# Data Analysis and Decision Making - I

Prof. Raghu Nandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

## Lecture – 32
## Multivariate statistical analysis

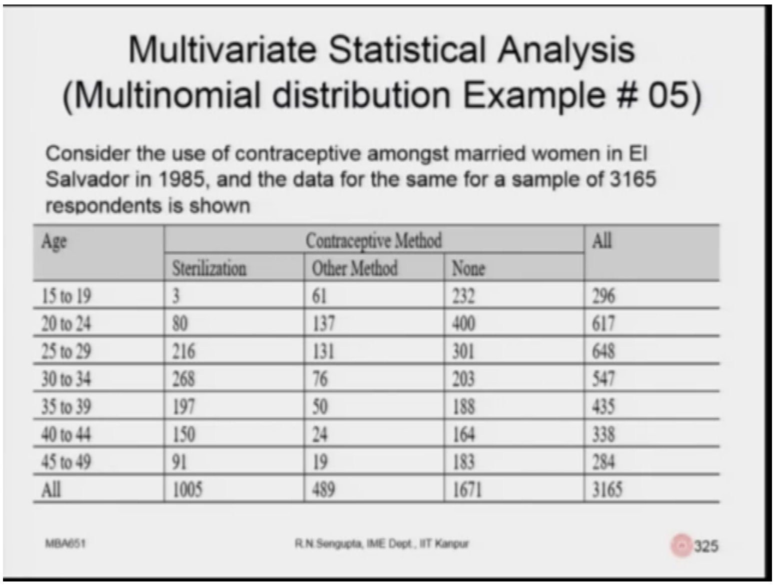
Welcome back my dear friends, a very good morning, good afternoon, good evening to all of you. This is the Data Analysis and Decision Making 1 course under NPTEL MOOC; and as you know this is a 12 week course and that is 60 lectures which is of 30 hours total combined. And each week we have 5 lectures each being of half an hour. So, we have completed the middle almost till 30 lectures we are going to start the 32nd 1 today.

So, if you remember we are discussing about after almost and the fagg in or 29th or 29. We started basically the multivariate statistical methods and we discussed that what was how to what was basically definition wise. What was the mean the standard deviation of the population or the variance of the population and that can be the second part; can basically be portraited or given the information as a covariance variance matrix which is a p cross p or k cross k metric. Were p and k are the number of independent variables and the mean are being basically given by a vector mu p size being p cross 1.

The first element has obviously; you know is the number of rows and second being the number of columns. Then the corresponding sample mean would be x bar p. So, these mu this covariance variance matrix which is a summation symbol or the x bar p all are bold either matrix or vectors. Then the corresponding part for the variance covariance matrix should be the capital S, which is the standard error square values. Then you have the correlation coefficient the counterpart in the case which is also p cross p or k cross k matrix.

And the counterpart in the sample would be capital R all are bold S is bold R is rho is bold then R is bold. Then we considered the multinomial model and that is basically the as is mentioned the polynomial expansion multinomial expansion the parameter the coefficients would be exactly be as discussed in the multinomial model. And we also discussed using the Pascal's triangle how the slices which you take basically depicts the parameters the coefficients. So, we will continue further on with the multinomial distribution.

(Refer Slide Time: 03:06)



## Multivariate Statistical Analysis (Multinomial distribution Example # 05)

Consider the use of contraceptive amongst married women in El Salvador in 1985, and the data for the same for a sample of 3165 respondents is shown

| Age | Contraceptive Method | | | All |
|---|---|---|---|---|
| | Sterilization | Other Method | None | |
| 15 to 19 | 3 | 61 | 232 | 296 |
| 20 to 24 | 80 | 137 | 400 | 617 |
| 25 to 29 | 216 | 131 | 301 | 648 |
| 30 to 34 | 268 | 76 | 203 | 547 |
| 35 to 39 | 197 | 50 | 188 | 435 |
| 40 to 44 | 150 | 24 | 164 | 338 |
| 45 to 49 | 91 | 19 | 183 | 284 |
| All | 1005 | 489 | 1671 | 3165 |

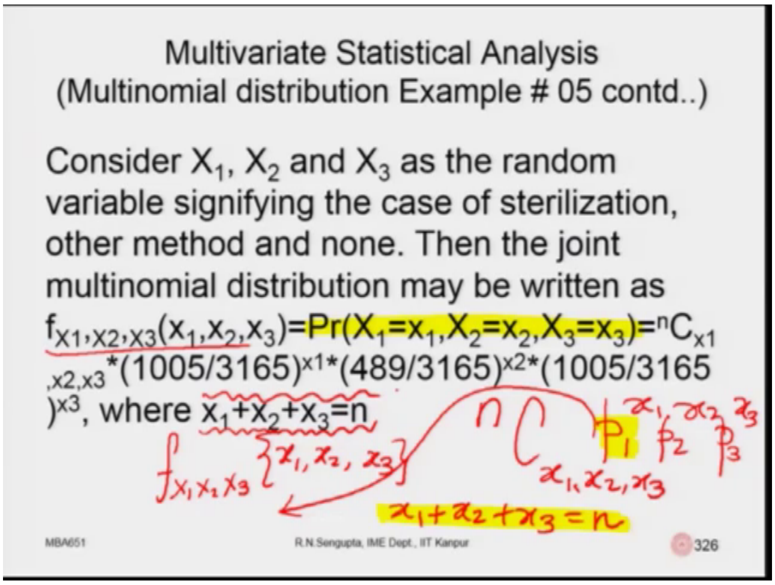
MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 325

So, consider these are actual data so based on the data's we are there to basically give you a picture. So, there till now for the multinomial they would has not been any problem, problem means some solved examples we just given examples and trying to give it in a pictorial format. To consider the use of contraceptive amongst married women in else and the verdure in the 19 year 1985 and the data for the same is basically there for a sample of 3,165. These are all actual data is taken from the net and the respondents gave their responses accordingly.

So, in the first column you have the age, so for the first group is 15 to 19 then 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49. The second one basic second third and fourth one are the respective set of women in that age category, who used contraceptive methods sterilization or they use other matter or they use none. So obviously, to give an example if you consider the age back at 35 to 39 the corresponding values of number of women who use sterilization method other methods and none is basically 197, 50 and 188 and correspondingly the values are given for each age bracket. And the last column is basically the summation of all the number of women for each age category. So, say for example, for the h category 20 to 24, the total number is 670.

So, you have these numbers and also have a look at the last row, which gives you in each category the total number of women irrespective of the age bracket, for sterilization is 1005 for other method is 489 and for none using any contraceptive method is 1671. So, you will basically try to use this data for the multinomial model.

(Refer Slide Time: 05:14)



Multivariate Statistical Analysis
(Multinomial distribution Example # 05 contd..)

Consider $X_1$, $X_2$ and $X_3$ as the random variable signifying the case of sterilization, other method and none. Then the joint multinomial distribution may be written as $f_{X1,X2,X3}(x_1,x_2,x_3)=Pr(X_1=x_1,X_2=x_2,X_3=x_3)={}^nC_{x1,x2,x3}*(1005/3165)^{x1}*(489/3165)^{x2}*(1005/3165)^{x3}$, where $x_1+x_2+x_3=n$

$f_{X_1X_2X_3}\{x_1, x_2, x_3\}$

${}^nC_{x_1,x_2,x_3}\, p_1^{x_1} p_2^{x_2} p_3^{x_3}$

$x_1+x_2+x_3=n$

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 326

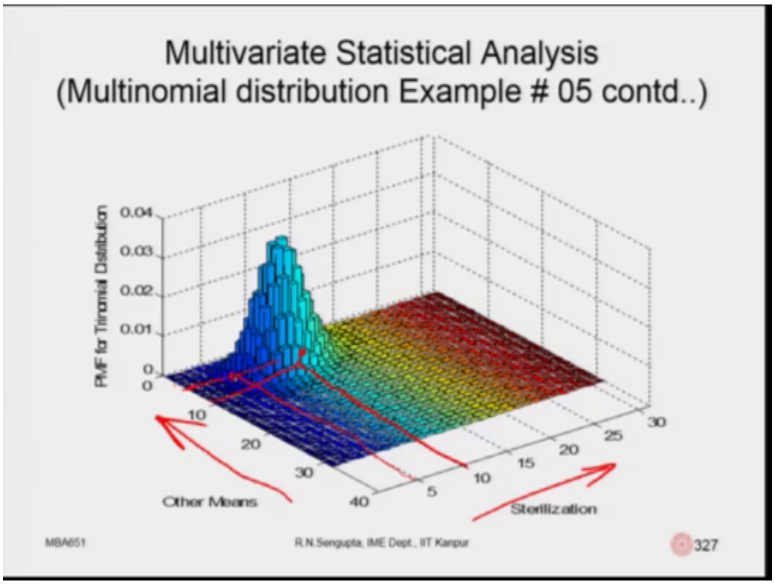
Now, consider X1, X2, X3 as the random variable signifying the three variables or the information which is the case of sterilization, which is a user other method and using none of the methods. So obviously, in this X1, X2, X3 would is the random variable which will depict that depending on the age.

Now, but I am basically considering all of them together age is not a factor now. So, then the joint multinomial distribution this is basically of three variables trinomial distribution may be written. So, this capital so technically these values So this is basically f capital X 1 X 2 X 3 x1 x2 x3. So, this is basically the PMF value actually it means what it actually means is this? That means, when capital X 1 which is the random variable takes the realest value a small x1, capital X 2 takes the realest value of small x 2 and capital X 3 takes the realest value of small x 3. Then the multinomial distribution is given like this so, again I use the red colour.

So, is n C x1 x2 x3 so obviously, that would be factorial n divided by factorial x1 factorial x 2 factorial x3 and so on and so forth. So obviously, remember this is important and the corresponding formula which you had was these were the coefficients. So, the probabilities would be p 1 x 1 p 2 x 2 p 3, x 3. So, now, if I ask you the question what is p 1 or p 2 or p 3 and you will rightly mention that. So, p 1 would basically be so I will take this. So p 1 would be the number of women who are using sterilization because x1 in corresponds to that number of people women who are using sterilization with respect to the total amount of woman who have been the respondents. So, that would; obviously, be 3161 35 is the total number. So, it will be 1005 divided by 3165.

Similarly, for p 2 it will be the number the ratios or the relative frequency or the chance of the number of women who are using other methods divided by the total number which is again 489 divided by 3165. And the last value would be corresponding to number of number for the women who are using none of the methods no sterilization methods. So, that number is relative frequency is 1005 divided by 3165. So, based on that you can find out this values the moment you put x1 x2 x 3; obviously, if you should remember this. So, I will just underline it. So, the sum of X1, X2, X3 is equal to n. So, you put it and solve the problems accordingly .

(Refer Slide Time: 08:47)

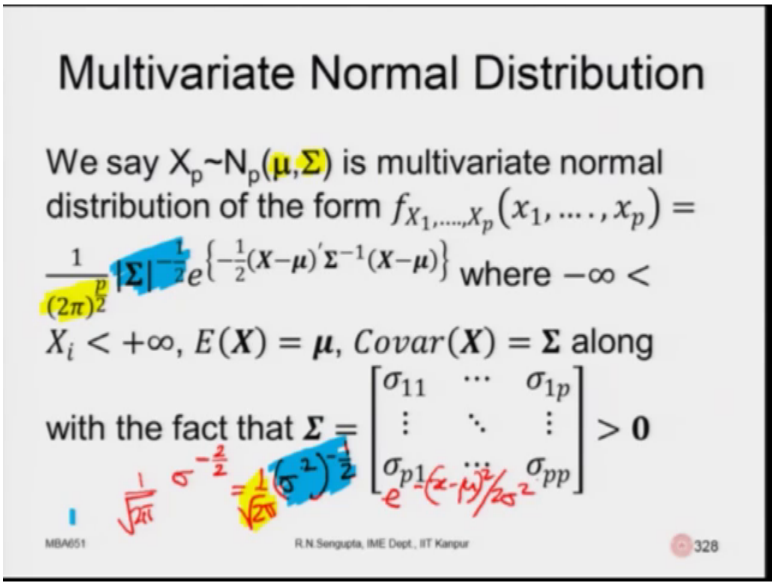


Multivariate Statistical Analysis
(Multinomial distribution Example # 05 contd..)

Now, here I have basically consider the trinomial distribution. So, in this case I know is remember because x1 and x2 if they are given x2 is immediately known because the sum is always n. So, I plot the sterilisation along the x axis. So, I am just marking it other means along say for example, the Y axis and basically if I consider the probability mass function for the trinomial distribution considering some values of x1 x2 x3 so, I keep changing probabilities are fixed remember p 1 p 2 p 3 are fixed. So, I keep changing X1, X2, X3 based on the fact that the sum of X1, X2, X3 is equal to n change that those value plot them. So, consider here is x1 and consider here is x2 so the value would be somewhere here.

So, and obviously, in that case x3 is already known because x3 would be n minus x1 minus x2. So, in really if I consider this 5 value for x2 take it along the x axis I take a long say for example, for x1 along this axis where they meant this height would be the corresponding PMF because immediately x1 x2 being known you have x3 because the sum is n sum means x1 plus x2 plus x3. So, there are very simple codes you can utilize in maths lab or r or whatever and you will get the information's according. Now, we considered the multinomial multivariate normal distribution. So, of p number of variables or k number variables. So, we say x suffix p, p mean the number of independent variables.

(Refer Slide Time: 10:46)



# Multivariate Normal Distribution

We say $X_p$~$N_p(\mu,\Sigma)$ is multivariate normal distribution of the form $f_{X_1,....,X_p}(x_1,....,x_p) = \frac{1}{(2\pi)^{\frac{p}{2}}}|\Sigma|^{-\frac{1}{2}}e^{\left\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right\}}$ where $-\infty < X_i < +\infty$, $E(X) = \mu$, $Covar(X) = \Sigma$ along with the fact that $\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix} > 0$

$\frac{1}{\sqrt{2\pi}}$ $\sigma^{-\frac{2}{2}}$ $= \frac{1}{\sqrt{2\pi}}(\sigma^2)^{-\frac{1}{2}}$ $e^{-(x-\mu)^2/2\sigma^2}$

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 328

P is not normally distributed with the mean value of mu. So, this is bold remember and the variance covariance matrix being given. So, we will say this is a multi variate normal distribution and the form would be this is the PDF value. So, f x1 to x n and these actually means the corresponding realize values for the random variable x1 to x p. So, that would be given by the falling distribution and remember. So, this is a very simple case if you remember the formula looks ominous threatening, but it is not because if you do a 1 to 1 comparison with the univariate case the information comes out very easily.

So, in the univariate case what you had? You had 1 by 2 pi so obviously, here this 2 is corresponding to square root which is here. Now why p because there as such p number of random variables, so obviously, p would come because you are multiplying the corresponding probabilities of PDF's. Now you had 1 by sigma which was the standard deviation. So, in this case sigma would be replaced by the corresponding variance covariance matrix.

So, technically I could have written this as it is it. So, look at this I should write it this should not be it is not the way how I wanted to be. So, it is 1 by sigma, sigma goes up it becomes minus and Y so. So, now, if I do it square so; obviously, root 2. So, in this case I write it as this part only as 1 by 2 pi sigma square to the power minus half and this we will see again a 1 to 1 correspondence. Sigma square is what? Sigma square is the variance covariance so, called value with this in the univariate case is 1 value only.

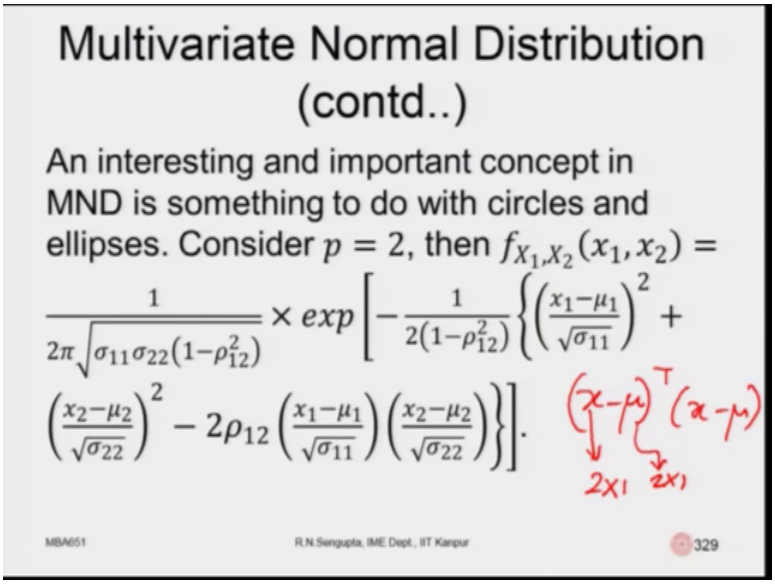
Now, if there are 2 random variables. So, the principal diagonal will be the first one 1 comma 1 would be the covariance of the first which with itself which is the variance of the first, the 2 comma 2 would be the covariance of the second with itself which is sigma square, square to the bar to a squeeze sigma base 2 to th e power 2 because that is the variance of the second and the of the diagonal element would be sigma 1 2; that means, 1 2 number is in the suffix and sigma 2 1 and they being symmetric. So, in this case you have the variance covariance matrix. So, that is here mod of that because you have to consider the this positive definite and to the power half because you are taking it to the numerator hence half here divided by 2 because you are taking the square.

So, I will use the highlighter first for the brown. So, this part has some oh my god. So, sorry colour of obviously, was could not be red. So, I use a lighter colour of would it will be possible yes. So, this is what I have for this case then I try to basically have a set of information here which matches here. Now in the case of the univariate so later on you had e to the power minus x minus mu whole square by 2 sigma square. So, this is the total formula. So, e to the power is here, so I will try to utilize the colour again let me check 1 colour let me read use the light green yes.

So, here again you will have e to the power. So, this becomes this is sigma square sigma square is what. So, sigma square this 2 is coming here with a minus sign the sigma square is basically, I should use try to use colour. So, it is easy for you have to understand. So, let me use dark blue as possible if it is not too problematic I am just dotting it. So, it does not get problematic. So, this part now finally, let me check any other colour orange is there yes. So, this is x minus the square value. So, this is the random value minus mu value this is this. Z because why this is a transpose x minus mu transpose into x minus mu because they are vectors so; obviously, you have to take the value values accordingly. So, this x is basically the random variables corresponding to the first second third fourth so on and so forth.

Now, as usual you will have x values ranging from minus 1 to plus infinity an expected value of x is nu covariance of x is the variance covariance matrix and the fact that the covariance variance matrix would always be positive definite. So, I will basically EJS this on highlight the important facts. So, this is the covariance of the first with itself which is the variance similarly covariance of second with itself which is 2 comma 2 element till the last one which is the covariance of pth 1 with itself which is the variance of the pth random variable and of the diagonal element those sigma 1 p suffix 1 p which is the covariance of first with p is equal to the coefficient covariance of pth to the one. So, this is a symmetric matrix.

(Refer Slide Time: 18:29)



## Multivariate Normal Distribution (contd..)

An interesting and important concept in MND is something to do with circles and ellipses. Consider $p = 2$, then $f_{X_1,X_2}(x_1,x_2) =$

$$\frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \times exp\left[-\frac{1}{2(1-\rho_{12}^2)}\left\{\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right\}\right].$$

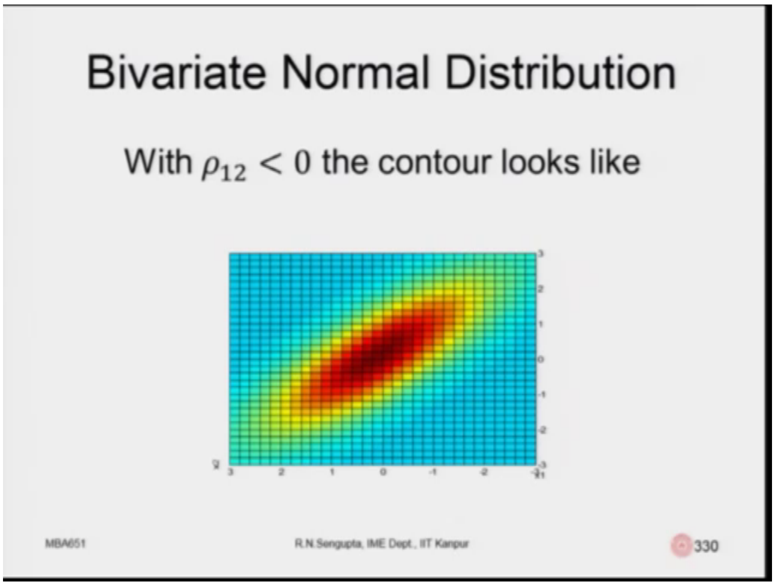
$$\underset{\substack{\downarrow \\ 2x1}}{(x-\mu)}^{T}\,\underset{\substack{\downarrow \\ 2x1}}{\phantom{)}}(x-\mu)$$

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 329

And interesting an important concept in multinomial normal distribution is sometimes to do with circles and ellipses. So, to which degrees they are dependent consider p is now 2 so, in this key p in case is p is to so obviously, it will be given by in this formula this is a set formula you can just expand whatever I given. So, again 2 pi comes the square root comes exponential comes outside minus 1 by 2 is there this square whole square. So, if you expand it x mi x1 minus mu 1 that matrix into x1 minus x2 into minus mu 2 that matrix if you multiply.

So, what I am actually a mean is this. So, in this case you have x minus mu transpose x minus mu. So, here x is of size 2 cross 1 similarly you will have this 2 cross 1 and you continue doing it and find out the value this is the formula given.

(Refer Slide Time: 19:37)



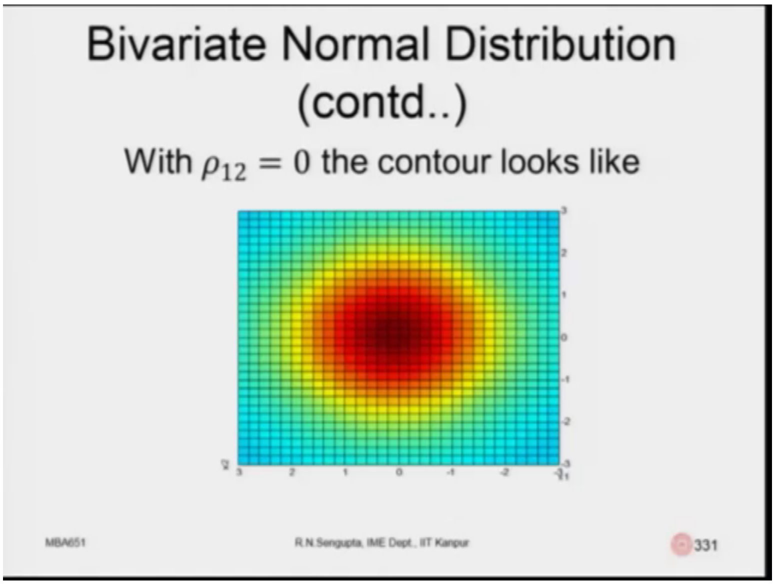
Bivariate Normal Distribution

With $\rho_{12} < 0$ the contour looks like

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 330

Now, the case is that we want to basically have the contours coming out in the picture. So, if encounters mean the relationship. So, if I am considering the relationship of the first variable and the second variable and I take slices of. So, multinomial bivariate normal distribution would look like a hillock symmetric hillock and if both of them are the correlation coefficients is 0. So, it will be if you take slices it would be circle. And if you they are positive related correlated and negative correlated they would be then in the first second first and the fourth or the second and the first and the third and the second and the fourth quadrant. So, this is what is being highlighted.

(Refer Slide Time: 20:30)



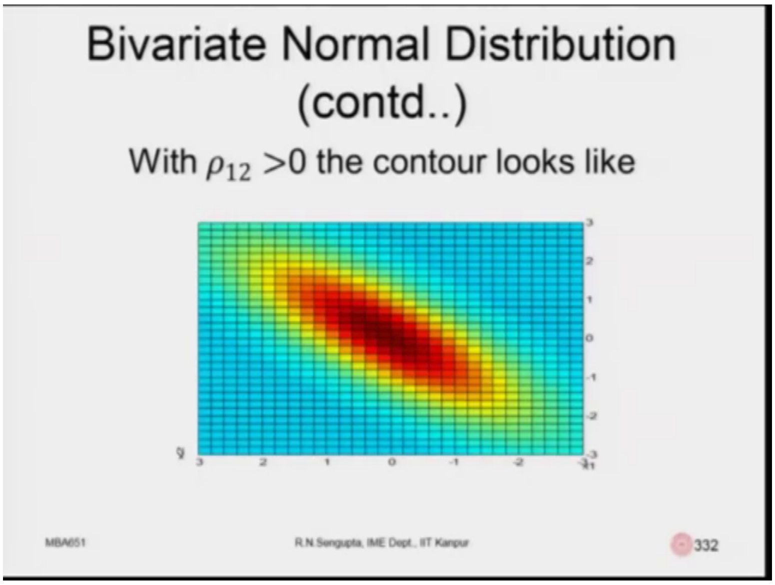
## Bivariate Normal Distribution (contd..)

With $\rho_{12} = 0$ the contour looks like

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 331

So, if I have the bivariate normal distribution row 1 to row 2 1 is 0. So, they are symmetric and if the hillock means basically you are looking from the top view. So, the concentric circles would be there more concentration on the values on the data point near the central line if you basically put a vertical line down from where I am looking. So, the points are concentrated and, but in general the relationship between random variable 1 and random variable 2 would be such that the correlation coefficient does not exist. If it does not exist; obviously, it means the covariance variance matrix would have only the principal diagonals of the diagonal elements are not there because the values are technically 0, but obviously, variants of first variants of second would remain.

(Refer Slide Time: 21:25)



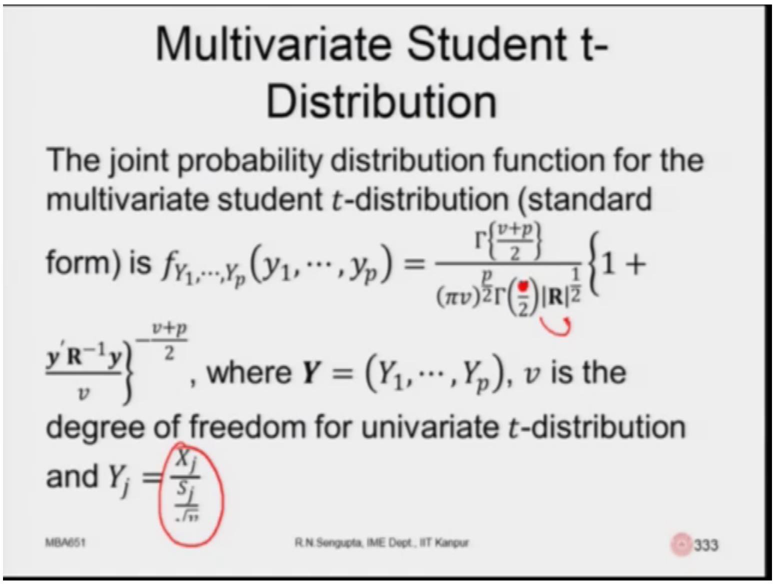
## Bivariate Normal Distribution (contd..)

With $\rho_{12}$ >0 the contour looks like

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 332

Similarly, in the counter would look for the case when the positive negative correlations are considered accordingly. So; obviously, you can imagine in a 3 dimension, 4 dimension, 5 dimension they would be spheres hyper spheres ellipse hyper ellipse and so on. And so these are ellipses with the major and minor axis accordingly and they would be ellipses. So, you can basically visualize how they would basically look like is this slice is say for example, you are considering the game of rugby. So, this is an elongated wall if you take slices principal diagonal principal axis at the major axis and minor axis major axis would be the 1 for which the variances had, minor axis being the case where the variance is low and it can be reversed major becomes minor, minor becomes major.

(Refer Slide Time: 22:21)



## Multivariate Student t-Distribution

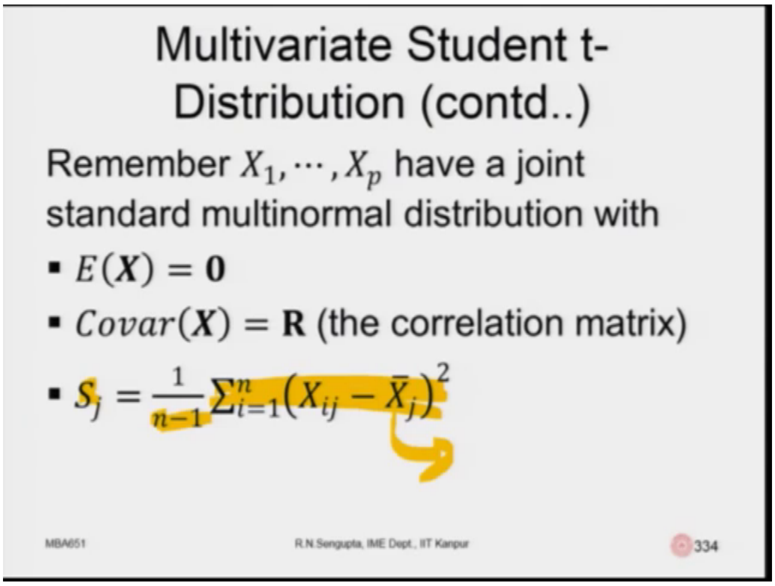
The joint probability distribution function for the multivariate student $t$-distribution (standard form) is $f_{Y_1,\cdots,Y_p}(y_1,\cdots,y_p) = \frac{\Gamma\left\{\frac{v+p}{2}\right\}}{(\pi v)^{\frac{p}{2}}\Gamma\left(\frac{v}{2}\right)|\mathbf{R}|^{\frac{1}{2}}}\left\{1 + \frac{\mathbf{y}'\mathbf{R}^{-1}\mathbf{y}}{v}\right\}^{-\frac{v+p}{2}}$, where $\boldsymbol{Y} = (Y_1,\cdots,Y_p)$, $v$ is the degree of freedom for univariate $t$-distribution and $Y_j = \frac{X_j}{\frac{S_j}{\sqrt{n}}}$

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 333

Now, if you remember we have considered the t distribution and that was for the univariate case. So, the t distribution which will consider now if the joint probability distribution function for the multivariate student t, and the standard form is given by the same formula, but we are just extending it to the multivariate case. So, here we will consider this value of so, called v is the degrees of freedom. So, there you had n.

So, these degrees of freedom and where y his has v degrees of freedom is a univariate t distribution and each of them are univariate Y 1 Y 2 Y 3 are univariate t distributions and you remember that how we have obtained the t distribution. So, you can basically normalize them. So, you are normalizing x i on x j which is the standard normal divided by the s which is the standard error that S dash and S concept divided by the square root of the degrees of freedom.

(Refer Slide Time: 23:31)



Multivariate Student t-Distribution (contd..)

Remember $X_1, \cdots, X_p$ have a joint standard multinormal distribution with

- $E(\boldsymbol{X}) = \mathbf{0}$
- $Covar(\boldsymbol{X}) = \mathbf{R}$ (the correlation matrix)
- $S_j = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$

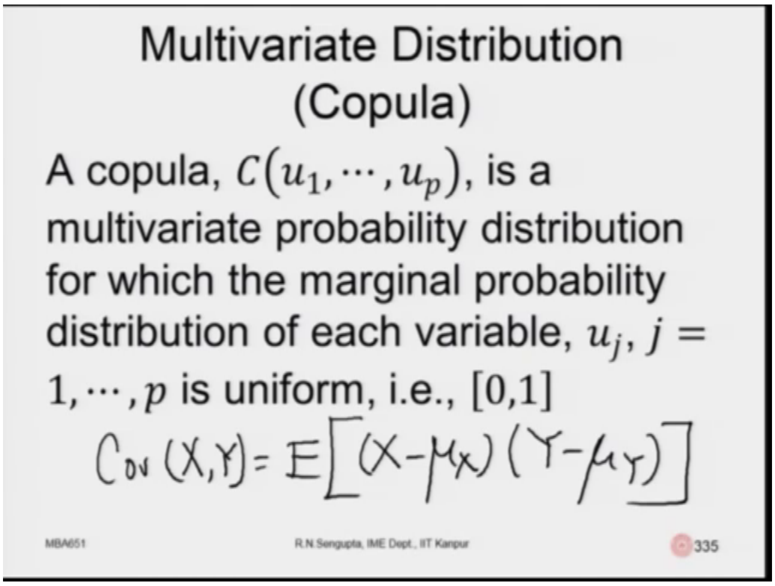
MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 334

And this is there if we remember this s dash and s. So, let me highlight S.So, this this n minus 1 is coming these values are there. So, here the population mean is not known. So, here remember x1 to x p have a giant standard multi normal distribution. So, x1 is normal x2 is normal x3 is normal till x p is normal. So, once you combine them you will find out that there are multinomial normal distribution will for a multi normal distribution. We will consider here in such a case that the expected values of x is 0.

In the standard case because it is a standard multinomial distribution; obviously, the mean value would be 0 and the variance would obviously, remember for each would be 1 and corner corresponding the fact you will have the variance covariance matrix depending on the correlation for each and every element ith j i to j j to i all these things, The covariance of x would be given by for the matrix R which is the correlation coefficient matrix. And in this case you will find out s naught s dash or s star because in that case the mean values are known in this case the mean values are not known.

So, you will use S without the dash and that would basically be. So, this should be square my apology should be squared that will be 1 by n minus 1 depending on the degrees of freedom, summation of all the squared values of the difference between the realest values on the mean menu. Now we will discuss something to do copula function, and here I will again come back to the copula function later on. So, let me give you a very brief background even though statistically it may not be adequate considering the whole set of information I would like to pass it on to you. So, consider the covariance.

(Refer Slide Time: 25:51)



Multivariate Distribution (Copula)

A copula, $C(u_1, \cdots, u_p)$, is a multivariate probability distribution for which the marginal probability distribution of each variable, $u_j$, $j = 1, \cdots, p$ is uniform, i.e., $[0,1]$

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

MBA651

R.N Sengupta, IME Dept., IIT Kanpur

335

So in the covariance function if you consider, the covariance function actually looks like this basically the expected value of the multiplication on the differences between X and its mean values and Y in its mean values. Now, consider if you consider this is sort of linear relationship, but consider that at the extremes either for a normal distribution or any other distribution, the relationship is such that per unit change in X or per unit change in Y, the rate of change of the value of the other random variable is happening at a faster rate, it is not linear anymore.

But if you consider the covariance concept, we consider that whatever the values of xs are if it changes by 1 unit then the Y value would change correspondingly or proportionally and the statement also remains true for the other case. When Y is at any extreme very high values or low very low values, you need not be very high positive it can be very low negative also.

So, as Y changes or exchanges the rate of change of the corresponding other random variables are not linear. So, in order to basically map the relationship between X and Y at the extremes we use the copula function. Copula function we basically sort of linkages or change linkages which basically gives you the relationship between X and Y. So, copula function is given by C and this you want to u p are the univariate case why they are univariate, I am going to come to that later.

So, C is a function of univariate numbers of distribution you want to u p, and it is basically a multivariate probability distribution copula by itself is a probability function for which the marginal probabilities are given in such a way that u j, j is equal to 1 2 3 4 till p. Because your p variables would be uniform; that means, each of the us are between 0 and 1, combining them we find out a copula function which is a function of u 1 to UPS are that we are able to map a linkage have a linkage between the random variables X and Y such that the extreme of the higher values. The relationship between x and Y are given which are much true in nature than with respect to the covariance variance matrix relationship which is linear in nature.

With this I will end the 32nd class and continue further discussion on the multivariate distributions. And for any queries please feel free to write to the on the forum and we will definitely answer it as soon as possible with all whatever information is needed. And I once again I am telling you because this will be repeating time and again especially for the later part after the 30th lecture which is to do with multivariate statistical methods that read the books. Books are absolutely plenty in the market plenty in the net. So, the set of books which we have suggested are more from a conceptual point clearing and obviously, people can after doing that it would be much easier for them to pick up any data set oriented book in multivariate statistics. And read them and base to the problems accordingly. With this I will end the class have a nice day and.

Thank you, very much.