Essentials of Data Science with R Software- 1

Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Lecture 49
Bivariate Normal Distribution

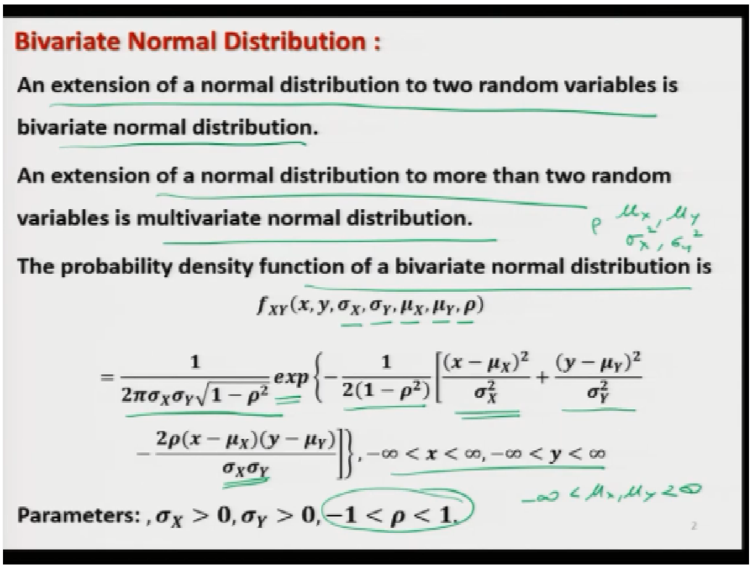
Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, up to now you have seen that in the last couple of lectures we have discussed and learned different types of concepts related to the bivariate random variables, we consider their joint probability density function, probability mass function marginal distribution, conditional distributions their expectations variance like as conditional expectations, conditional mean, conditional variance etcetera.

And now I am sure that you should not have any problem in finding out any of such quantities or you can apply or you can at least think that under what type of condition you can use the concepts of bivariate random variable. Well, there are many, many probability mass functions and probability density functions which are multivariate and now discussing all of them here it is practically difficult for me.

So, but in order to give you an idea that how this multivariate and bivariate probability functions look like and how do we handle them, I have just taken the example of the most popular bivariate normal distribution and whatever I am going to define in the bivariate normal distribution that can be extended to a multivariate case but I will try to give you some idea but after that I will very honestly request you that if you want to learn about this multivariate distribution more you will have to look into the books.

Well, handling them into the R software is not difficult and that is what I will try to show you in the class today in this lecture today. So let us try to begin our lecture and try to see how these bivariate normal distributions look like and what are the properties.

(Refer Slide Time: 2:33)



Bivariate Normal Distribution :

An extension of a normal distribution to two random variables is bivariate normal distribution.

An extension of a normal distribution to more than two random variables is multivariate normal distribution.

ρ μ_X, μ_Y σ_X^2, σ_Y^2

The probability density function of a bivariate normal distribution is

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

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

$-\infty < \mu_X, \mu_Y < \infty$

Parameters: , $\sigma_X > 0, \sigma_Y > 0, -1 < \rho < 1.$

So, let us begin, so now you can see here we are going to handle with the topic of bivariate normal distributions. So, obviously this bivariate normal distribution is simply an extension of a normal distribution that you have done in detail and this is actually called as bivariate normal distribution and in case if you try to extend it more, beyond two variables then we can also define the multivariate normal distribution which is also very well defined.

Now, in case if you try to recall the univariate normal distribution there were two parameters, one was mean, and another was variance. So, now you can think that what can be the possible parameters of a bivariate normal distributions, so obviously when you are trying to think about bivariate normal distribution there are going to be two random variables and they are going to have their own properties in terms of their parameters.

So, if I say that if I have here two random variables X and Y then they will have their mean, mean of X, mean of Y, variance of X σ_X^2 , variance of Y σ_Y^2 and do you think that there should be some other parameter? Yes, obviously X and Y are bivariate so we expect that they are interrelated, they are correlated, so there should be one parameter which is going to describe the joint behavior and that is the correlation coefficient between X and Y.

So, the probability density function of a bivariate normal distribution is characterized by such five parameters σx, σy, mu x, mu y and rho and the form of this probability density function is here like this, well this cannot be just extended from the univariate normal but surely I can inform you that if you try to rewrite the things possibly this bivariate normal distribution can be very easily extended to a multivariate normal distribution, where we are going to indicate the random variables by a random vector mean, so by mean vector and variances by a covariance matrix.

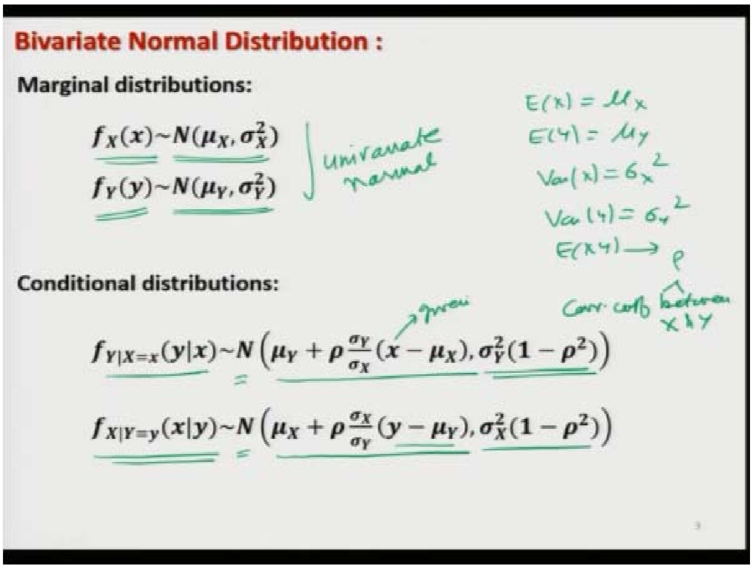
So, this form is like this

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

and $-\infty < \mu_x < \infty$, $-\infty < \mu_y < \infty$ and σ_X^2 and σ_Y^2 they will be lying between 0 and ∞ and correlation coefficient rho will be lying between -1 and +1.

Well at this moment we do not know what are this μ_x, μ_y, σ_X^2 , σ_Y^2 and although, I am informing you beforehand that once you try to find out the properties of the bivariate normal distribution they will come out to be the as a mean or variance or correlation coefficient.

(Refer Slide Time: 6:23)



Bivariate Normal Distribution :

Marginal distributions:

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

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

univariate normal

$E(x) = \mu_x$

$E(y) = \mu_y$

$Var(x) = \sigma_x^2$

$Var(y) = \sigma_y^2$

$E(xy) \longrightarrow \rho$

Corr. coeff between X & Y

Conditional distributions:

given

$$\underline{f_{Y|X=x}(y|x)}\sim N\left(\underline{\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}, \underline{\sigma_Y^2(1-\rho^2)}\right)$$

$$\underline{f_{X|Y=y}(x|y)}\sim N\left(\underline{\mu_X+\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)}, \underline{\sigma_X^2(1-\rho^2)}\right)$$

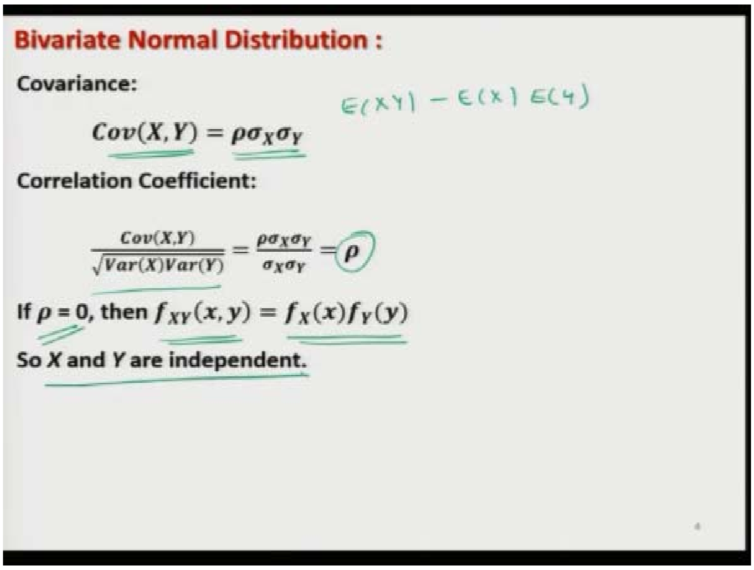
So, now I will not give you here the proof because the proof is available in almost all the statistics book, but my main objective in this data science course is to give you the property so that you can use it in applications. So, in case if you try to find out the marginal distribution of x and y then they are going to be univariate normal, with their respective mean and respective variance.

So, for example, marginal distribution of X is going to be N(μ_x , σ_X^2) and marginal of y will be univariate N(μ_y ,σ_Y^2) and if you try to find out here expected value of x this will come out to be μ_x, expected value of y will come out to be as μ_y.

And similarly, if you try to find out here the variance of x this will come out to be σ_X^2 , variance of y will come out to be here σ_Y^2 and if you try to find out the correlation coefficient and you know how to compute the correlation coefficient you will have to compute different types of expectation, including expected value of X, Y and then using those things you can see here that this ρ will come out to be the correlation coefficient between X and Y. So, this is going to be correlation coefficient between X and Y. So, that algebra I am not showing you here, but I am simply informing you the result.

Similarly, if you want to find the conditional distributions, so the conditional distribution of Y given x will also be normal but it will have a different mean and different variance. So, the conditional mean is going to be in this case $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$. So obviously, you must understand that here x is given. And the conditional variance will come out to be $\sigma_Y^2(1 - \rho^2)$ and similarly if you try to find out the conditional distribution of X given y this will come out to be a univariate normal with the mean $\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$ will come out to be $\sigma_X^2(1 - \rho^2)$. So, this is the conditional distribution.

(Refer Slide Time: 9:05)



Bivariate Normal Distribution :

Covariance:

$$Cov(X,Y) = \rho\sigma_X\sigma_Y$$

$E(XY) - E(X)\,E(Y)$

Correlation Coefficient:

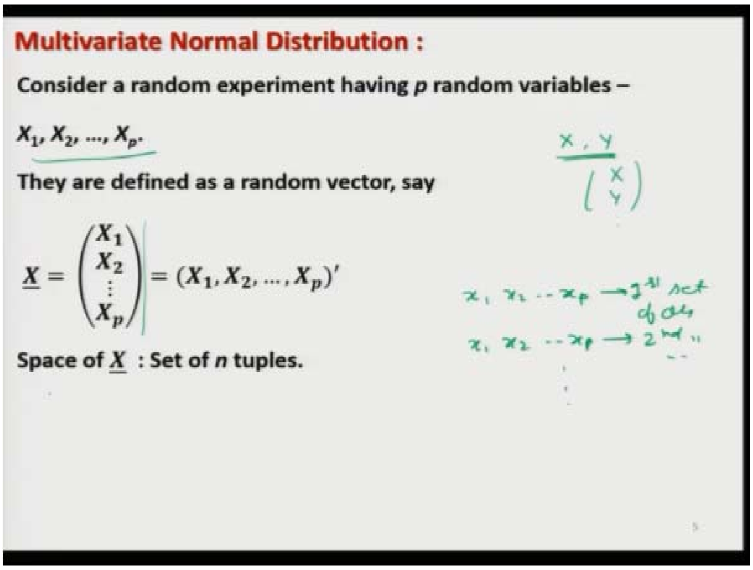
$$\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\rho\sigma_X\sigma_Y}{\sigma_X\sigma_Y} = \rho$$

If $\rho = 0$, then $f_{XY}(x,y) = f_X(x)f_Y(y)$

So X and Y are independent.

And similarly, if you try to find out the covariance between X and Y that is expected value of XY minus expected value of X into expected value of Y, if you try to simplify it this will come out to be $\rho\sigma_X\sigma_Y$ and since σ_X and σ_Y are the standard deviation of X and Y, so I can write down here that this covariance of X, Y divided by square root of variance of x into variance of y this will come out to be here rho. So, in case if rho is 0 then this joint probability density function can be expressed as the product of the marginal density functions of X and Y. So, in this case X and Y are independent, so these are very important properties of this bivariate normal distribution.

(Refer Slide Time: 9:59)



Multivariate Normal Distribution :

Consider a random experiment having p random variables –

$X_1, X_2, \ldots, X_p$.

They are defined as a random vector, say

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = (X_1, X_2, \ldots, X_p)'$$

Space of $\underline{X}$: Set of n tuples.

X, Y

$\begin{pmatrix} X \\ Y \end{pmatrix}$

$x_1, x_2 \cdots x_p \to$ 1st set of obs

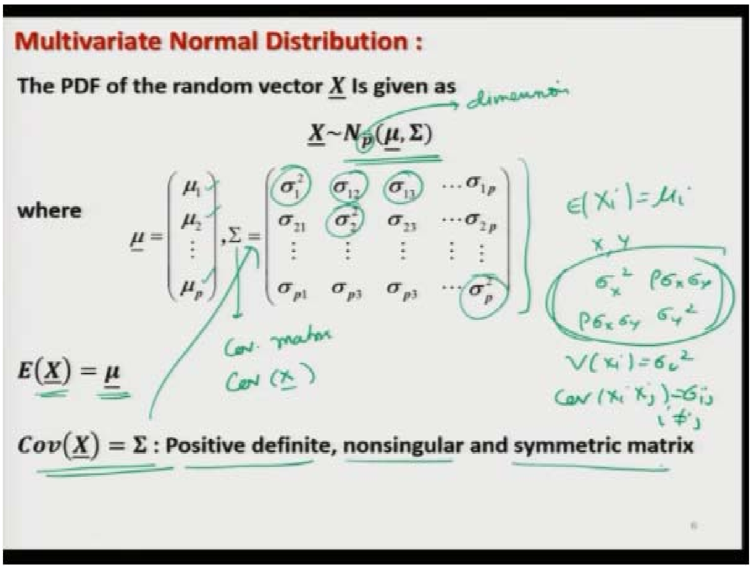
$x_1, x_2 \cdots x_p \to$ 2nd "

And this property like conditional distribution, conditional mean, conditional variance they are very useful when we are trying to use the Bayesian inference and we are trying to find out different types of results in real application when we are trying to fix one variable and then I am trying to find out the effect of the other variable.

Just to give you some idea, now I can show you here that here you have taken two random variable X and Y but suppose if I try to take a p random variable $X_1, X_2, \ldots, X_p$ so this pair of random variable X and Y can be expressed here as say as a random vector X, Y and this can be extended and I can express this $X_1, X_2, \ldots, X_p$ as a random vector here $X_1, X_2, \ldots, X_p$ which is a column vector of order p cross 1.

And when you are trying to get the observations on these variables you are going to set of $X_1, X_2, \ldots, X_p$ this is going to be the first set of observation and then you will have one more $X_1, X_2, \ldots, X_p$ which is different from the first set this will be your second set of observations, so and then you will try to repeat it and time. So, this space of X is going to be the set of n tuples.

(Refer Slide Time: 11:33)



Multivariate Normal Distribution :

The PDF of the random vector $\underline{X}$ Is given as

$$\underline{X} \sim N_{p}(\underline{\mu}, \Sigma)$$

dimension

where

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \;\vdots \\ \sigma_{p1} & \sigma_{p3} & \sigma_{p3} & \cdots \sigma_p^2 \end{pmatrix}$$

Cov. matrix

Cov ($\underline{X}$)

$E(X_i) = \mu_i$

X, Y

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

$V(X_i) = \sigma_i^2$

$Cov(X_i X_j) = \sigma_{ij}$ $i \neq j$

$E(\underline{X}) = \underline{\mu}$

$Cov(\underline{X}) = \Sigma$: Positive definite, nonsingular and symmetric matrix

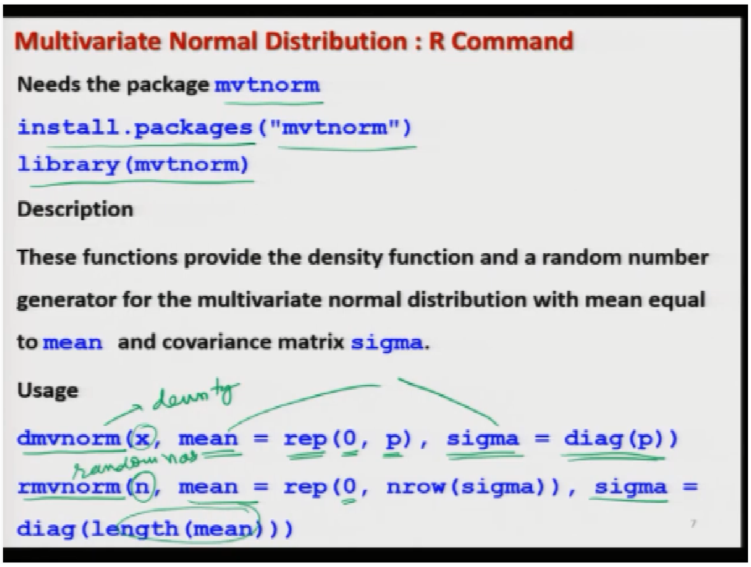
And just like you have written the probability density function of a bivariate normal density function similarly, we can also write down the probability density function of a multivariate normal distribution and that is indicated symbolically by $N_p(\mu_x, \Sigma)$, where this p is going to indicate the dimension, means how many variables are there in X_1, X_2,…, X_p and vector is going to indicate the means mu1, mu2…, mup where we assume that each of this X_i has got the mean μ_i.

So, in case if you try to find out the expected value of this random vector this will come out to be as a mean vector and similarly when you are trying to consider here X and Y you can write down their covariance matrix X here σ_X^2 , σ_Y^2 and the covariance will be $\rho\sigma_X\sigma_Y$, so they will be occurring on the off diagonal elements and variance will be occurring in the diagonal element.

So, similarly you can extend it and in case if I assume that variance of Xi is sigma i square and covariance between X_i, X_j this is equal to say σ_{ij}, i not equal to j, then I can extend this covariance matrix to a multivariate setup with p random variables and this can be written as like this which I have given here, so you can see in the diagonals we have the variance of X_1, X_2,…, X_p $\sigma_1^2, \sigma_2^2, ..., \sigma_p^2$ and on the off diagonal elements so we have here σ_{12} which is indicating the covariance between X_1, X_2 we have a σ_{13} that is indicating the covariance between X_1 and X_3 and that is going to be a symmetric matrix, why? Because covariance between X, Y is the same as covariance between Y and X.

So, this is now here instead of having a variance we have here a covariance matrix and in case if you try to find out the covariance matrix for this random vector X this will come out to be as sigma which is given here and this is a positive definite matrix, non singular matrix and symmetric matrix.

(Refer Slide Time: 13:57)



Multivariate Normal Distribution : R Command

Needs the package mvtnorm

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

Description

These functions provide the density function and a random number generator for the multivariate normal distribution with mean equal to mean and covariance matrix sigma.

Usage

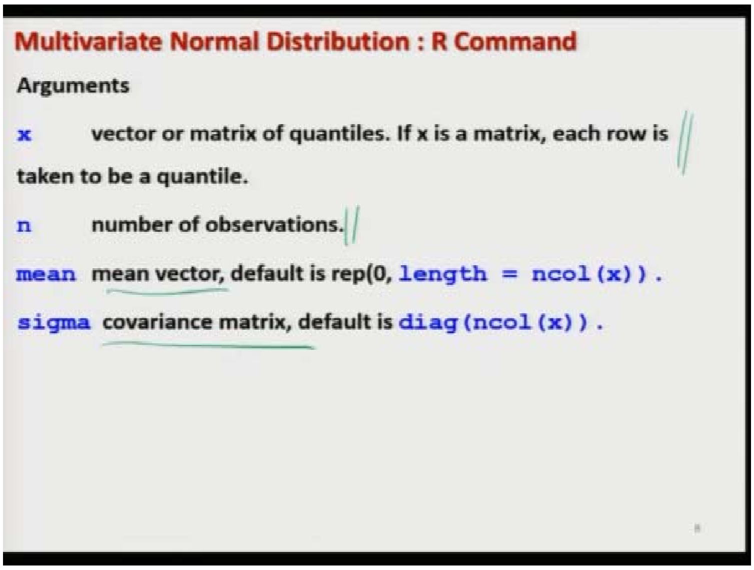
```
dmvnorm(x, mean = rep(0, p), sigma = diag(p))
rmvnorm(n, mean = rep(0, nrow(sigma)), sigma =
diag(length(mean)))
```

So, these are some properties but now I am more interested in how are you going to handle this bivariate or in general the multivariate normal distribution in R software. So, in order to handle it we need a special package whose name is mvtnorm, so you need to first download the package, you have to install the package using the command install dot packages and within parenthesis you have to write within double quote mvtnorm and then you have to load it on your computer by using the command library.

So, this function provides you the density function, random numbers etc., just like you have used earlier in all the cases. So, in case if you try to write down the command here dmvnorm this will give you the density and if you use here the command here rmvnorm this will generate the random numbers. So here the command is like this that you have to give here the data vector here x and then I am trying to take here the mean vector as 0, 0, 0 so I am trying to use here the command rep so that is repeating the value 0 p number of times and the covariance matrix that is given by the parameter sigma and this is going to be any matrix but I have taken here as say diagonal matrix of order p.

So, well I have taken it just for the sake of illustration otherwise for this mean and sigma you can take any choice and similarly here also if you want to generate the random numbers you have to give here the value of n that how many values you are going to generate and then you have to specify the mean, you have to specify the sigma according to your need and requirement, I have given you here just for the sake of simplicity mean as to be 0 for all the variables and the sigma is going to be here in the diagonal matrix of the order of the length of mean vector.

(Refer Slide Time: 16:15)



Multivariate Normal Distribution : R Command

Arguments

`x` vector or matrix of quantiles. If x is a matrix, each row is taken to be a quantile.

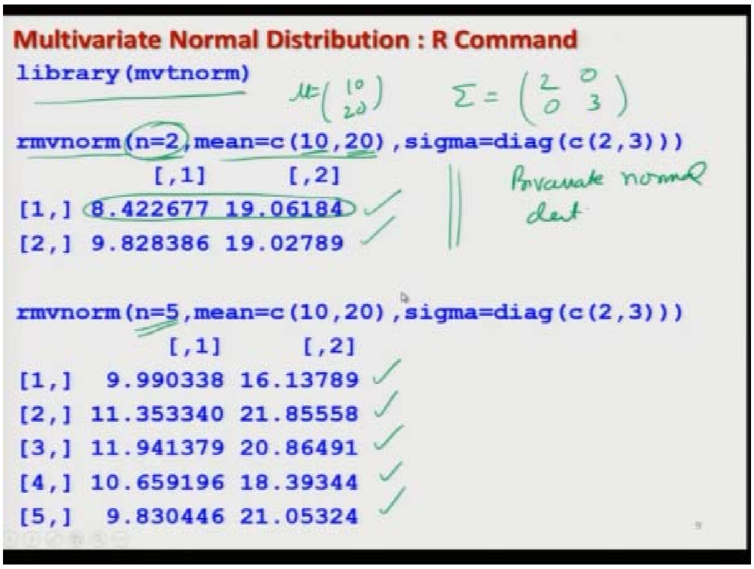
`n` number of observations.

`mean` mean vector, default is rep(0, `length = ncol(x)).`

`sigma` covariance matrix, default is `diag(ncol(x)).`

So, let me try to explain you these things, so x here is the vector of or the matrix of quantiles, n is the number of observation, mean is the mean vector and sigma is the covariance matrix.

(Refer Slide Time: 16:27)



Multivariate Normal Distribution : R Command

$\mu = \begin{pmatrix} 10 \\ 20 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$

Bivariate normal dist

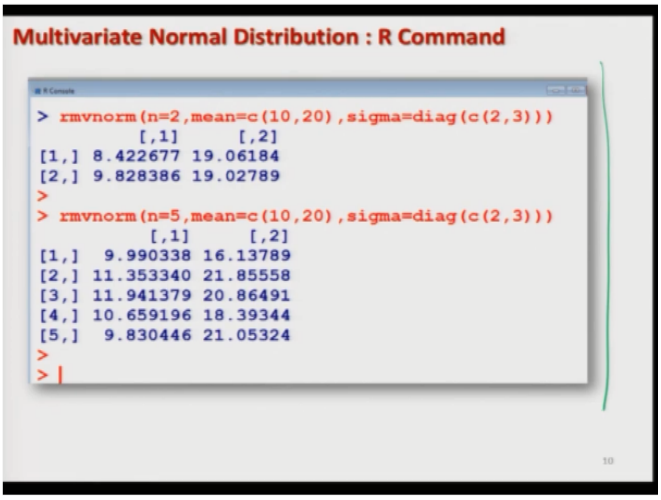
```
library(mvtnorm)

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

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 ✓
```

Anyway, so now let me try to show you that how are you going to get the random numbers in this case. So first you have to load the library mvtnorm and after that you have to just write the command here rmvnorm, n is equal to 2 for example, I want to get here two sets of observation and now I am giving here the mean as say like as 10 and 20, so mean here is like here mean 10 and 20 and sigma I am trying to give it here as a diagonal matrix 2 and 3 and off the elements are 0 but you can choose anything actually that is up to you, the only condition is that Σ has to be a positive definite matrix and symmetric.

So, if you try to do here you will get here this type of outcome, so you can see here this is a bivariate normal distribution because you have taken here only here two values in the mean vector, so that is indicating that this is a bi vector and you have given only the value of μ_x and μ_y. So, you can see here the first set of observation here is like this, these are the random numbers generated and similarly because you have given here n is equal to 2, so there are one and here two sets of observations.

And similarly, if you try to give here n equal to 5 then you will get here one, two, three, four, five sets of observation and every set of observation will be containing two random numbers but if you want to make it here more you can give it very easily without any problem.

(Refer Slide Time: 18:09)



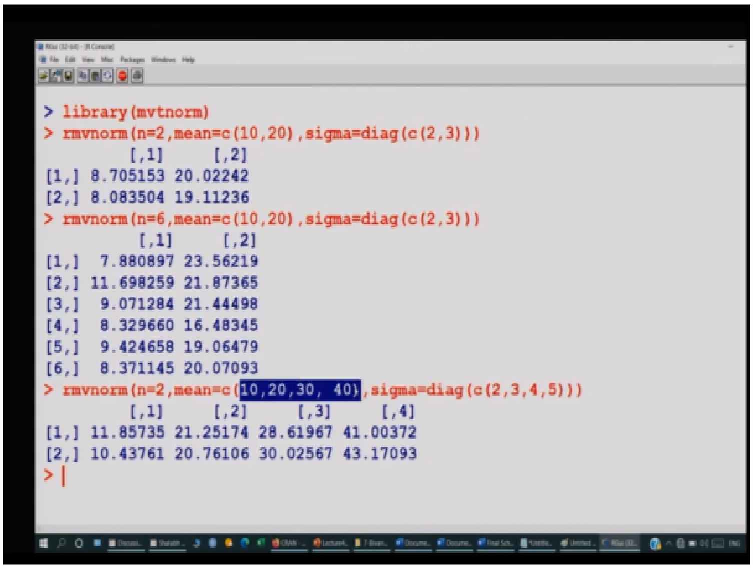
Multivariate Normal Distribution : R Command

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

10

So now, let us come to the R console and I try to show you that how these things are going to happen, but here is the screenshot of the same observations which I have shown you but definitely when I am going to do it on the R console these things are not going to be repeated.

(Refer Slide Time: 18:35)



```
> library(mvtnorm)
> rmvnorm(n=2,mean=c(10,20),sigma=diag(c(2,3)))
         [,1]     [,2]
[1,] 8.705153 20.02242
[2,] 8.083504 19.11236
> rmvnorm(n=6,mean=c(10,20),sigma=diag(c(2,3)))
          [,1]     [,2]
[1,]  7.880897 23.56219
[2,] 11.698259 21.87365
[3,]  9.071284 21.44498
[4,]  8.329660 16.48345
[5,]  9.424658 19.06479
[6,]  8.371145 20.07093
> rmvnorm(n=2,mean=c(10,20,30, 40),sigma=diag(c(2,3,4,5)))
         [,1]     [,2]     [,3]     [,4]
[1,] 11.85735 21.25174 28.61967 41.00372
[2,] 10.43761 20.76106 30.02567 43.17093
> |
```

So, first let me try to copy this command here. So, now first I need to load this library mvtnorm, well this package is already there on my computer otherwise you will have to install it on your computer first. So let me load this library and if I try to give here this command here rmvnorm you can see here that with the mean 10, 20 it is trying to give me here two values and if you try to generate here more values here six values so you can see here you are getting here six sets of observation and every set of observation has two values, one corresponding to x and one corresponding to y.

Similarly, if I try to say increase here the order of this mean vector, that means the number of variables suppose if I try to make the mean vector to be 10, 20, 30, 40 and I try to make this covariance matrix as the diagonal matrix consisting of elements 2, 3, 4, 5 you can see here now you are getting two observation but every observation has got four values.

So this is we try to execute through the R software for the multivariate normal distribution, although if you try to take it here only two values here 10 and 20 which I have done here, then it is bivariate normal and if you try to make it here more than two up to whatever order you want this will become a multivariate normal and this is how we try to compute the density random variables etcetera, and you can do the same exercise what I have shown you in so many earlier probability density functions.

So, they are now very simple and straight forward things, so my idea in this lecture was only to give you an idea that how this bivariate or multivariate probability functions look like and what is the difference in the way we try to handle them. For example, earlier we were finding only the mean variance and other movements but when we are trying to go for bivariate or multivariate we are also interested in the marginal distributions, conditional distributions and whatever we have done here.

For example, we have taken only here two variable but now if you try to take here a vector suppose, this vector has suppose 20 value so you can take a sub vector also and sub vector will itself be a random vector and then you can find out the marginal, conditional, etc., instead of conditioning only on a single univariate random variable you can condition it on a sub vector also.

Sub vector is also a vector so I can find out the for example, conditional distribution of x given y where x and y both are some vector quantities so those things can be done and I can promise you these things are not difficult. At the moment you are starting so I do not want to make the life complicated, but my idea is to give you the confidence, yes, if you can understand univariate normal bivariate normal then you can very easily understand the multivariate normal also.

And similarly, we also have generalization of the binomial distribution also as multinomial distribution that also goes exactly on the same lines but surely, I will not go into that much detail but I will try to take up a new topic in the next lecture. So, now we are concluding with our random variables and their probability functions and their properties, now it is the time that you try to revise once again all the lectures and try to see whether you have understood the concepts or not and more important part when you are trying to look at them from the data science point of view, you have to understand what is the meaning of their values, what are they trying to understand and how you are going to compute them in the R software.

So, that was my objective as I said the title of the course is also essentials of data sciences because unless and until you understand these things you cannot surely work in the data science, so I would request you that to try to revise it and I will see you in the next lecture till then goodbye.