

Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 31
Multivariate statistical analysis

Welcome my dear friends, a very good morning, good afternoon, good evening to all of you. Welcome to this NPTEL MOOC course titled Data Analysis and Decision Making. And as you know this is the course for 12 weeks, and this total number of hours is 30 hours, total number of lectures is 60, and each week we have 5 lectures, each being of half an hour.

And if you remember we have just finished 30 lectures that means, we have just completed half of the whole course and that means, half of the whole course took about the whole portion of very keeping it to very brief the whole portion of univariate statistical methods in a very simple way. And if you remember that in the 30th lecture the beginning, for about 15 minutes I gave our wrap up of what we have discussed, and how we have proceeded.

So, if you have understood the all the concepts in the 30 lectures related the univariate statistics, I do understand many of the things were done in a manner so as to give a brief outline. But based on the fact that he have inculcate in within yourself an interest for these topics, please read the book please read any good textbook, especially for the colleges, for the engineering colleges for statistics, which are more applied in nature. Because this is this I do remember mentioning it time and again this is not a theoretical DADM one not a statistics course. It is more to do with the application, and how the concept are utilized.

And then also I urged in the 29th lecture in the 30th lecture also please give as feedback to the to us about how the course is going on, how you are feeling comfortable with the level of coverage, whether they are becoming interesting, the coverages are fine, whether the examples are fine, such that we and whether how the assignments are such that we can also learn and make this course in a much better fashion for all of you, who really want to learn.

Now, if we remember that in the 30th lecture I was basically discussing the different type of matrix or different type of characteristics hope for the multivariate statistical distribution. We discussed the mean, we discussed the standard deviation, we discussed the sample mean, we discussed the sample standard deviation, which was the error, and what were the symbols like mu, the variance covariance matrix, then x bar being for the sample counterpart S being the counterpart in the sample for the variances. They would be counterpart for the for the skewness kurtosis also.

So, now if you remember, I did discuss about the correlation matrix, correlation matrix being the relationship between the random variables and correlation matrix would be in the range of minus 1 to plus 1, 0 being the case when in they are not correlated, greater than 0, and till plus 1 would be the case when they positivity correlated. And less than 0 and greater than minus 1 would be the case when they are negatively correlated.

(Refer Slide Time: 03:52)

Multivariate Statistical Analysis (Important Definitions contd...)

3) Correlation coefficient matrix: $\rho_{p \times p} = \begin{pmatrix} 1 & \dots & \rho_{1,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,1} & \dots & 1 \end{pmatrix}$, while the sample counterpart is

$$R_{p \times p} = \begin{pmatrix} 1 & \dots & r_{1,p} \\ \vdots & \ddots & \vdots \\ r_{p,1} & \dots & 1 \end{pmatrix}$$

4) Mean: $E(X_j) = \mu_j = \sum_{x_j} x_j Pr(X_j = x_j)$, or $E(X_j) = \mu_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j f(x_j) dx_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j dF_{x_j}(x_j)$, while the sample counterpart is $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{i,j}$, for $j = 1, \dots, p$.

DADM-1
R.N. Sengupta, IIM Dept., IIT Kanpur
311

So, the correlation matrix again would be of size p cross p, where the principal diagonal would, obviously would be the correlation first element would be the correlation of the first with itself. Second element would be the correlation of second with itself so on and so forth. So, the principal diagonal would all be 1 and the half the diagonal element would be the correlation coefficient existing between the ith and the jth. The upper half and the lower half would be the correlation of the jth to the ith, so they would be symmetric to each other.

So, while the sample counterpart would be given by capital R, which is again size p cross p . And again the sample counterpart the correlation for the sample the principal diagonal again would be 1. When the half that the diagonal in mint would be the correlation coefficient from the sample perspective for the i th to the j th, in there upper half and then the lower half. Across in the principal diagonal would be the correlation coefficient of the j th to the i th, and their symmetric if you remember.

Now, if you want to find out the mean of the j th random variable and every discrete case, you will basically multiply the corresponding realized value X_j . So, here it is, so you will basically sum them, and what you will sum is basically this the realize value, which is X_j for all j . So, obviously it will be X_{j_1}, j_2, j_3, j_4 so on and so forth. j_1 with the first reading, j_2 to be the second reading, j_3 be the third reading with j denotes the random variable number, and 1, 2, 3, 4, basically denotes the realest value, multiplied by this corresponding probability. So, this is what it is p probability of X_j is equal to x_j small x_j , which is basically probability of capital X_j is equal to small x_{j_1} or a small x_{j_2} depending on the case. And you will basically multiply that corresponding value probability with the realest value.

So, if I write down the formula, so it will be x_j say for example, let me use the index of 1 probability X_j is equal to x_{j_1} , so for all. So, it would be x_{j_1} probability X_j is equal to x_{j_1} plus I am basically writing here below the second part would be x_{j_2} probability X_j is equal to x_{j_2} , and will go on till that number of terms, and you add them up that becomes the expected value. While if it is a continuous case you have to basically integrate it, so the integration would be in the case let me erase it.

(Refer Slide Time: 07:36)

Multivariate Statistical Analysis (Important Definitions contd...)

3) Correlation coefficient matrix: $\rho_{p \times p} = \begin{pmatrix} 1 & \dots & \rho_{1,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,1} & \dots & 1 \end{pmatrix}$, while the sample counterpart is

$$R_{p \times p} = \begin{pmatrix} 1 & \dots & r_{1,p} \\ \vdots & \ddots & \vdots \\ r_{p,1} & \dots & 1 \end{pmatrix}$$

Handwritten in red: $\int x_j f(x_j) dx_j$

4) Mean: $E(X_j) = \mu_j = \sum_{x_j} x_j Pr(X_j = x_j)$, or $E(X_j) = \mu_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j f(x_j) dx_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j dF_{X_j}(x_j)$, while the sample counterpart is $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{i,j}$, for $j = 1, \dots, p$.

DADM-1 R.N. Sengupta, IME Dept., IIT Kanpur 311

For the case when it is a continuous distribution integrated for all x_j , and it will be $x_j f$ of $x_j dx_j$. So, basically you find out the probabilities for all and find out the sum, which is integration exactly like this. And he will find out the expected value. While the sample counterpart is n as I have already told that would be the sample average. So, multiply the corresponding values for $X_{i,j}$ s. So i being from 1 to n number of observations add them up divide by 1 by n here 1 by n , basically means the corresponding relative frequency or the probability for the $X_{i,j}$ th element. So, if all of them are of equal proportions the probability would be 1 by n add them up divide by n , and get \bar{X}_j .

(Refer Slide Time: 08:44)

Multivariate Statistical Analysis (Important Definitions contd...)

5) Covariance: $Cov(X_{j_1}, X_{j_2}) = E[(X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2}))] = \sigma_{j_1, j_2} = \sum_{x_{j_1}, x_{j_2}} (X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2})) Pr(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2})$, or $Cov(X_{j_1}, X_{j_2}) = E[(X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2}))] = \sigma_{j_1, j_2} = \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} (X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2})) f(x_{j_1}, x_{j_2}) dx_{j_1} dx_{j_2} = \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} (X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2})) dF_{X_{j_1}, X_{j_2}}(x_{j_1}, x_{j_2})$, while the sample counterpart is $s_{j_1, j_2} = \frac{1}{n-1} \sum_{i=1}^n (X_{i,j_1} - \bar{X}_{j_1})(X_{i,j_2} - \bar{X}_{j_2})$, for $j_1, j_2 = 1, \dots, p$.

6) Correlation coefficient: $corr(X_{j_1}, X_{j_2}) = \frac{Cov(X_{j_1}, X_{j_2})}{\sqrt{Var(X_{j_1})} \sqrt{Var(X_{j_2})}}$, while the sample counterpart is $r_{j_1, j_2} = \left(\frac{s_{j_1, j_2}}{\sqrt{s_{j_1, j_1}} \sqrt{s_{j_2, j_2}}} \right)$ for $j_1, j_2 = 1, \dots, p$.

DADM-1 R.N. Sengupta, IME Dept., IIT Kanpur 312

Now, if I consider the covariances, so covariance existing between j th 1 and j th 2 j suffix 1 and j suffix 2, basically denotes the random variable given by the notation of j 1 and j 2 that if obviously we know that the covariance concept gives us, the formula that it is equal to the expected value of the multiplication of the differences. So, what are these differences in case the differences are, so if there are two random variables in the first case the differences are with the random variable X_{j1} with respect to its mean value, so all the real value as they are changing I take the mean may the differences.

And in the second case it is basically again the differences of the all the random variables of the j 2 with respect to its own mean value, which is $x_{\mu j2}$ or \bar{X}_{j2} whatever the case is. So that is given by the covariances and the corresponding variances are, when both X_{j1} and x_{j2} are same. So, they would basically be the principal diagonal and we can find them out.

And in the case when you have the continuous sum for the continuous case, you will basically double integrate there will be double integration. Only remember one important thing the integration would be for the case, when you want to find out X_{j1} minus is the expected value multiplied by X_{j2} minus is expected value, but the c to the f of x would, now be replaced by f of x_{j1} and j 2, because that is the binomial distribution based on the fact that X_1 and X_2 a random variable. So, based on that you find out the covariance is for the continuous case.

And similarly the counterpart the covariances in the in the for the sample would be again given by 1 by n minus 1 provided you would use 1 degrees of freedom, and the multiplicative term are X_{j1} minus is sample mean multiplied by X_{j2} minus it is corresponding sample mean. Similarly, covariances can be found out by the ratios of the covariance of $X_{j1} X_{1,j1}$, and j 2 divided by the square root in one case there would be 2 square roots. So, the square root would be the variance of the of j 1; and second one would be the variance of j 2, which is the other one.

So, while the sample counterpart which will be denoted by small r . So, if you remember the capital R that was basically the matrix corresponding to the correlation coefficient for all the p number of variables. So, this sample counterpart r_{j1} and r_{j2} would be equal to the covariance of the samples which is in the numerator. So, this would be mimicking this, this one would be mimicking this one, and this one would mimicking this one. So,

once you have this issue you can find out the correlation coefficient existing between two random variables.

(Refer Slide Time: 12:36)

**Multivariate Statistical Analysis
(Important Definitions contd...)**

7) Co-skewness: $E\{(X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2}))(X_{j_3} - E(X_{j_3}))\} = \sum_{x_{j_1}, x_{j_2}, x_{j_3}} (X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2}))(X_{j_3} - E(X_{j_3}))Pr(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}, X_{j_3} = x_{j_3})$ or

$E(X_{j_1}) \int_{x_{j_2, \min}}^{x_{j_2, \max}} \int_{x_{j_3, \min}}^{x_{j_3, \max}} (X_{j_2} - E(X_{j_2}))(X_{j_3} - E(X_{j_3})) f(x_{j_1}, x_{j_2}, x_{j_3}) dx_{j_2} dx_{j_3}$, for $j_1, j_2, j_3 = 1, \dots, p$.

Note: Co-skewness is related to skewness as covariance is related to variance

8) Skew relation: $\frac{E\{(X_{j_1} - E(X_{j_1}))(X_{j_2} - E(X_{j_2}))(X_{j_3} - E(X_{j_3}))\}}{\sqrt{E(X_{j_1} - E(X_{j_1}))^2} \sqrt{E(X_{j_2} - E(X_{j_2}))^2} \sqrt{E(X_{j_3} - E(X_{j_3}))^2}}$, for $j_1, j_2, j_3 = 1, \dots, p$.

DADM-1 R.N. Sengupta, IIM Dept., IT Kanpur 313

So, now we see the co skewness, obviously we did not go into details, but I am just giving of this definitions just for your own interest. So, the co-skewness basically consists of the case where you find out the third moment help based on the fact that I want to find out the co-skewness between j_1 , and j_2 , and j_3 .

And in the case when if it is discrete case you will basically find out the differences between the j_1 in its mean value j_2 and its mean value j_3 , n is mean value multiply them. And then multiply by the corresponding multivariate values when X_{j_1} is equal to X_{j_1} X_{j_1} small x_{j_1} , capital X_{j_2} is equal to small x_{j_2} and capital X_{j_3} is equal to small x_{j_3} , and based on that fact how the probability changes we multiply the probability to find out the co-skewness. In the case if it is a continuous distribution rather than finding on the sum we replace it with the corresponding integration. So, the integration is given; I am just highlighting the relevant part.

So, this is for the second. If you see the colouring scheme will understand, which variable and what are the counter points or the formulas based on which you are doing the calculation. And obviously this is the joint distribution of X_1, X_2, X_3 . And remember co-skewness is related to skewness in the same way as covariance is related to the variance. When we found out the skewed relationship, the correlation coefficient on

skewed relationship again is basically the co-skewness divided by the square root 3 terms.

So, the first square root is basically the standard deviation of the jth j_1 , then the second element is basically the standard deviation of j_2 , and the 3rd element is basically the standard deviation of j_3 based on that we would divide. The skewness by this values, which are technically the standard deviation of 1, 2, 3 multiplied you divide, and you get the skew relations.

(Refer Slide Time: 16:00)

**Multivariate Statistical Analysis
(Important Definitions contd...)**

9) Co-kurtosis:
$$E\{[X_{j_1} - E(X_{j_1})][X_{j_2} - E(X_{j_2})][X_{j_3} - E(X_{j_3})][X_{j_4} - E(X_{j_4})]\} = \sum_{x_{j_1}, x_{j_2}, x_{j_3}, x_{j_4}} [X_{j_1} - E(X_{j_1})][X_{j_2} - E(X_{j_2})][X_{j_3} - E(X_{j_3})][X_{j_4} - E(X_{j_4})] Pr(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}, X_{j_3} = x_{j_3}, X_{j_4} = x_{j_4})$$
 or
$$\int_{x_{j_4, \min}}^{x_{j_4, \max}} \int_{x_{j_3, \min}}^{x_{j_3, \max}} \int_{x_{j_2, \min}}^{x_{j_2, \max}} \int_{x_{j_1, \min}}^{x_{j_1, \max}} [X_{j_1} - E(X_{j_1})][X_{j_2} - E(X_{j_2})][X_{j_3} - E(X_{j_3})][X_{j_4} - E(X_{j_4})] f(x_{j_1}, x_{j_2}, x_{j_3}, x_{j_4}) dx_{j_1} dx_{j_2} dx_{j_3} dx_{j_4}$$
 for $j_1, j_2, j_3, j_4 = 1, \dots, p$.

Note: Co-kurtosis is related to kurtosis as covariance is related to variance

10) Kurtic relation:
$$\frac{E\{[X_{j_1} - E(X_{j_1})][X_{j_2} - E(X_{j_2})][X_{j_3} - E(X_{j_3})][X_{j_4} - E(X_{j_4})]\}}{\sqrt{E\{[X_{j_1} - E(X_{j_1})]^2\}} \sqrt{E\{[X_{j_2} - E(X_{j_2})]^2\}} \sqrt{E\{[X_{j_3} - E(X_{j_3})]^2\}} \sqrt{E\{[X_{j_4} - E(X_{j_4})]^2\}}}$$
 for $j_1, j_2, j_3, j_4 = 1, \dots, p$.

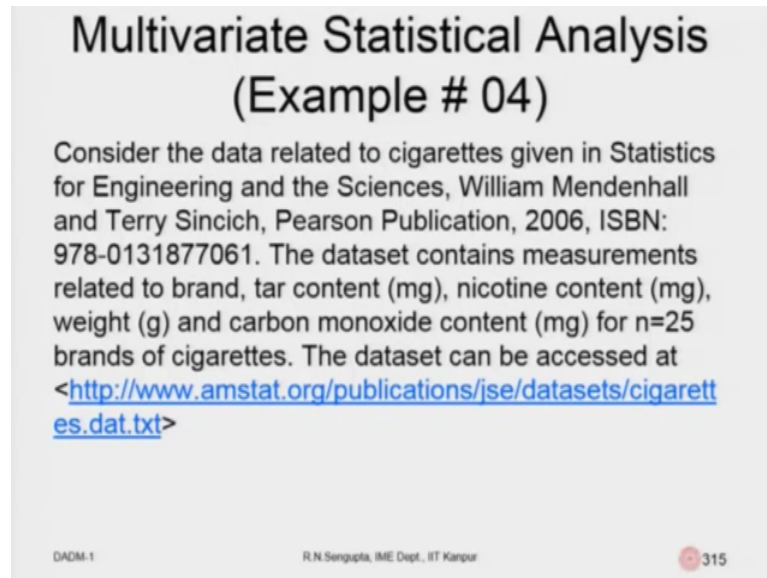
DADM-1 R.N. Sengupta, IIM Dept., IT Kanpur 314

Now, co-kurtosis basically is again the same thing, but you own is increase the number of random variables. So, in skewness your 3, now your 4 X 1, X 2, X 3, X 4. So, in the case if it is discrete, again you follow the same rule find out the difference between the random variable the realize value n is expected value. So, this will be true for the 1st random variable, 2nd random variable 3rd random variable, 4th random variable. In the case if it is discrete, you multiply them by the corresponding probability add them up. In the continuous case you integrate it, but the joint distribution of all the four random variables, so you should be considered so that we get the value of the co co-kurtosis.

So, similarly co-kurtosis is related to kurtosis as covariance is related to variance, and I did mention that so, in order to basically bring a smile or the understanding how things are done. And then again you have the Kurtic relationship Kurtic relationship would basically be the same concept the co-skew, co-kurtosis would basically be on the

numerator. And in the denominator, obviously you will have the standard division of the 1st standard division, 2nd standard division of the 3rd standard division of the 4th and based on that you will basically do your problem.

(Refer Slide Time: 17:25)



**Multivariate Statistical Analysis
(Example # 04)**

Consider the data related to cigarettes given in Statistics for Engineering and the Sciences, William Mendenhall and Terry Sincich, Pearson Publication, 2006, ISBN: 978-0131877061. The dataset contains measurements related to brand, tar content (mg), nicotine content (mg), weight (g) and carbon monoxide content (mg) for n=25 brands of cigarettes. The dataset can be accessed at <http://www.amstat.org/publications/jse/datasets/cigarettes.dat.txt>

DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 315

So, now we will consider the data set related to cigarette smoking, which is given in Statistics for Engineers and Sciences by William Mendenhall and Terry Sincich. So, this is a Pearson Publication, 2006 the dataset contains measurement related to brand tar, content, nicotine content then the weight and carbon monoxide content for n for 25 brands of cigarette. So, the data set can basically be access accessed in the American statistical website, which is amstat dot org, and you have the publication based on that you can find out the data.

(Refer Slide Time: 18:09)

Multivariate Statistical Analysis (Example # 04 contd..)

Considering the variables, $p = 4$, i.e., X_1, X_2, X_3 and X_4 as tar content (mg), nicotine content (mg), weight (g) and carbon monoxide content (mg) respectively, one obtains

$$R = \begin{pmatrix} 1 & 0.9766 & 0.4908 & 0.9575 \\ 0.9766 & 1 & 0.5002 & 0.9259 \\ 0.4908 & 0.5002 & 1 & 0.4640 \\ 0.9575 & 0.9259 & 0.4640 & 1 \end{pmatrix}$$

DADM-1 R.N. Sengupta, IIM Dept., IT Kanpur 316

So, consider there are basically 4 variables $p = 1, 2, 3, 4$, which are X_1, X_2, X_3 and X_4 , which are basically the tar content, the nicotine content, the weight and carbon monoxide content. So, based on that if we have for the 252 data sets port them. Find out the correlation coefficients and the values as rightly shown in this diagram. The principle diagonal is all 1, because they are the correlation of the first to first, second to second, third to third, fourth to fourth. And the other diagonal element are the correlation existing between in the first to second, first to third, first to fourth. Similarly, for second to first would be the counterpart on the mirror image. Then it would be second to third and second to fourth.

Similarly, you can find out all the correlations they would be mirror image. So, as pointed out. So, let me show this one this one a same, this one delete this one, this one this one I am just showing you the mirror image concepts I am sorry this will be the case, my apologies. So, they are the mirror image, and you have found out. So, they can basically give you the relationship correlation coefficient existing between the terms.

(Refer Slide Time: 20:08)

Multivariate Statistical Analysis (Example # 04 contd..)

Furthermore utilizing R we get the multiple correlation

coefficient vector as $\begin{pmatrix} 0.9867 \\ 0.9774 \\ 0.5001 \\ 0.9584 \end{pmatrix}$ while the corresponding R^2

values are 0.9720, 0.9554, 0.5366 and 0.9174.

DADM-1 R.N Sengupta, IME Dept., IIT Kanpur 317

Furthermore, utilizing R we get the multiple correlation coefficient factors so which are given as 0.9867 till the first one till 0.9584, while the corresponding R square values. What are these I will going to come back to that later on. And there corresponding R square values are 0 I will only read the first two decimals which is 0.97,0.95,0.53,0.91 and based on that you can do the calculations accordingly.

(Refer Slide Time: 20:42)

Multivariate Statistical Analysis (Example # 04 contd..)

Simple calculations would also yield the partial correlation

coefficient matrix as $\begin{pmatrix} 1 & 0.8199 & -0.0141 & -0.6556 \\ 0.8199 & 1 & -0.1092 & 0.1465 \\ -0.0141 & -0.1092 & 1 & 0.0072 \\ -0.6556 & 0.1465 & 0.0072 & 1 \end{pmatrix}$.

To double verify the calculation we may also calculate the partial regression coefficients.

DADM-1 R.N Sengupta, IME Dept., IIT Kanpur 318

Simple calculation would also yield the partial correlation. Now, remember the important fact the partial correlation coefficient and the full correlation coefficients are different.

So, partial correlation is exactly something to do the fact that I am keeping all the other variables fixed. And basically I am trying to change the one of the variables and based on that I am trying to find out the partial correlation coefficient. So, partial correlations would basically the relationship between two random variables provided others are kept fixed. So, the partial correlation coefficient matrix is given. Again it is a mirror image.

So, the first one first one, second one second one, I am just highlighting it third one third one. So, these are the elements in the partial correlation coefficient matrix to dark, but double verify the calculations. We may also calculate the partial regression coefficients, and solve our problem, but I am not going to come into this details as of now.

(Refer Slide Time: 22:14)

**Multivariate Statistical Analysis
(Multinomial distribution)**

- Suppose X_1, \dots, X_p be p jointly distributed random variable each of which is discrete, non-negative and integer valued
- Then the joint probability mass function of X_1, \dots, X_p is called the multinomial distribution and is of the form $\binom{n}{x_1, \dots, x_p} p_1^{x_1} p_2^{x_2} \dots p_p^{x_p}$

$\sum p_1 + p_2 + \dots + p_p = 1$

DADM-1 R.N. Sengupta, IIM Dept., IT Kanpur 319

So, we will consider the multinomial distribution for the multivariate case. So, suppose X_1 to X_p be p jointly distributed random variables each of which is discrete, non negative and integer valued. Then the joint probability mass function. So, now, what you have if you remember the binomial case, there were two random variables. Now, in the multinomial case you are basically in a trinomial you have 3 random variables.

And the Bernoulli trials and all these things would hold true that means the values of the random variables X_1, X_2, X_3 would be coming out the corresponding probabilities would be independent from through to through or roll to roll on the outcomes. And the corresponding sum would also be 1 and that means, if their probabilities p_1, p_2, p_3 some of that p_1, p_2, p_3 would be 1, because there has to be any one outcome out of the

three considering is a trinomial distribution. And we will find out the corresponding distribution for the case for the for the multinomial case.

So, as I was saying suppose X_1 to X_p are p jointly distributed random variables each of which is discrete non-negative and integer valued. Then the joint probability mass function of X_1 to X_p is called the multinomial distribution on the form. So, here the form is. So, it is whatever I written I am just trying to highlight. This is $n C x_1, x_2, \dots, x_p$. And probably places p_1 to the power x_1 , one to read p_2 to the power x_2 , which is p_p to the power x_p .

And the sum of p_1 plus p_2 plus equal to 1 that is true in this formula, which I have written is for the case of this distribution. So, this is the multinomial case. So, probabilities are given by p_1, p_2, p_3 to p_p or p_k whatever it is. And the corresponding numbers are basically given as x_1 and x_2, x_3, x_4 to x_n .

(Refer Slide Time: 24:44)

Multivariate Statistical Analysis
(Multinomial distribution contd..)

For a better appreciation of the multinomial distribution consider the polynomial coefficients of the expansion of the multinomial expansion, $\{p_1x_1 + \dots + p_px_p\}^n$

$\{p_1x_1 + p_2x_2 + \dots + p_px_p\}^n$

DADM-1 R.N.Sengupta, IIM Dept., IIT Kanpur 320

So, for a, but better appreciation on the multinomial distribution consider the polynomial coefficient of expansion the multinomial case. So, polymer expansion would be for p_1 sorry I am using the highlighter my mistake. So, $p_1 x_1$ plus $p_2 x_2$ plus $p_p x_p$ expanded in the multinomial case and you will get the multivariate distribution.

(Refer Slide Time: 25:43)

Multivariate Statistical Analysis (Multinomial distribution contd..)

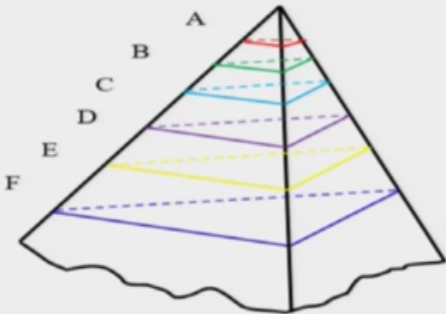
A closer look at this expansion will make it immediately evident how the polynomial coefficients of the multinomial expansion correspond to the multinomial distribution discussed above. Another interesting analogy for multinomial distribution can be made from the Pascal Pyramid

DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 321

Now, a closer look at the expansion will make it immediately evident, how the polynomial coefficient the multinomial expansion corresponds to the multinomial distribution as case. And another interesting fact is basically its analogy with the multinomial distribution has with the Pascal pyramid.

(Refer Slide Time: 26:08)

Multivariate Statistical Analysis (Multinomial distribution contd..)



DADM-1 R.N.Sengupta, IME Dept., IIT Kanpur 322

So, Pascal pyramid is a pyramid and with three sides it is looked like this. So, the slices which are there in red colour, green colour, light blue, violet, yellow, dark blue are as you expand the numbers higher on. So, as you go for the higher numbers of the Pascal values

as you go down technically they would basically correspond to the fact that you have higher values of n.

(Refer Slide Time: 26:34)

Multivariate Statistical Analysis
(Multinomial distribution contd..)

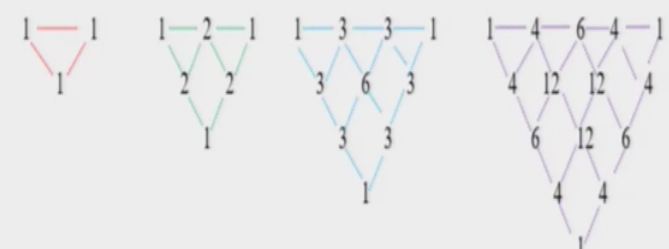
If we view the slices (as represented by A, B, C, D, E, F and so on) of the Pascal Pyramid as flat triangular plates, then the numbers depicted on them are as follows

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 323

And if you basically take the slices as I mentioned, as represented by A, B, C, D, E, F of the Pascal pyramid as a flat triangular plate. So, they are basically a pyramid of three sides the base is also a triangle not the square one.

(Refer Slide Time: 26:57)

Multivariate Statistical Analysis
(Multinomial distribution contd..)



The image shows four triangular slices of Pascal's triangle, each with numbers and connecting lines. The first slice has 1s at the top and bottom corners. The second slice has 1s at the top corners and 2s in the middle. The third slice has 1s at the top corners, 3s in the middle, and 3s at the bottom corners. The fourth slice has 1s at the top corners, 4s in the middle, 6s in the middle, 4s at the bottom corners, and 1 at the bottom center.

MBA651 R.N.Sengupta, IME Dept., IIT Kanpur 324

Then the numbers depicted on them would be as follows. So, they would basically give you the numbers for the polynomial expansion. The first one would be corners we do 1,

1, 1. Second corners which is the green slice which you have taken would be corners would be 1, 1, 1 and the midpoints would be 2, 2, 2. So, they are corresponding to the coefficients. In the third case blue one you will have corners again 1, 1, 1, And the midpoints would be again between the sides of the triangle 3, 3, 3; and the middle point where the intersections happens for the 3, 6. So, they are basically corresponding to the coefficients of the multinomial polynomial expansion. Similarly, for the other case triangles for the fourth level ABCD, so we will basic again the corner points as 1 and the in between points of 4, 6, and 12 would basically corresponding to the or the values of the polynomial coefficients.

So, with this I will close the 30th lecture. And continue the discussions of the multivariate distribution in the corresponding one, and go a little bit slow, because there are lot of concept to be covered. And we will try to basically make it more hands on depending on definitely at the feedback which you give, and I am sure it will be helpful for all of us to make it in the right fashion. Thank you very much with this your attention, and have a nice day.

Thank you.