tail and their probabilities were one by two and one by two right. So, with the values of the random variable there is always an associated probability and this probability functions are going to explain as the behavior of the probability as the values of the random variable change. So, in the case of continuous random variable we try to define a probability density function by a mathematical function.

So, so for a function f(x) if it is to be a probability density function of a continuous random variable x it needs to satisfy the following two conditions. f(x) should be greater than 0 for all values of x that means at every value of this random variable this value f(x) has to be greater than or equal to 0 and the integral over the whole range of x should be equal to 1. So, here I am trying to write an integral $\int_{-\infty}^{\infty} f(x)dx = 1$. So, I am assuming that my random variable is taking value between minus infinity and plus infinity, but in case if x is trying to take the values between here a and here b greater than a and less than b then this integral will become here a and a to b right. So, these are the two condition which had to be satisfied, but before we try to go into the mathematical aspect let me try to give you here an idea that how are you going to understand this function.

Because in practice if you try to see nobody is going to come and explain you that ok in this case this is going to follow the normal probability density function in this case it is going to follow a gamma density function etcetera etcetera. This is only you who has to look at that data and you have to take a call and decision that which probability density function is to be used in this case so that you can draw the correct statistical inference and if you try to change the probability density function the way you are trying to draw the statistical inference will also change. So, now in practice how are you going to get a fair idea about the probability density function right ok. Couple of lectures back if you recall that we had discussed the histogram and in histogram what do we do that we have a data set and we try to classify them into different intervals and the frequency the absolute frequency or the relative frequency in those interval is plotted on the graph and which gives us the histogram. So, if you try to look here for example, if I try to see here in this if you try to look here in this histogram suppose if I take here this is here your x and here it is f(x) right.

So, if you try to create this histogram and if you try to take here the midpoint of each of the bar of the histogram like this here and if you try to join these points by a smooth curve means you start drawing the curve by putting a pen on the paper and draw it without lifting your pen. So, if I try to and then try to join this midpoint. So, if I try to

gradually draw it will look something like this here you can see here like this and this curve can be viewed that if the width of these bins or this bars is reduced then you can imagine that as the width of the bar is converging towards 0 it is tending towards 0 then what will happen all these midpoints will start look like here these dots and you are simply trying to join this actually these dots here and this is what I have represented here in this line you can see. So, this is actually a probability curve and this curve is being indicated by the probability density function when we are trying to handle a real set of data right. So, this is I am trying to explain you that how are you going to look at this probability density function from a data and then by this curve you can have a fair idea ok.

This curve resemble with the probability density curve of this particular distribution. So, I can start working with that distribution and if you try to see there is a one to one correspondence between the relative frequency and the probability. So, if you try to see in a particular say this interval there is some frequency and what we are trying to plot here this is essentially here the relative frequency. So, relative frequency is the total number of points which are lying in this bin or in the inside this bar or the total number of points which are lying in this interval divided by the total number of observations and if you try to see or if you try to recall the definition of probability which you studied in your elementary classes it was the total number of favorable point divided by the total number of points right. So, that is why this midpoints or these heights they are going to indicate the relative frequency which in turn is indicating the probability of the event right.

So, this is how we in practice we try to approximate the probability density function from a histogram right. So, as I said the relative frequency is an estimate of the probability that a measurement falls in the interval. For example, if this is here like this is here the relative frequency then this is an estimate of the probability and we know what is the total sum of relative frequency this is always 1. So, that is why the total area under this here curve this will always be equal to 1. And so, in case if you want to estimate the probability of an event what you have to do you have to simply find out the area of the curve which is lying under that interval.

Suppose if I want to find out what is the probability of an event falling in this range A to B. So, basically I have to calculate this area of these 3 bars. So, this is one to one means relation between the probability theory and the relative frequency. And this I am trying to explain you because whenever you are trying to work with the multivariate procedures. So, usually there will be more than one variable and yeah then you will have to get some idea that what is the probability distribution of this random variable and based on that your these computations are going to depend upon.

So, that is why by plotting such a data you can have a fair idea. After this concept of probability density function we have a concept of cumulative distribution function which

is indicated by here CDF, C for cumulative, D for distribution and F for function right. So, this actually here the sort of total probability up to certain point right. So, this CDF or cumulative distribution function or it is popularly called as only distribution function it is indicated by capital F. So, this capital F or the CDF of a random variable X is defined for any real number small x as the probability that capital X is less than equal to x right.

So, you can see here suppose if I try to show you here in this curve also. Suppose if I try to clear this screen and if I try to show you here that probability up to this particular point this is here the this area. Suppose it is my here point number A. So, I can see here this is the probability that X less than equal to A. So, this will indicate the value of CDF at point *a* right and I can write in down here like this.

And similarly if I try to take here the area up to here this particular point. Suppose this point here is my here *b*. So, if I want to cover this area then this is going to be nothing, but probability that X is less than equal to B and this is going to be the value of the CDF of random variable X at point *b*. So, this is exactly the same concept which is being explained here probability that X is less than equal to small x and this indicated by here F of X right. Sometime more specifically we write in down here $F_X(x)$ right.

So, this $F_X(x) = P(X \leq x)$. So, this is the concept of CDF right. So, in case if you want to find out the CDF for a continuous random variable then how to get it done you can know you know that in for example, in this case if you want to find out such areas this areas can be obtained by the integration right. And based on that we also have a definition that when do we say that our random variable is discrete or continuous. So, we say that a random variable X is continuous if there is a function small f(x) such that for all $x \in R$, $F_X(x) = \int_{-\infty}^{x} f(t)dt$ right.

So, basically your here F here is the PDF in the case of continuous random variable. So, you are simply trying to integrate from minus infinity to a particular point and then it will give you here the value of the probability up to the point is small x and this is simply your here CDF. So, F(x) is here the CDF of x and small f(x) is the probability density function of x and in case if you are if you know the CDF then you can differentiate it at a small x or say differentiate it with respect to x and then you will get here the value of PDF and this can be obtained for all x that are the continuity points of x. So, this is the mathematical part I just wanted to give you some idea so that you understand that there is always a mathematical concept which is associated with the concept which I am trying to explain based on the data right. And now this PDF and CDF they can be used in different ways for example, whenever you are trying to conduct any statistical analysis you are interested in finding out different types of probabilities.

So, if you know the PDF then also you can find out different probabilities if you know the CDF then also you can find different types of probabilities. So, just for the sake of

your convenience I try I am trying to give you here one result that if a random variable x has a CDF capital X and suppose if there are two points small x_1 and small x_2 such that x_1 is less than x_2 there are some known values known constant and if you want to find out the probability that capital X random variable lies between small x_1 and small x_2 value then this can be obtained here as a difference of the CDF at x_2 and x_1 that is $F(x_2) - F(x_1)$ right. And this can be expressed as a integral say small x_1 to small x_2 f(x)dx right. And definitely one very important result for you to understand that in the case of a continuous random variable the probability at particular point is 0 right. This can be obtained here like this probability of $P(x = x_0)$ that is the integral $\int_{x_0}^{x_0} f(x)dx$ which is equal to 0 right.

Now in practice you will always be confused that how to get the observation on the continuous random variable. So, you see in practice we are always getting the observation which are discrete, but it depends on the nature of the variable that decide to take it as continuous or not right. And now let me try to give you one simple example to explain you the utility of this probability density function. So, for example, there is some continuous random variable X and whose probability density function is given by here like this

$$f(x) = \begin{cases} \exp(-x), x \geq 0 \\ 0, x < 0 \end{cases}$$

And suppose if you want to find out the probability that capital $a < X < b$ then this can be obtained here as a integral $\int_a^b f(x)dx$.

And if you try to plot it then you will see here this curve of exponential of minus x is here like this you can see here. And if you want to find out here that x is lying between a and b suppose this is here point a and this is here point b. So, this is the shaded area which is going to give you this integral or this probability right. Now let me try to give one more example so that I can show you that how this probability functions are formulated and how they are used. Well, I am trying to take a very simple example so that you can understand them.

In practice the examples are not usually going to be so simple. Suppose I want to do some analysis about the waiting time for the train that different persons arrive at the metro railway station and suppose the metro railways or the metro trains which are frequently going and I want to know that how much time a person has to wait before the train arrives right. So, suppose my random variable is the waiting time for the train and suppose the train arrives after every 20 minutes. So, what will happen? If a person arrives in the at the station the person has to wait between 0 and 20 minutes. If the train is already there on the platform the person can catch it immediately and the waiting time will be 0.

But suppose the person arrives at the platform and when the train is just leaving then the person has to wait for 20 minutes right. So, this waiting time of a particular person is a random value or a random variable which is trying to take the value or any value which are contained in the interval 0 to 20. So, suppose I can say here that we can think that ok there can be a mathematical function f(x) which can take a value here k if x is lying between 0 and 20 and 0 otherwise. And this f(x) we can through this f(x) we can design the probability density function and k is some unknown constant. Another question is this if the k is unknown how are you going to use it.

So, now we try to utilize the properties of probability density function and we try to find the value of k so that in turn we can have a probability density function which can be used for further application. So, we know from the probability of from the property of probability density function that the integration over the whole range of x should be equal to 1. So, I try to make it here $\int_0^{20} f(x)dx = 1$ and if you try to solve it this will come out to be 20 k=1 this implies that $k = 1/20$. So, now my pdf is given by here like this

$$f(x) = \begin{cases} \frac{1}{20}, 0 \leq x < 20 \\ 0, otherwise \end{cases}$$

Now you have got here this a PDF and now you can use it for different jobs for computing different types of probabilities right.

Now before I try to move forward let me try to also give you an idea here how are you going to find out the CDF from a given PDF. So, if you try to see here the CDF $F(x) = \int_0^x f(t)dt$ and so now here it is f(t) here is 1/20 dt and if you try to solve it this is coming out to be coming out to be here x/20. So, now suppose we are interested in calculating the probability of a waiting time between 15 and 20 minutes. So, now that means, x is lying between 15 and 20 which can be written here as a $F(20) - F(15)$ and now using this value I can write down here directly (20-15)/20 which is 0.25.

So, that means, if a person arrives at the railway station then there are 25 percent chances that the person will have to wait between 15 and 20 minutes right. So, now after this continuous random variable let me try to give you some idea about the probability mass function also. So, the probability mass function is defined for a discrete random variable right. So, we know that the probability distribution of a random variable x is a description of the probabilities associated with the possible values of x. The difference between continuous and discrete random variable is that the continuous random variable is trying to take the value in an interval whereas, the discrete random variable takes the value at points.

So, that is why for a discrete random variable the distribution of probability is specified by a list of the possible values and in certain cases it is easier for us to find out the mathematical formula which can describe different probabilities that one can compute different types of probabilities and in some cases we try to define it only by giving the value of the random variable and the corresponding probability right. So, suppose there is a discrete random variable x which takes k different values. Now in this case the probability function is called a probability mass function or say pmf. So, p is coming from probability m from mass and f from function. So, the probability mass function of x is given by here probability that x takes the value x_i which is equal to suppose here p_i for each $i = 1,2, \dots k$.

So, for x equal to x1 the probability is p1 for x equal to x2 the probability is p2 and so on. So, in this case also similar to the two conditions in the case of probability density function similar conditions are here for the probability mass functions also. So, we assume that each value of the pi is between 0 and 1 and the sum over the whole sample space of this pi is equal to 1 right. So, these are the two condition for a function to become a probability mass function and similarly in the case of a discrete random variable also we can define the cumulative distribution function like this F(x) will be equal to summation i goes from say p i right and it is here an indicator function means we will take the value 1 if xi is less than equal to x and 0 otherwise. So, the difference between the CDF of discrete and continuous random variable is that in the case of discrete random variable the CDF is always a step function like this one this this this like this whereas, in the case of a continuous random variable this will always be a continuous increasing curve like this thing right.

And then yeah means I am not giving you here different properties of the CDF because that I am not going to consider in the further lectures. So, but my strong recommendation to you all is that please try to look into the book and try to at least read the basic concept of PDA PMF and the cumulative distribution function for discrete and continuous random variable right. So, now we come to an end to this lecture and now you can see that we have considered here the probability function for discrete and continuous random variable. But in this case we have taken only the univariate case and now these definitions can be extended to bivariate tri variate or in general a multivariate random variable also. And the same concept probability density function probability mass function the way they have been defined here they will be extended to bivariate or say higher order random variables.

And if you try to think how to get it done for example, if you want to extend the definition of probability density function. So, you have taken here a condition like as integral $\int_{-\infty}^{\infty} f(x)dx = 1$. Now if you try to extend this suppose you have more than one variable x and y now you will have double integral over x integral over y and the

probability density function will also involve two random variables. So, instead of f(x) this will become f(x,y) and when you are trying to integrate that will become double integral f(x,y) dx dy. And this can be extended forward extended to a p cross 1 random variable also that you have p integrations $f(x_1, x_2, \ldots x_p)dx_1\, dx_2 \ldots dx_p$.

So, my basic objective in this lecture and in the forthcoming lecture is not to introduce you with the univariate random variable or univariate probability distributions. But my objective is that I want to extend this concept to a multivariate setup. In this lecture I have considered only the scalars when we are trying to consider the bivariate and multivariate then we are going to consider the vectors and matrices. As I said I have given you here a very brief introduction to the concepts like PDF, PMF or say CDFs. They are not sufficient to understand the entire probability theory.

But as you know that I have also a limitation of time in this course. So, my request will be that you please try to pick up a good book and try to read this concept and believe me once you understand the basic concept I can promise you it will be a very easy job for you to understand this lecture and finding out different types of probabilities, finding understanding probability with distribution functions etcetera. And there is a long list of probability density functions and probability mass function which have been defined in the statistics. But surely we are not going to do all of them here I will try to do here only those PDFs which I will be needing in the forthcoming lectures. But if you want to expand your knowledge you need to look into the books.

So, you try to look into the books, try to take some examples from the book, try to solve them in theory and I will see you in the next lecture where I will try to explain you with the bivariate and multivariate concepts related to the random variable. So, you try to practice, you try to read and I will see you in the next lecture till then good bye. Thank you.