Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 06

Lecture – 25

Bivariate Normal and Multivariate Normal Distributions in R

Hello friends, welcome to the course multivariate procedures with R. You can recall that in the last lecture we had talked about the normal distribution. First, we discussed the theoretical properties and then in the last lecture we had seen its implementation in the R software. But in case if you want to extend it to more than one variable that means from univariate to say bivariate normal or in case if you have more than two random variables then multivariate normal then how to get it done. And as we have discussed many time that in real life most of the real data is multivariate and there we will need the assumptions of multivariate normal. So, my objective in this lecture and in the past two lectures was to begin with a univariate normal give you a fair idea how are you going to do it and then now finally extend it to a multivariate setup.

So, this is what I am going to do in this lecture now. I am assuming that you have understood the univariate normal distribution properly and based on that I will try to give you the probability density function of bivariate normal distribution and there I will try to show you that how the number of parameters get increased. For example, in the case of univariate normal distribution you have only two parameters mean and variance. But now in the case of bivariate you have two variables.

So, for each variable you have mean variance that means four parameters. But there is one more that is the interaction between the two variable x and y that is the joint effect of the random variable that also has to be incorporated. So that can be incorporated through how? If you recall we had discussed the concept of covariance, correlation coefficient. So now one parameter related to either covariance or correlation coefficient will also get introduced. Now in case if you extend it to a multivariate setup then each of the variable will have its mean, variance as well as their joint variations.

They will come into existence. So, this is how we are going to learn this lecture. But before starting the lecture, I would like to clarify this bivariate normal and multivariate normal I am going to discuss here only their properties. But they have a very strong mathematical background and in the course of multivariate analysis we try to give all type of mathematical derivation that how to obtain those things like as marginal distribution, conditional distribution, their mean, their variances etc. But here I am not going to do it.

My objective is this. I am trying to do here the multivariate analysis only that much amount which I require in the lectures in future. So let us begin this lecture and try to understand about bivariate normal, multivariate normal along with their application in the R software. So let us begin our lecture. So now if you try to recall we had considered the univariate normal say x following normal mu sigma square and its probability density function say is given as 1 over sigma root 2 pi exponential of here minus 1 over 2x minus mu upon sigma whole square right where this x and mu they lie between minus infinity and plus infinity and sigma square is greater than 0 and both mu and sigma square they are the parameters.

So parameters are those things if you know them then you know the complete details about the probability distribution. So now we try to extend it to a bivariate case that means there are two variables. Suppose I can see here there is another variable here y which is following here normal mu sigma square. They may have same mean, different mean but now we also want that we want to consider their joint effect right. So now the extension of a normal distribution to two random variable is bivariate normal distribution and the probability density function of a bivariate normal distribution is given here like this.

There are two random variables x and y. So we have here now the parameters here sigma x, sigma y, mu x, mu y and rho and it is given here like this. It looks like normal but it is different because there are couple of things more. So because there is here one parameter here which is rho. So as you have learned that mu indicates the mean, sigma square indicates the variance.

Similarly rho indicates the correlation coefficient. This I will try to show you later but this is how the joint effect enters into the normal distribution. So this PDF is given by 1 upon 2 pi sigma x, sigma y square root of 1 minus rho square, exponential of minus 1/2, 1 minus rho square, x minus mu x whole square upon sigma x square, y minus mu y whole square upon sigma y square minus twice of rho, x minus mu x, y minus mu y, divided by sigma x sigma y, right. Where x and y both lies between minus infinity and plus infinity, mu x mu y they are also lying between minus infinity and plus infinity and you can see here the parameters here are mu x mu y sigma x sigma y and here rho, right. So, sigma x and sigma y square they are lying between 0 and infinity and rho is the correlation coefficient so it is lying between minus 1 and plus 1.

$$f_{XY}(x, y, \sigma_X, \sigma_Y, \mu_X, \mu_Y, \rho)$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}, -\infty < x < \infty, -\infty < y < \infty$$

Parameters: $-\infty < \mu_X < \infty, -\infty < \mu_Y < \infty, \sigma_X > 0, \sigma_Y > 0,$

$$-1 < \rho < 1.$$

So this is how we try to define this bivariate normal distribution function. Now I am giving you a small exercise. Try to substitute here rho is equal to 0 and then try to see what happens. Whatever you will get that you can see this f of x y this can be represented here as say f of x and f of y. Please try to do it yourself and if you recall what this mean that is indicating that the two random variables x and y are independent and if you try to recall when we understood the concept of stochastic independence then we had seen that if you want to judge whether two random variables are stochastically independent of each other then we have to check whether their joint density function f x y can be expressed as their marginal density function f x into f y.

$$f_X(x) \sim N(\mu_X, \sigma_X^2)$$

$$f_Y(y) \sim N(\mu_Y, \sigma_Y^2)$$

If you try to see if rho becomes 0 that means they are independent and this joint density function can be expressed as the product of marginal of x and marginal of y which are univariate normal distribution. But anyway this is only for your exercise. Now in this bivariate normal distribution in case if I try to find out the marginal distributions well you know how to find out them f of x will be minus infinity to infinity f of XY x y d y and if you try to solve it this will come out to be normal mu x sigma x square. And similarly if you try to find out the marginal density of y this will come out to be here normal distribution with mean mu y and variance sigma y square. And here similarly you can find out the conditional distribution also like as f of x given y and f of y given x anyway I will try to handle it in the multivariate case I will show you alright.

This I am trying to give you only some idea how can you extend the concept of univariate into say bivariate and then to multivariate. Now you can see here if I try to find out here the covariance between x and y in this case this will come out to be rho into sigma x into sigma y. So in case if you try to find out the correlation coefficient in the case of bivariate normal distribution this can be obtained as a covariance of x y divided by square root of variance of x and variance of y that is standard deviation of x and standard deviation of y. That is the same definition if you recall we had done earlier when we were trying to do the concepts in descriptive statistics. So now this covariance of x y can be written here as a rho into sigma x sigma y and in the denominator these are the standard deviation sigma x sigma y both get cancelled and this comes out to be here rho.

So now if you try to see earlier we did not knew what is rho but now we have seen that rho is representing the correlation coefficient right and in case if rho is equal to here 0 then you can verify that f of x y can be expressed as the product of f of x and f of y right. So this x and y are independent. Now the same concept will be extended to a multivariate case also. So now let us try to extend this bivariate case to a multivariate case where we are considering that there are p random variables something like x1, x2, … xp. Now this p random variables can be expressed in the form of a random vector.

Vector is the vector of vectors and matrices. So I try to express this p variables x1, x2, …xp here as a x underscore which is here like this right and the space of x is going to be the set of n tuples. Now let me try to show you here that in case if I say that x follows normal mu sigma then what are the changes? Well I am going pretty slow because I am not giving you the mathematical details but I want to explain you right. So now you can see here mu here is a mean vector which is here like this right and what is here sigma. In the case of bivariate you can extend this covariance matrix as say here sigma x square sigma y square and here say covariance between x and y which is rho sigma x sigma y like this right.

$$\frac{\boldsymbol{Cov(X,Y)}}{\sqrt{\boldsymbol{Var(X)Var(Y)}}} = \frac{\boldsymbol{\rho\sigma_X\sigma_Y}}{\boldsymbol{\sigma_X\sigma_Y}} = \boldsymbol{\rho}$$

This is indicated here as a sigma x y. So this covariance matrix can be written here as a sigma x square sigma y square sigma x y, sigma y x. Now I try to extend it to a multivariate case. The covariance matrix here sigma can be expressed here as a sigma 1 square which is the variance of x1 then first you try to look at the diagonal elements. So, sigma 2 square which is the variance of x2 then we have here variance of xp sigma y square p.

$$\underline{\boldsymbol{X}} = \begin{pmatrix} \underline{\boldsymbol{X}}_1 \\ \underline{\boldsymbol{X}}_2 \end{pmatrix} \sim \boldsymbol{N_p(\underline{\mu}, \Sigma)} \quad \textbf{where} \quad \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots\sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots\sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \ \vdots \\ \sigma_{p1} & \sigma_{p3} & \sigma_{p3} & \cdots\sigma_p^2 \end{pmatrix}$$

Now on the off-diagonal elements we are trying to consider here the covariances. So sigma 12 is the covariance between x1 and x2, sigma 1 3 is the covariance between x1 and x3 and sigma 1 p is the covariance between x1 and xp. And similarly, if you try to consider the second-row sigma 2 3 is the covariance between x2 and x3 and similarly here sigma 2p is the covariance between x2 and xp right. And definitely because you know that the covariance between x y is the same as covariance between y and x right. So, the lower diagonal elements like this one this is going to be the same as the upper diagonal elements that means sigma 12 will be the same as sigma 21, sigma 13 will be equal to the same as sigma 31 and so on.

So, this is how we try to create the sigma matrix. Now in case if you want to write down the probability density function of this normal p this, I can write down here as say 1 over 2 pi raise to power here p by 2 determinant of sigma raise to power of 1 by 2 exponential of minus 1 over 2 x minus mu transpose sigma inverse x minus mu like this right. This is the pdf of multivariate normal distribution of a random vector x which is of order p cross 1 well you can find these expressions in the book that is not a big deal right. And there is a proper proof and derivation that how you obtain such a thing, but anyway I am not going into that correction. So, this is about the this is how this multivariate normal distribution is given.

Now what I try to do is the following means I try to clear here this writing so that I can show you it here more clearly. Now I try to partition it into two sub vectors. You can see here now I am writing in the blue color so you simply have to look into the blue color. So, x is now being partitioned into two variables sub vectors x1 and x2 it is here like this. You can see here that if I try to remove this thing and I am it is something like this somewhere I am trying to write this say x1 here see here I can write down here say x cube and then after that x cube plus 1 up to here xb right like this x1 x2 x cube x cube plus 1 up to x cube plus 2 up to here xb and then I try to divide it here.

So this is your here x1 sub vector and this is your here x2 sub vector. So this is what I am trying to do here. Now once you try to do this type of division then your mean vector and covariance matrix they have to be suitably partitioned. So I am assuming here that x1 sub vector has a order q cross 1 and this x2 vector is of going to be of order p minus q right and now so you try to partition this mu here like this somewhere here. And similarly this covariance matrix will also be partitioned all the variables in the sub vector x1 that means x1 x2 x cube they will be here like this.

So, the covariance matrix of the sub vector x1 will be here like this it is the covariance matrix of x1. The covariance matrix of the sub vector x2 this will be your here like this and this is here the covariance between the sub vectors x1 and x2 and this is here the covariance between sub vector x bar 2 and x bar 1 right. Now what will be the order of sigma 11? Sigma 11 has q elements. So its order will be q by q. Sigma 22 is based on x bar x underscore 2 sub vector which has P minus q elements so the order of the sigma 22 will be P minus q and similarly this sigma 12 will be q into p minus q.

Orders of $\underline{\boldsymbol{X}}_{\boldsymbol{1}}: \boldsymbol{q \times 1}, \underline{\boldsymbol{X}}_{\boldsymbol{2}}: \boldsymbol{(p-q) \times 1}$

Orders of $\boldsymbol{\Sigma_{11}}: \boldsymbol{q \times q}\,, \boldsymbol{\Sigma_{22}}: \boldsymbol{(p-q) \times (p-q)}, \boldsymbol{\Sigma_{12}}: \boldsymbol{q \times (p-q)}$

Now as we had seen earlier that in the case of univariate normal distribution that mu was the mean and sigma square was the variance. Similarly in the case of multivariate normal distribution this mean vector mu actually is the mean vector. So you can see this is a direct extension of the univariate case to a multivariate case and sigma is here the covariance matrix of here x right. So you can also see that this is also a sort of extension of the bivariate normal distribution to multivariate. In the case of bivariate there was only one of diagonal elements sigma xy but now you have pair wise covariance, covariances for all the variables.

Now this sigma is a matrix so we have to assume that it is a positive definite matrix. It is non-singular and it is a symmetric matrix. Non-singular you know that the determinant of the sigma is not equal to 0 and symmetric means that the diagonal and off-diagonal elements are the same and positive definite means if you try to take a quadratic form something like x transpose sigma x then this is going to be greater than 0. Well, I am not going into the details of this definition because I believe that you are aware of these concept from the matrix theory. But symbolically we write this positive definite as sigma greater than 0.

So, wherever I simply write this matrix greater than 0 this indicates that the sigma or that matrix is assumed to be a positive definite matrix. Now as you had found the marginal distributions in the case of bivariate normal distribution similarly in the case of multivariate normal distribution you can also find out the marginal distribution. I am not going to give you here the proof or derivation but I am simply showing you here the direct result that in case if you try to recall that in the case of bivariate normal this f x y had a marginal f x and f y and f x was normal mu x sigma x square and f y was normal mu y sigma y square. So in the case of multivariate normal distribution the same thing extends that each of xi will have a normal mu i sigma i square for each i goes from 1 to p. Now similarly if you try to look at this random sub vectors say x1 underscore and x2 underscore the marginal distribution of this sub vectors are found to be normal the order here is q and the mean vector here is mu 1 and covariance matrix sigma 11 and similarly in the case of x2 this is normal p minus q that is the dimension of the multivariate normal distribution and its mean vector is mu 2 and covariance matrix here is sigma 22.

So, you can see here whatever you had done in the case of bivariate normal that if you have f x y then the marginal are say univariate normal. So now the same thing is happening here that if you try to create here two sub matrices then the sub matrices are also distributed as multivariate normal and their mean vector and covariance matrix they are suitably divided. Now similarly if you try to find out the conditional distributions you can recall that when we did this concept we had denoted the x given y equal to y right here like this it was something like here f of x y and then it was the marginal of here y and so on. So the similar concept is used here in the case of sub vectors also and if I try to consider here the same set of that x has been partitioned into two sub vectors x1 and x2 then the conditional distribution of x1 given x2 it is here obtained here like this multivariate normal distribution the order of the distribution here is Q and the mean vector here is mu 1, but it has one more term mu 1 plus sigma 12 sigma 22 inverse x2 minus mu 2 right and you can see here because it is conditional so this x2 is actually here known and the covariance matrix is coming out to be in this format sigma 11 minus sigma 12 sigma 22 inverse sigma 21. In many books you will find that there is a standard notation for this expression it is sigma 11.2

So, the way it is written it is something like this suppose if I write here sigma 11 dot 2 so first you have to write sigma 11 then you have to write minus and then you have to write this sigma 12 sigma 12 once again. So what we try to do here I try to write down here sigma 12 because the order has to match with the sigma 11 because sigma 11 this is sigma 12 now there is here sigma 22 inverse because the orders have to match and then it has to be here sigma 21 so that the orders match here. So this is how this symbol is given so but you can see here that this new mean vector this can be called here as say mu 1 star and anyway this covariance matrix is already renamed as sigma 11 dot 2. So you can see here that the conditional distribution of the sub vector x1 given another sub vector x2 is again a multivariate normal distribution with certain mean vector and certain covariance matrices and the same thing happens when you try to consider the conditional distribution of x2 given x1. So this comes out to be here the mean vector here is mu 2 plus sigma 2 1 sigma 11 whole inverse x1 minus mu 1 actually this can be just obtained just by interchanging the 1 and 2 here and here and similarly if you try to see the covariance matrix will come out to be here like this.

Now if I ask you how can you express this covariance matrix in the symbol like you have used here sigma 11 dot 2 you can see here this can be written as here as sigma 22 dot 1 that is a standard symbol. So now you can see here if I try to write down this mean vector here as say mu 2 star then once again I can say that the conditional distribution of sub vector x2 given sub vector x1 is again multivariate normal distribution with mean vector mu 2 star and covariance matrix sigma 22 dot 1. So these type of different probability distributions can be found without any problem. Now let us try to understand that how are we going to work in the R software with this multivariate normal distribution. Now you can see that bivariate normal distribution is also a particular case of multivariate normal and even including the univariate also.

In order to understand or implement the multivariate normal distribution in the R exactly in the same way as we had considered in the case of univariate normal, but in univariate normal we do not need any additional package and that is available in the base package, but for multivariate normal distribution you need to have here another package which is mvtnorm. So, you need to install this package first by using the command install dot packages and within parenthesis within double quotes mvtnorm and then you have to load this library mvtnorm. So, this function provides the density function and a random number generator for the multivariate normal distribution with mean equal to like here is here mean and covariance matrix is equal to sigma. So this is what you have to write down here. For example you remember that in the case of univariate normal density function you had used the command like dnorm which was used to create density and rnorm to create the random numbers to generate the random numbers.

```
install.packages("mvtnorm")
library(mvtnorm)
```

Similarly here also we have the command here dmvnorm then here this the values of the x data vector what you want and then here the mean for example you can give it anything here but just to be in the safe side it is saying rep 0 to p that means you try to repeat 0, p times and sigma is the here the diagonal matrix you can see here. Well you can choose it anything whatever you want, but just for the sake of simplicity I am trying to take it here sigma 1, sigma 2, sigma p. And if you want to generate the random numbers from the multivariate normal then you have to use the command here rmvnorm and then n here is the number of data vector that you want to generate and again this mean here is given like this, sigma here is given here like this. Anyway these are your choice actually because you need to define the mean vector and covariance matrix and you already have learnt in the topic of matrices theory that how can you write different types of matrices in the R software.

```
dmvnorm(x, mean = rep(0, p), sigma = diag(p))
rmvnorm(n, mean = rep(0, nrow(sigma)), sigma =
diag(length(mean)))
```

So, the same thing you have to use here anyway. So these are the means options which I just explained you x is the vector of matrices of quantiles, n is the number of observations, mean is the mean vector and sigma is the covariance matrix and the default sigma is taken here as a diagonal matrix of the ncol(x), ncol you recall it was the number of columns in the matrix x which is p actually in our case right. And mean vector also this is here length ncol(x) that means whatever is the number of columns. If you have x1, x2, x3 p random variable then your mean vector has mu1, mu2, mup right. Now let me try to give you here some examples. So first I try to load this library using the command library mvtnorm then suppose I try to generate here two data vectors well I am not taking here a large number of data vector because of the limitation on of the space on my slides otherwise the font size will become very small and then it will be difficult for you to understand and follow.

So, I try to use here the command rmvnorm n is equal to 2 that means I want to have two data vectors and I want that this data vector x and y is like normal bivariate normal with mean vector here 10 and 20. So I am giving here mean is equal to c 10, 20 and the covariance matrix I want here this is a diagonal matrix say like as here 2, 3 and here 0, 0 and once I try to generate this thing you can see here that this type of observation is produced. Now how to interpret it? Now if you try to see every this column well there are going to be total four number of values why because two random numbers on two variables each right. So you can see here every this here row that is going to express one random number. So you want here two sets of random number from bivariate normal so there are going to be here two rows which are indicating the observation.

```
rmvnorm(n=2,mean=c(10,20),sigma=diag(c(2,3)))
         [,1]     [,2]
[1,] 8.422677 19.06184
[2,] 9.828386 19.02789
```

For example, this observation is on say here x this observation is on y in the first data vector and in the second data vector this is the observation on here x and this observation is on y. Now I try to repeat the same thing for by generating the five observations. So I simply have to change here n equal to 5 and remaining parameter I make the same so that there is no confusion. So you can see here 1, 2, 3, 4 and here 5, these 5 sets of data vectors they are generated right and this is here the screen shot of the same operation. But now let me try to show you these things on the R console also so you that so that you get convinced right.

```
rmvnorm(n=5,mean=c(10,20),sigma=diag(c(2,3)))
          [,1]     [,2]
[1,]  9.990338 16.13789
[2,] 11.353340 21.85558
[3,] 11.941379 20.86491
[4,] 10.659196 18.39344
[5,]  9.830446 21.05324
```

So I try to now load this library here I already have installed this package on my computer so you can see it very easily here and now I try to say here use this command here and control c and then you can see here it is like this right. So now you have got here two sets of random numbers well they will not match with my whether numbers on my slide because they are random and every time you generate they will give you a different value. Similarly if you try to change here the number you want to generate suppose you want to generate here 10 numbers. So you can see here that now there are here means 10 rows. Now suppose if you want to generate here some random numbers from this 4 cross 1 data vector.

Suppose I try to make now the mean vector as 10, 20, 30, 40 and suppose I try to make the diagonal matrix also of 4 by 4. So now you can see here suppose if I make it here it will come out to be like this. So now there are here 4 observations on the first data vector, 4 observation on the second data vector, 4 observation on the third data vector and similarly 4 observation on the 10th data vector and if you try to repeat it here you can see here the values will be different you will not get the same value. Well if you get that means the concept of random is violated. So now we come to an end to this lecture and you can see here what we have now done.

We wanted to reach to this multivariate normal distribution but I believe if I give you it here directly then it will become little bit complicated for you to understand that is why I have taken a little bit longer route but definitely we are going to use each and everything. We have considered the univariate normal distribution so you will see later on that we are going to use it and finally we have come to the multivariate normal distribution and definitely we are going to use it in the further lectures. Well I would like to address once again that I have considered only the normal distribution here but there are many univariate distribution multivariate distributions which I have not considered here but in case if you wish there are books on multivariate distribution where they try to compile different probability density function probability mass functions and also we have not done any mathematical proof because this is not the objective of this course but we are trying to do it because these things are going to create a foundation for the topics which we are going to handle in the future For example you can see here now I am using the concept like matrix, diagonal which I had given in the earlier lectures. So now I would say that you please try to understand this concept.

Try to write down the sub vectors and sub matrices yourself with appropriate order. Whenever you are trying to deal with the multivariate analysis with vectors and matrices one common problem is the which occurs including the programming that is the order of the vectors and matrices. So that you have to learn how to give it properly correctly because only then the R will understand it but your concept should be clear that why you are giving Q by Q or P minus Q by Q and then I have given you here the result but if you have a good background in mathematical statistics you can go through with the books and try to find out that how these expressions have been found and then try to see how you can express them in the multivariate normal in R. One more important point to understand. I have shown you in the R software only the multivariate normal distribution with some mean vector mu and covariance matrix sigma.

I have not considered here marginal distributions or conditional distribution and the question comes how are you going to generate the random numbers for them. Well, that is pretty straightforward. If you try to see the marginal distributions are again multivariate normal. You simply have to change your mean vector and covariance matrices and similarly in the case of conditional distributions you have to change only the mean vector and covariance matrix instead of mu. For example, we have used the symbol mu star say mu 1 star or mu 2 star and instead of covariance matrix sigma you have to use sigma 1 1 dot 2 or sigma 2 2 dot 1 and they themselves are proper matrices.

So there is absolutely no issue. Wherever you want whenever you need you can generate the random numbers on the marginal as well as conditional distributions and this is what you have to keep in mind. So I request you that you try to take some example, practice them, be comfortable with these commands and try to understand what R is trying to do with the multivariate normal distribution. So you practice and I will see you in the next lecture with more topics till then goodbye. Thank you.