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

Lecture 30
Multivariate Statistical Analysis

Very warm welcome my dear students, a very good morning, good afternoon, good evening to all of you. This is the DADM which is Data Analysis and Decision Making -I lecture series under NPTEL MOOC. And as you know this is a 12 week course for 30 hours, total number of lectures is 60, each week we have 5 lectures and each lecture is for half an hour. And I am Raghu Nandan Sengupta from the IME department, IIT Kanpur.

So, we are in the thirtieth lecture; that means, we are going to end the 6th week and; obviously, I am sure you are doing your assignments properly trying to understand that any queries based on the assignment please send them on the forum we will try to answer it as soon as possible. And please read the books if possible, they are definitely in depth. But, generally go through the books, try to get some data sets in different fields as I said and in the net is full of datasets trust me.

Play with them, try to plot the distribution, try to plot the cumulative distribution function, try to plot the mean, how the mean looks, median looks, mode looks. Do some take some data, do some have this testing, interval estimation, point estimation. Try to plot the different distributions for the discrete case, then the continuous case for the F-distribution, T-distribution, chi square all these things.

So, what I am telling is basically if you remember on the last 2 lectures in the 29th also I gave a brief preview that where we are heading, what we have covered. And if you remember in the 30th lecture, I did say when I will just try to basic take a stock of the situation, get your feedback, understand how things are going.

So, it will be much easier for me and definitely for my TAs to understand the course deliverance from your point of view and how it can be fine tune and make much better. So, initially part we did discuss about probability and then we did discuss about the simple data interpretation, trying to draw using the pie chart, histogram, spatial diagrams

Then how you can basically draw the stack diagrams and all these things for different data sets. Then how to find out the cumulative distribution function, the pdf, the pmf, what is the mean, what is the mode, what is the median and what are the significance of them.

Then we went into trying to basically find out what is the expected value, variance, I did not give the formula for kurtosis and skewness because they are the higher moments, but definitely I did give the skewness kurtosis concept. And I will come back to that later also in the multivariate case.

So, then we discussed about probability in general sense, what is the axiomatic axioms of probability, then what is a relative frequency concept and what is the general definition of probability and difference between axiomatic probability definition, what are the probabilities are properties of probability. We that also then we went into different type of discrete distributions, the uniform discrete, the partial distribution, the hyper geometric.

We later on consider the geometric distribution, negative binomial and I gave how the distributions look like, what is the pdf of that. Obviously, I did not go to a cdf because you can just simply find them. Then I gave few very simple examples for each and we also discussed the relationship of many of the cases the binomial, then hyper geometric and all these things as sample size increases.

Then we went into the continuous distribution, in the continuous distribution case I mentioned about the exponential distribution, normal distribution and I also mean discussed about the log normal distribution. And I have spent a lot of time on normal distribution I understand, I remember. Then we discussed what is the concept of base theorem, conditional probability, why condition probabilities required and when a simple problems were discussed.

So, as I am saying that for each of the concept we went to small problems. Then we went into the areas of what is a population, what is the sample, different type of sampling techniques (Refer Time: 05:14) sequential sampling concept, judgmental sampling concept, cluster sampling concept and all these things.

Then we came into the 3 main distributions the 3 being F-distribution, chi square distribution and F-distribution. We discussed how the distributions look like, what is the concept of degrees of freedom, what is the pdf of the distribution, what is the expected value of the distribution, what is the variance of the distribution.

I gave only the end result, did not go through the derivation of proof, because that is not required aside this is not a hardcore theoretical statistic course. Then we went into discussing the concept of point estimation, in the concept of we also discussed 2 of the inequalities Markov inequality and Chebyshev's inequality. Then for the point in distribution I did discuss what is unbiasness, what is consistency, what are the significance of them; then, for point distribution for different distribution as such;

We and also hypothesis testing before that we came up with the different rules. Rules means that what is the best estimate for the population mean given the, what is a given? The sample is there and population means not known. Both in the case when the standard deviation for the population is known, standardization population is not known.

Then we went into find out that what is the best estimate for the sample variance or standard error square for a sample given the population mean known and population mean not known. We and we also understood what how we lose one degrees of freedom, then we consider the different rules 4 different rules based on which you can proceed and formulate and formulate the different type of F, chi and T-distribution from the for the sample distribution case.

Then we consider the point distribution, simple examples the how the sample means can be taken as the best estimate for the population mean for the say for example, from the normal distribution, Poisson distribution, exponential distribution. Then we went into interval estimation, discussed about lower control limit, upper control limit, what is the concept of level of significance. Level of confidence that whole area being 1 minus alpha and the left and right portion of that band which is between lower and upper control limit both.

So, left and right of that portion the total area would be alpha because the total area under the pdf of the pmf should definitely be 1. Then we also considered that given the, we given the main fact that we want to find out something to do with the mean for the population. We understood in details that how we could use the Z-distribution or the T-

distribution. In the Z-distribution there is no degrees of freedom concept, in the T-distribution there is a degrees of freedom concept because, we lose one degrees of freedom and minus 1 if it is only to do with one population.

Then for the trying to find out the ratios of the standard deviation on the variant square for the population, we took the concept that you can use the chi square distribution with the n minus 1 degrees of freedom, chi square degrees distribution with only n degrees of freedom. So, the second case with n degrees of freedom being true when the population mean is known as you know.

And we use s dash square if you remember, then the chi square with n minus 1 degrees of freedom would be used for the case when it is s without the dash square because, you are losing one degrees of freedom. Then we went into trying to understand that how the F-distribution could be used in the case when we are trying to compare 2 different variances from 2 different populations. And that can have F- distribution with m comma n degrees of freedom provided, but the population mean from both the populations are known.

And in the case the population means from both the distributions are unknown we will use the F-distribution with m minus 1 and n minus 1 degrees of freedom. Then in the case when you want to find out the difference of the means provided the sample variances are known or not known, then we will basically be utilizing the Z-distribution or the T-distribution.

In Z-distribution again we know that there is no loss of degrees of freedom, but in the T-distribution when we are comparing the differences between 2 means of 2 populations the degrees of freedom lost would be n plus m minus 2. n being the sample size for the first population from the first population and m being the sample size from the second population. So, the total degrees of freedom lost would be 1 1 in each.

Then later on we went with the hypothesis testing, for the hypothesis testing we use the same concept as being used in the interval estimation. But, in the hypothesis testing we have 3 different outcomes of the rules based on the estimation problem. One would be in the case that when we had the less than type, greater than type, that is the second one and in not equal to is the third one.

But; obviously, before that we understood what is the level of significance, what is the type one error, type 2 error, alpha, beta given h naught and h a h naught being the null hypothesis h a being the alternative hypothesis. So, there were errors and I have explained that with a simple example like you being a banker, you are trying to give a loan to people with certain scores and how it can be differentiated that alpha and beta are the errors.

Then in again in following the same policy trying to find out something too with the sample mean we use the Z-distribution of the less than type, greater than type, not equal to. Then we use trying to find out something to the mean or the difference of the mean, we use either the Z-distribution or the T-distribution with degrees of freedom being n minus 1, if there is only 1 population and n plus m minus 2, if there are 2 population.

Again of the less than type, greater than type, not equal to then for when we went with the variances we consider the ratios of the variances provided the population mean were known or not known. If they are known then; obviously, use the chi squared with n degrees of freedom again less than type, greater than type, not equal to.

Then for the case when you want to find out something to do with the standard deviation of the variance of one population given the population mean unknown, then again we use the chi square with n minus 1 degrees of freedom; that means, we are losing 1 degrees of freedom of the less than type, greater than type and not equal to.

Then when we came to the F-distribution when you are trying to find out the ratios of the variances of 2 different populations given the population mean in both the cases being known. We use the F-distribution with n comma n and for the less than type, greater than type and not equal to. When we came into case of trying to find out something to do with the variances ratios of the variances provided the population means are not known then we use the F-distribution.

Again with n minus 1 and m minus 1 being the degrees of freedom again for the less than type, greater than type, not equal to. Now, in all this case for the Z, for the T, for the chi, for the F, I mentioned the degrees of freedom. But, I did not mention anything I did mention again before that, but I did not mention for the time being that what would be in that value of 1 minus alpha.

So, if you remember if it is a less than type or greater than type we will use the value of 1 minus alpha or alpha accordingly. And if it is of not equal to then it will basically be 1 minus alpha by 2 and alpha by 2 corresponding to the fact that we are trying to quickly divide the area on the left hand side our right hand side in equal proportion such that the total area and it is up to 1.

So, how does it add up to 1? Alpha by 2 plus 1 minus alpha plus alpha by 2 that total area should be 1. Then later on after this hyper testing we went to the multiple linear forecasting methods. In the forecasting methods we considered that given different type of values of Y we use the weighted method, weighted simple moving average method and weights were given to the past data.

Then we use the exponential smoothing where the weights are given both for the predicted data and the actual data for 1 period, 2 period, 3 period depending on how we want to basically be accurate. But, we always remember that the sum of the weights is equal to 1 and basically we are trying to base minimize the sum of the square as with the errors. Hence, we will try to partially differentiate the parameters with respect to the partial differentiate the sum of the square of the errors with respect to the parameters, put them to 0 and basic you solve them to find out the alpha hats or beta hats whatever is required.

Then we went into trying to consider the trend effect. So, there was a trend, trend this seasonality can also be considered. So, we considered all these things, but with the main focus being that the parameter should be such that they are able to minimize the sum of the square of the errors 0.1. And then the fact was that the sum of the parameters should be 1. Then in the multiple linear regression con concept by the way we did not consider in an inherently any distribution for the forecasting case.

Then we went into the case of multiple linear regressions, for the simple linear regression we considered only one X independent variable and one dependent variable. We consider alpha as the point where the line the average line. Average line being Y is equal to m X plus see the example which I kept giving time and again.

So, that where it cuts the Y axis that was the value of alpha and beta was basically the rate of change of the value of Y with respect X, as X changes by 1 unit. Then we assumed; obviously, there is an error and error I would have a particular distribution

which was a mean value of 0 and a variance of 1 or sigma square and the distribution was normal in case.

And then in the later I will come to the assumptions in totality, then we went to the multiple linear regression, we considered that there are more than one X, that is X1, X2, X3 till Xk, suffix k or X suffix p. And; obviously, there is a alpha or a beta naught depending upon the problem has been formulated. And; obviously, there would be an error, epsilon and utilizing all these things we will try to basically predict for Y which is the dependent variable.

Then; obviously, we had few assumptions, the assumptions were that the rank of the matrix was k or p of the X matrix and what is that we did not remember. That all of them should be independent on each other and none of the rows, none of the columns should be expressed as a linear combination of the rest.

We considered the covariance's between the error themselves from time period 1, 2, 3, 4 would be 0 that they are independent. The covariance's between X's and epsilon or the errors would be 0 also; that means, the errors do not affect the X's. And; obviously, as it means the ranks or rank of the x matrix is k; that means, they are independent. Hence the covariance is existing between all the X's between themselves is also 0.

So, based on that we found out what is the mean value of Y, we found out what is the variance of Y. And then I had the main focus was basically to minimize the sum of the squares of the errors, differentiate with respect to alpha and beta, put it to 0 and then find out the values of the estimate of alpha and beta.

So, those are denoted by alpha hat, beta hat whatever it is. But, this partial differentiation rather than differentiation partial differentiation was important. Because, you want to basically find out the rate of change of these functions keeping the change of Y with respect to any one X considering the rest of the x values are not changing, they are fixed.

And then in the last 2 class in the 29th class and the 28th class we spend a lot of time and trying to understand what are the q plots, how normality could be understood. And how the one to one correspondence is correspondence between the cdf of a normal distribution and cdf of a non unknown distribution can be mapped in order to basically find all the quantile-quantile plots.

And then in the in the in the fag end of the 29 th lecture, I started basically giving 3 different examples of a dietician of a person who is a realized state agent who wants to basically find out the price of the house. And third being a faculty, who wants to find out how does the prices of different stock move. Are they moving in tandem? Considering they are from the same sector on all these things.

We consider these 3 examples and I told you also in all these 3 examples the main focus was to basically find out some relationship or find out some distribution based on the fact that there are more than one number of random variables. And they are basically X1 to Xp or Xk suffix Xk or a suffix Xp which would give us a whole lot on information related to the pdf of some distribution, related to multiple linear regression. Related to say for example, other methods of anova, manova, factor analysis, cluster analysis, conjoin analysis, structural equation modeling all this will be coming and.

Here now I will basically say the overall jump of the quantum of concept of in the multivariate case starting from today which is the 30th lecture would be a little bit different. So, I would rather maybe in many of the cases I would try to go slow, a bit slow and even if an I am fast please bear with me.

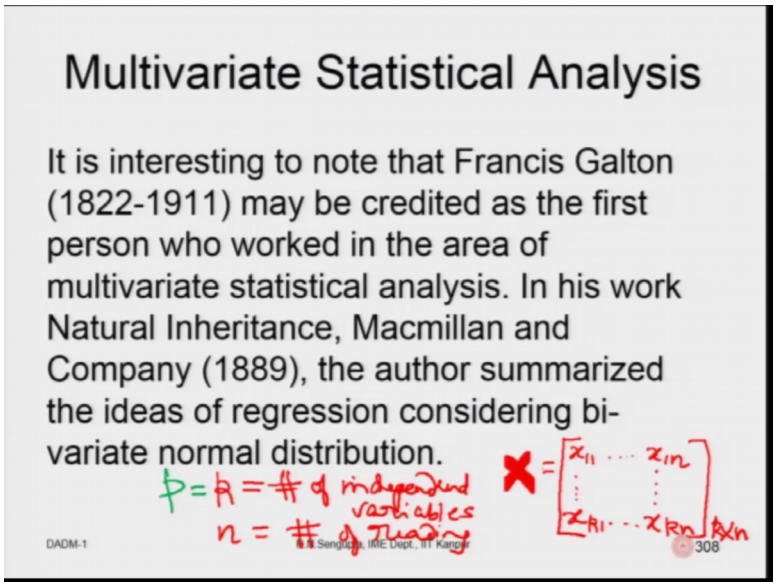
We will try to basically cover the basic concepts which will give a lot of information for the students who are taking this course in order to basically build up their concepts after somebody goes to this DADM-I course. Obviously, I am not saying this is the best set of topics where I am are going to cover, I am able to give the best examples.

But, I am trying to basically in calculate the interest in all of you such that you can pick up any good books and start reading it such that it will give you a good base for the area of data analysis, decision making, multivariate analysis, multiple linear regression, structural equation modeling, conjoint analysis, factor analysis and all these things.

So, with this it is basically a brief wrap up of the all the 29 classes which you have covered. So; obviously, should it would have been at the middle of the 30th lecture. But, I thought I will just give you a brief background so; that means, once we are you are able to appreciate where you stand and how the teaching has been going on you would basically feel much more comfortable. Then what things you are going to face and how you are going to tackle the problems accordingly.

So, coming back to the multi weight statistical methods if you remember I did not mention that in the last slide of the 29th class that given one X, which is the matrix which has basically p number of random variables. All of them are independent of each other and each periodic number of random variables have basically n number of reading.

(Refer Slide Time: 21:08)



So, it will be as so, capital X which is a vector I should write it like this which hopefully this is I am able to denote this as a bold that will be equal to. So, the nomenclature may change, but the general essence remains the same. So, it will be X 1 1 till 1 X 1 n and this will be X k 1 X k n.

So, there are basically if we consider if you consider the number of rows and number of columns it will be of the size k cross n. So, k or in many of the examples later on you will see p. So, k is basically the number of independent variables and n is number of reading and this in many of the problems this will be given as p also.

So, whether use p or whether use k would hardly make any change in the understanding. So, given this matrix X we want to basically predict or find out from pdf or pmf of the actual set of observations. Because, there are different random variables which affect the reading and we want to find out the relationship.

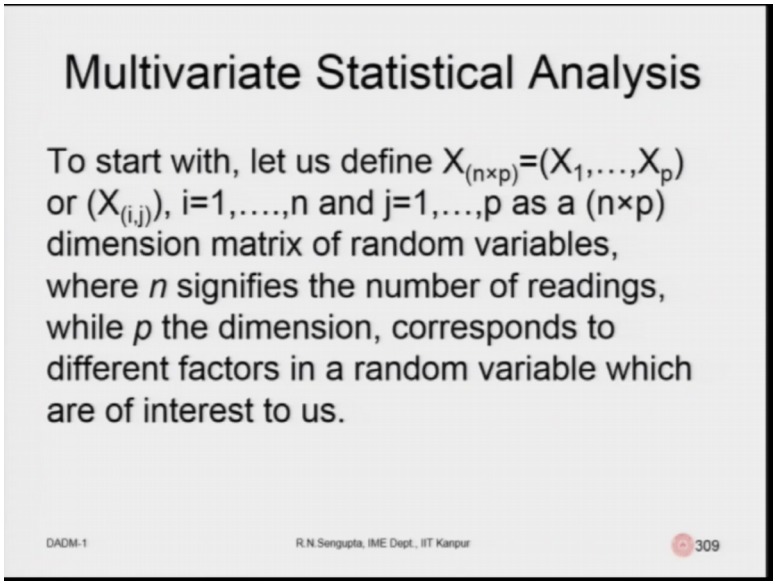
So, it is interesting to note that Francis Galton may be credited as the first person who worked in the area of multivariate statistical analysis. In his work natural inheritance

which was published by Macmillan and company. The author basically summarized the ideas of regression considering bi-variate normal distribution and he considered the children born to parents and depending on their heights.

Basically the multivariate, the bi-variate, multi normal distribution was basically plotted or given an idea such that one could understand that given 2 variables which are normal, how you can basically find out the interrelationship between them which is a bi-variate normal distribution.

So, to start with again I will rehash. So, let us consider this matrix X so, all these are bold remember X is a bold is a matrix, while these X1 till X p are basically the vectors. So, they would also be bold.

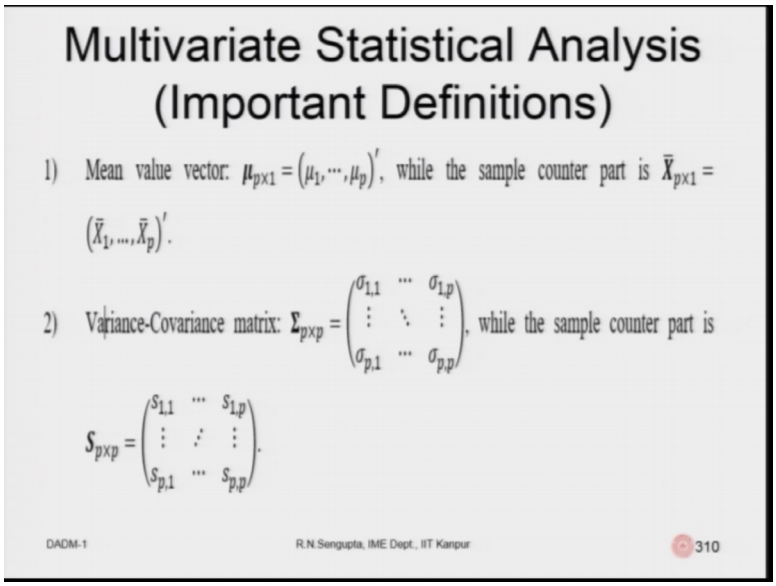
(Refer Slide Time: 24:17)



So, let us define X which is of size n cross p which consists of p number of random variables. So, each element will be denoted by X i j, i is basically from one to n j is equal to 1 to p. So, p is the number of random variables, n is the number of readings.

So, here X i j, i is equal to 1 to n, j is equal to 1 to p is basically the n cross p dimension matrix of random variables, where n signifies the number of readings, where p basically signifies the dimension. So, and it corresponds to different factors in a random variable which are of interest to us based on which we will try to basically do our studies.

(Refer Slide Time: 25:01)



So, we will basically some nomenclature and definitions. So, definitions will be coming time and again and will try to basically give an example of the problems. Rather than going to the details solving of them because, if the consist of lot of theorems which I do not want to go and proves. Obviously, they were proves in the univariate case also. But, we skip that and here also we will skip that, but, but the initial pace would be more of the discussions of the distributions of the concepts with solved examples and the results analysis.

So, the mean value of this multivariate some distribution for each one X 1 to X p, they would be mean value given by mu 1 to mu p. So, we will basically denote mu 1 to mu p collectively as a vector of mu, size p cross 1, where the first element is the mean value of the first random variable X 1. Second mean value is basically the second and the first and the actually the mean of the second random variable and so, on and so, forth.

While when we consider the sample counterpart in the sample counterpart would basically have the sample means. The sample means would be given by X 1 bar for the first one, X 2 bar for the second one till the last one which is X n bar for the X bar know X p bar, sorry my mistake not X n bar X p bar for the p'th variable. Now, the variance covariance matrix technically would be given by matrix of size p cross p and the p cross p would basically have elements along the principle diagonal. It will be the covariance of

the first I am talking about the elements. So, 1 comma 1, then 2 comma 2, 3 comma 3, till p p comma p which is the principal diagonal.

The first element in the principal diagonal would be the covariance of the first with itself which is basically the variance of the first. The second element which basically would be 2 cross 2, 2 comma 2 would be the covariance of the second with itself which is the variance of the second. Till the last one would be which will be the covariance of the p'th one with itself. And that would basically the variance of the p'th one and the of the diagonal element would be all mirror image.

So, the i j element would be exactly equal to the j i element, where i j element would be the covariance's of the i'th and the j'th one, random variable. Similarly, the j i element would be the same and the values would be the exactly the same so, they are just mirror images. While when we consider the sample counterpart we will basically take this in the square of the standard errors and will use the same formula whether using s dash, s would basically depend on whether the mean value is known not known.

So, again it will be a p cross p s where matrix, principal diagonal would exactly be the standard error squares. The first element would be basically s 1 1, which is the standard errors square or the covariances between the first 2 first which is basically the variance standard error square or the sample variance of the first.

Similarly, the 2 comma 2 would be the sample variance of the second and the last element would be the sample variance for the nth p'th one. And again the off the diagonal element would be mirror image of each other, they would basically if the covariance's for the sample 1 and sample 2. And similarly the mirror image would be the covariance of the sample 2 to sample 1.

So, similarly if you find out the covariance of the i'th and j'th, it would be exactly equal to the covariance of the j'th and the i'th from the sample counterpart and that these matrices for the case for the sample would be denoted by capital S which is bold and; obviously, it will be size p cross p.

So, with this I will close this 30th lecture and continue more discussion about the results or the or the definitions nomenclature for the multivariate case in a much better way such

that you appreciate as we proceed with the discussion of multivariate statistical analysis have a nice day and.

Thank you very much.