INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR

NPTEL
National Programme
on
Technology Enhanced Learning

Applied Multivariate Statistical Modeling

Prof. J. Maiti
Department of Industrial Engineering and Management
IIT Kharagpur

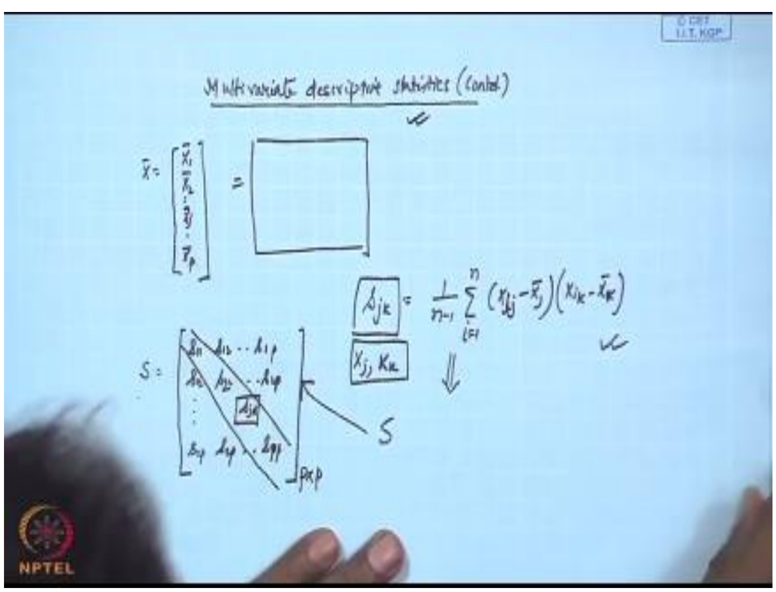
Lecture – 09

Topic

Multivariate Descriptive Statistics (Contd.)

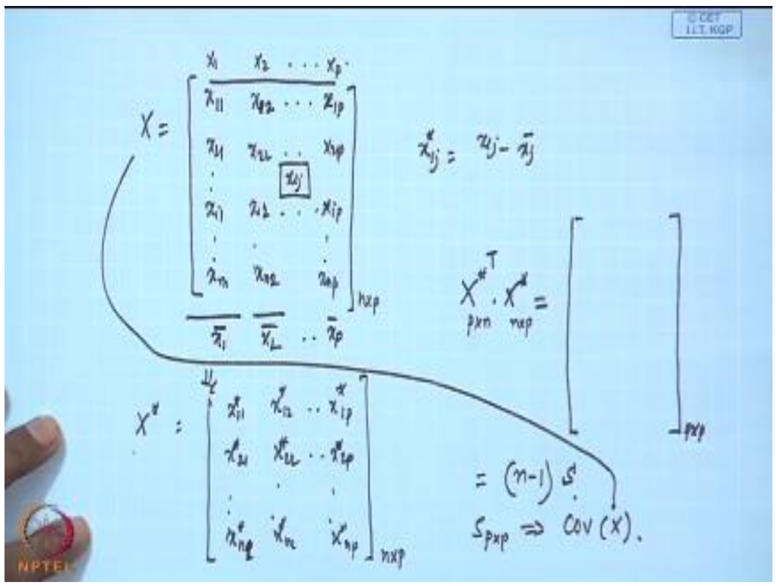
Good afternoon, we will continue multivariate descriptive statistics today.

(Refer Slide Time: 00:24)



Last class I have given you s_{jk} the formula to compute the covariance between two variables X_j and X_k, and this is the sample covariance computation formula. Now, we want to use matrix here, primarily matrix multiplication to compute S, where S is this one, that PxP matrix diagonal elements will be the variance component and up diagonal elements will be the covariance component. We will compute these from the data matrix.

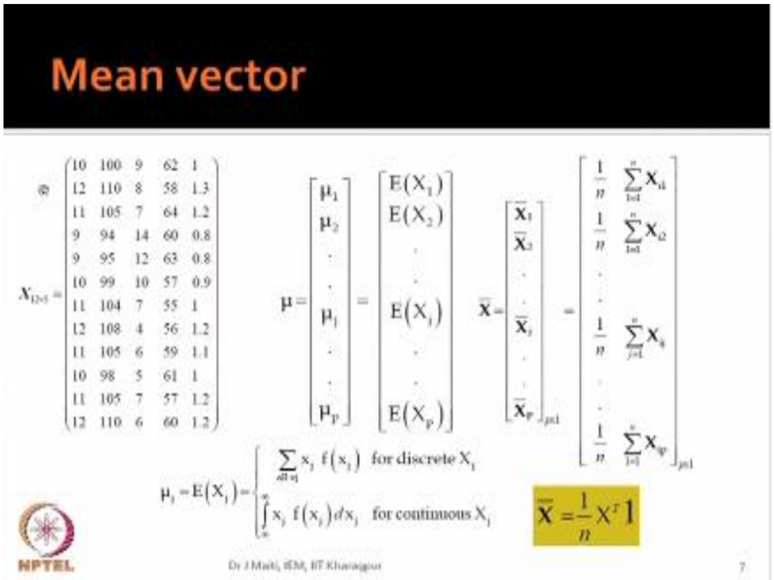
(Refer Slide Time: 01:20)



Thus data matrix last class you have seen the data matrix like this that x_{11}, x_{21} like x_{i1}, x_{n1}, that is observation on variable x_1. Similarly, observation of variable x_2 is x_{12}, x_{22} like x_{i2} then x_{n2}, so you take all p variables then x_{1p}, x_{2p}, x_{ip} and like this x_{np} this is our nxp matrix. Now, let us consider a general observation here, which is x_{ij}, so if we create one more general observation x^*_{ij}, whichis x_{ij}- $\bar{x}_{j,}$ then the resultant if, what you do, if you subtract each of the elements by its respective mean.

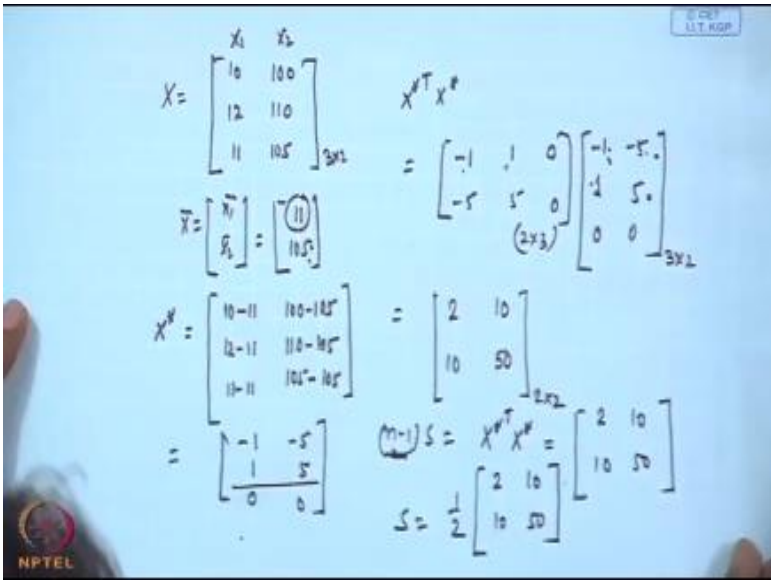
For example, for this one, if you subtract by $\bar{x}_1$, for these it is $\bar{x}_2$ like these it is $\bar{x}_p$, then we will create another matrix, which we are denoting like this x* this is your x^*_{11}, x^*_{21}, so like this x^*_{n1} then x^*_{12}, x^*_{22} like this x^*_{n2} so x^*_{1p}, x^*_{2p} like this x^*_{np}, okay. Now, if you create these x^{*T} what will be the order of this matrix pxn and take a dot product with x* this resultant matrix nxp, the resultant matrix will be pxp cross matrix, okay which is nothing but (n-1)S, where S is the covariance matrix. So, S is px p that is the covariance of x, where you have taken this calculated this x, okay so in matrix multi manipulation you are you are able to find the covariance matrix in just one go for all matrices, all the variables, okay.

(Refer Slide Time: 04:53)



Now, you solve one small problem here. For example, suppose this one you see that you take this one that first 10, 12, 11 100, 110 and 105, this data matrix.

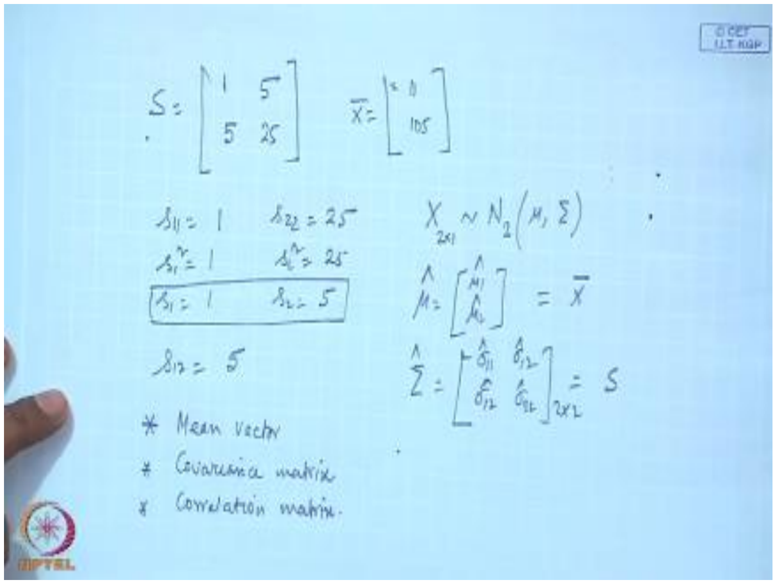
(Refer Slide Time: 05:20)



X=10, 12, 11, 100, 110 and 105. So, what will be your $\bar{x}_1$ or if I say $\bar{x}$= $\bar{x}_1$ and $\bar{x}_2$, what will be its value it will be 10+ 12, that is 33/3, 11 and this one will be 100, 110, 105. Now, what you are creating, you are creating x*, what is this x* you want that each element on x_1 will be subtracted by its mean 11. Similarly, each element of x_2 will subtracted by its mean 105, so then this one will be 10- 11, 12- 11,11- 11, second one will be 100-105, 110- 105, 105-105, what is happening here, you are getting this is10-11-1+1 0, then 100 minus that is -5 + 5 0.

So, sincerely that zero is element is coming here. Okay, let us see that what will happen if we do like this x^{*T} x*, what will happen here minus -1 1 0, -5 5 0, multiplied by a x*-1 1 0, -5 5 0. If you multiply what will happen, this is basically 2x3 matrix, this one is 3x2, so you want to get matrix called 2x2. So, this time this -1x-1, that is +1, +1+0 so this will be this 2. Now, second one -1x-5 that is +5, 1x5, +5 that is 10, so this one -5x5, 5x5+5, 10 and this one 25+25 that is 50, okay. So, we say (n-1)S=x^{*T}x*, so which is here 2, 10, 10, 50. So, what is n value here 3, n is 3 so n-1 is 2, so S will be 1/2 by 2 10, 10 50.

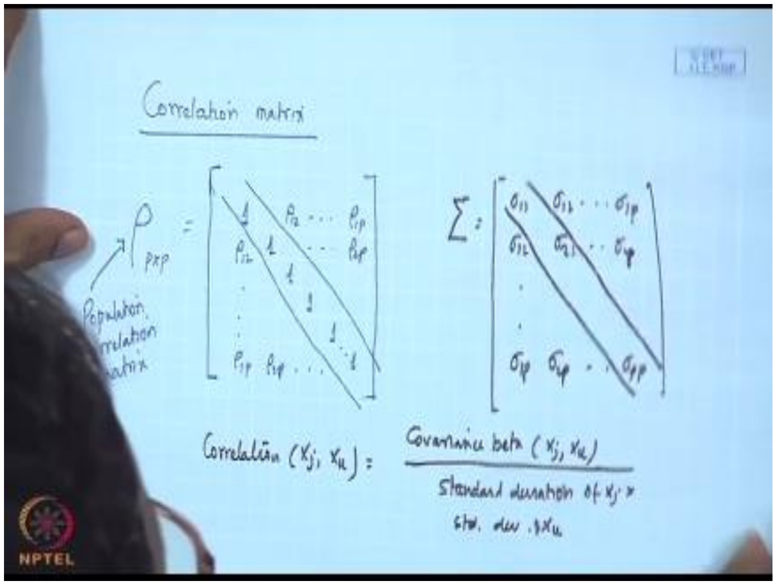
(Refer Slide Time: 08:57)



Then what is this value now, our S is that means 1, 5, 5, 25, so you have already computed $\bar{x}$ which is your 11, 105, and your S is this. So, that mean s_{11} is1, s_{22} is 25, what does it mean $s1^2$=1 and $s2^2$=25 s1=1, s2=5, that is a standard deviation both the cases and s_{12} which is 5 the covariance between x_1 and x_2, okay. So, if I say that my population, normal population multivariate normal that is 2x1 this is the variable, so it is basically N_2 μ and σ then my μ is μ_1 and μ_2 and σ will be 2x2 σ11, σ12, σ_{12}, σ_{22}, this is 2x 2.

So, we can now say, this is our $\bar{x}$ the estimate S in this manner we will precede. Okay, so, that please remember multivariate descriptive statistics has three components, one is mean vector, second one is covariance matrix, third one is correlation matrix. Now, we will discuss about correlation matrix.

(Refer Slide Time: 11:21)

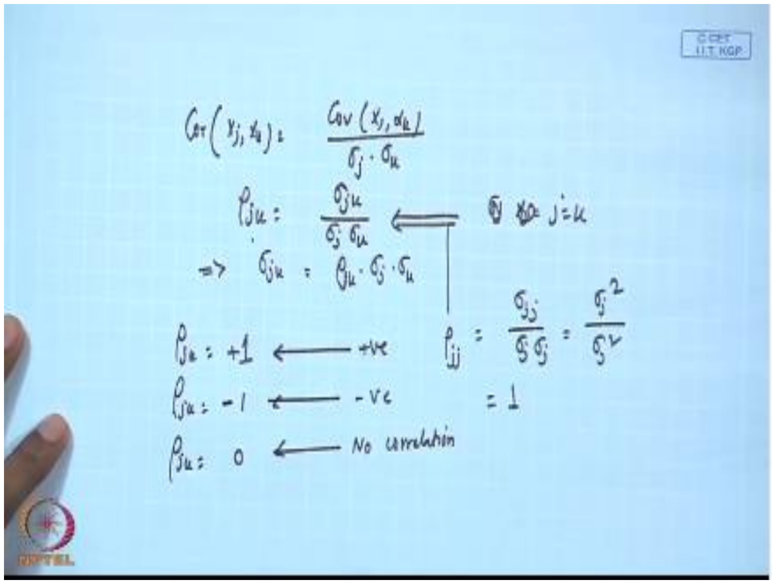


Now, population correlation matrix is denoted by ρ, this is population correlation matrix, correct. If there are p variable basically, what I mean to say that the population characterized by p variable, then will you get pxp matrix for the population correlation matrix, like pxp for some population covariance matrix. Duty here is that your diagonal element will be 1, this is the correlation of the same variable with it.

And up diagonal variable element will be writing like this ρ_{12}, like ρ_{1p} here also ρ_{12}, ρ_{2p} so like this ρ_{1p} ρ_{2p} it will continue like this, correct. Now, if this is the case can we find out a relationship between ρ and σ what is σ population covariance matrix that we have seen σ_{11},σ_{12},σ_{1p},σ_{12}, σ_{22}, σ_{2p} like this σ_{1p}, σ_{2p}, σ_{pp}. So, the cross of the matter is the diagonal elements are variance that is means, the same variable varying with it that is similarly here, diagonal element is the correlation, okay.

So, if you see what is the correlation between x_j and x_k, then you can write this one as covariance between (x_j, x_k) by standard deviation of x_j times standard deviation of x_k.

(Refer Slide Time: 14:22)

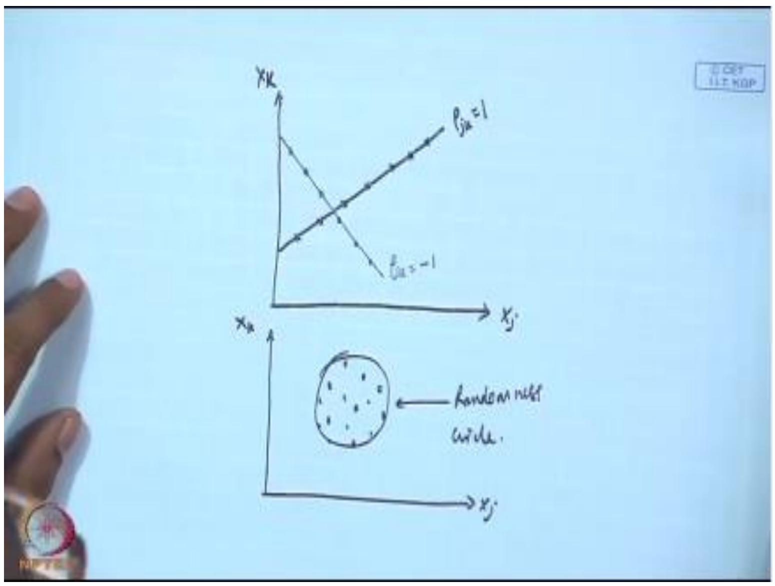


So, mathematical what we will write basically mathematically that correlation of (x_j, x_k) is cov of (x_j,x_k)/$\sigma_j.\sigma_k$. Now, you have seen from this figure, we are saying that correlation between j and k is somewhere like this ρ_{jk}, here also we are talking about that covariance between this is this like this. So, then if we use this same notation I can write $\rho_{jk}=\sigma_{jk}/\sigma_j.\sigma_k$, that is the relationship. What is the relationship then covariance between two variables is the correlation between the two variables times its standard deviations, correct. Now, what will happen to this correlation when σ_j, when σ_j=k what I mean to say j=k. You see what will happen to this that means it will be jj, j=jj, then σ_{jj}/σ_j and we have discussed earlier that σ_{jj} is nothing but σ_j^2.

So, that by σ_j^2 which is 1 and as a result you are getting all 1, okay, so conceptually that the same variable you are trying to find out the covariance between the same variable, then that will be give the variance and here what you are doing, you are basically standardizing it by dividing the standard deviation to the covariance component. So, as a result this standardization effect is bringing you that all the diagonal elements will be 1. Okay, now same thing will happen for sample data also.

Now what do mean by suppose correlation between j and k is +1, correlation between j k is -1, correlation between j k is 0. Okay, this one is saying that perfectly positively correlated and one stands for the perfect correlation, this minus stands for negative, that so that being that is positively correlated, this is negatively correlated, and this one is having no correlation positive, negative and no correlation.

(Refer Slide Time: 17:41)

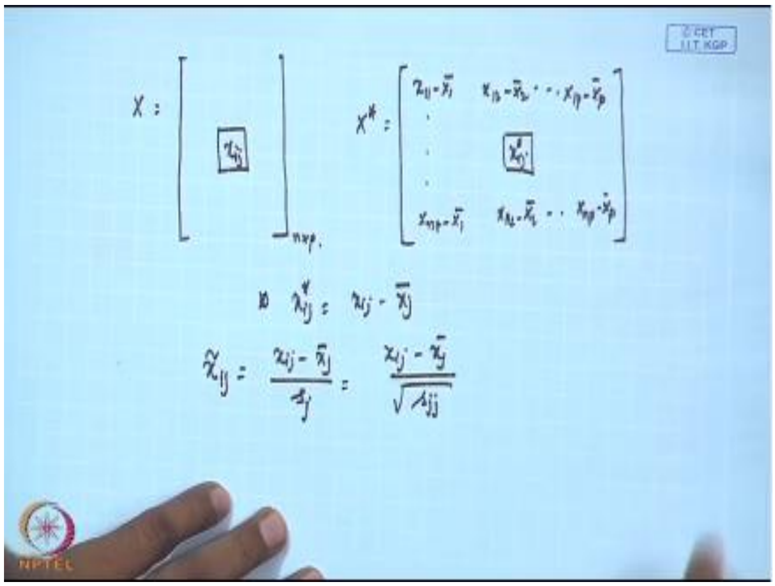


And if you draw this, suppose you have two variables this side suppose x_j and this side x_k and if you draw scatter plot for positive correlation, you may get like this that is $\sigma_{jk}=1$ okay, and negatively correlation means x_j will increase x_k will decrease or vice versa. So, you can think like this, here what is happening in this case this $\sigma_{jk}=-1$. Assuming that all the values of x_j and x_k falling on this line, that is why negative that mean here when x_j is increasing, x_k is decreasing.

And in the in this case when x_j is increasing, x_k also increasing, all are falling under the positive 1 is possible when it will increase both the variable co vary in the same direction, negative means in the opposite direction and1 is possible, when you will find a perfect straight line if you do curve fitting and when you will get 0, suppose your points are like this, this is your x_j, this y axis is x_k, points are like this you cannot find circle, this points are resembles circle its totally random.

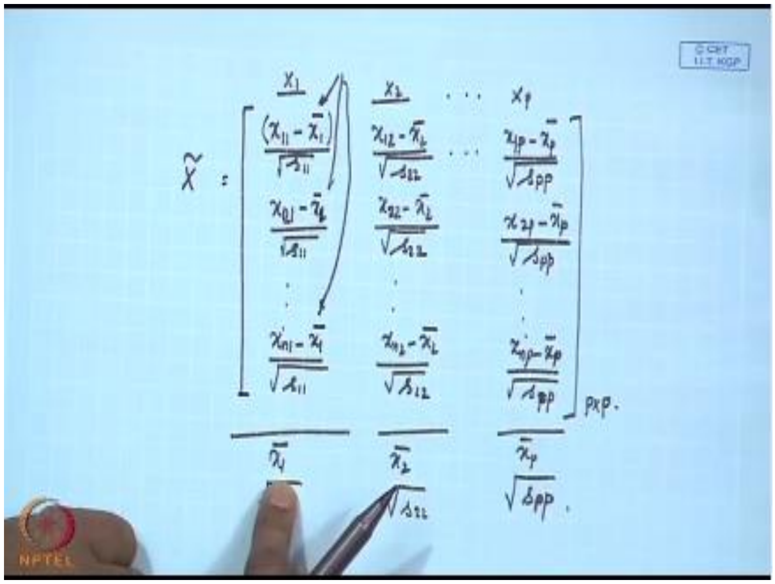
So when you find this type of randomness, that mean it is basically resembling a circle, there is no relation because you see you take any direction you will not get any pattern here, now that is the meaning of correlation coefficient like σ_{jk}.

(Refer Slide Time: 20:16)



You convert it into the same how do you calculate? You have data set like x, where same dataset nxp and I have already given you that we have converted this x*, where each of the observation was subtracted by its mean like this, this manner x_{n1}- $\bar{x}_1$, x_{n2}- $\bar{x}_2$, x_{np}- $\bar{x}_p$, this was we created earlier. And here we created one general x_{ij}, here also we have created that x_{ij}*, where x^*_{ij} is x_{ij}- $\bar{x}_j$. Now, let us create another variable, let us write like this $x^{\sim}{}_{ij}$, this one is x_{ij}- $\bar{x}_j$ /s_j what are you doing you are first finding out the mean subtracted values, and then dividing it by the corresponding standard deviation. I can write it like this, x_{ij}- $\bar{x}_j$ $\sqrt{s_{jj}}$, okay if you follow these.

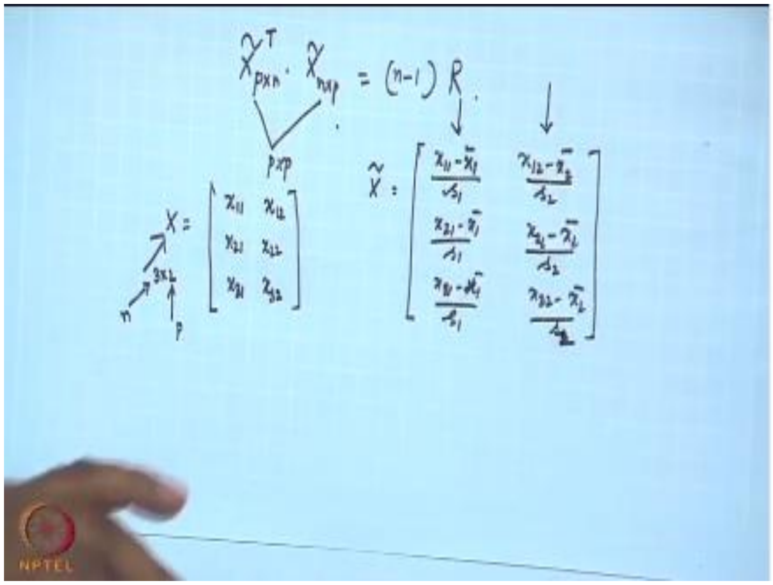
(Refer Slide Time: 22:16)



Then you will create a matrix which is x tilde, this one will be look like this x_{11}- $\bar{x}_1$ these divided by s_{11} square root then, x_{21}- $\bar{x}_1/\sqrt{s_{11}}$ same manner all the observation in the x variable is x_{n1}- $\bar{x}_1/\sqrt{s_{11}}$, for x_2 what you will do x_{12}- $\bar{x}_2/\sqrt{s_{22}}$, this is s_{22}- $\bar{x}_2/s_{22}$ like this, x_{n2}- $\bar{x}_2/s_{22}$. So, if you go in this manner for p^{th} variable then you will write x_{1p}- $\bar{x}_p/\sqrt{s_{pp}}$ then x_{2p}- $\bar{x}_p/\sqrt{s_{pp}}$ same manner x_{np}- $\bar{x}_p/\sqrt{s_{pp}}$ this is a transform at data matrix pxp, or when it is x_1, all will be 11, x_2 all will be 22, when it is p all will be x_{pp}.

See the similarity is all these are x_1, when each observation is subtracted by the same mean vector of the variable 1, and each that sub resultant quantity is divided by $\sqrt{s_{11}}$. Similarly, here $\bar{x}_2\sqrt{s_{22}}$ similarly here $\bar{x}_p\sqrt{s_{pp}}$, okay so you see that earlier ultimately, if you see the covariance and correlation relationship, you have found out that see jk, when you say jk, we will basically the covariance component is divided by the corresponding standard deviation. So, in order to achieve this, what we are doing here, we are now dividing each of the observation by the corresponding standard deviation.

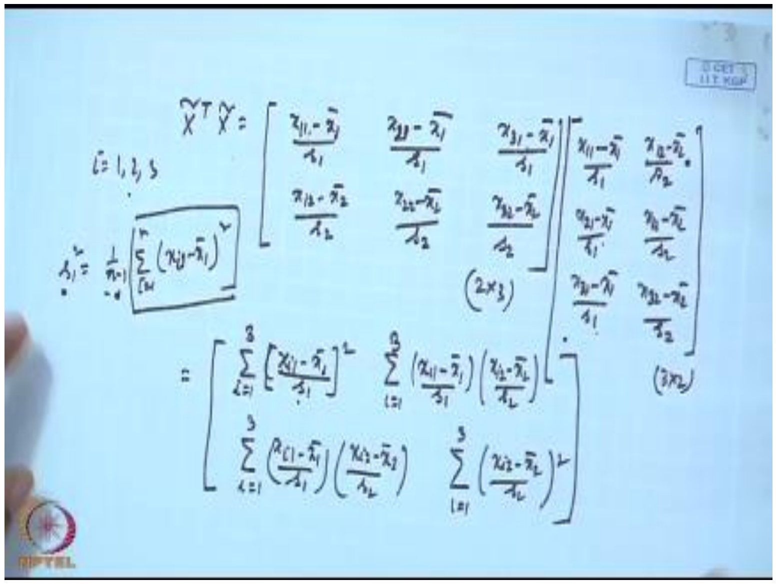
(Refer Slide Time: 25:46)



Now, if you find out this one with this so x^T then it will be pxn, if you give the transpose here x tilde transpose dot product x tilde, that will be your nxp. So, the resultant quantity will be your pxp getting me pxp then what will happen? No, this will not be identity matrix, you see you will get this one as (n-1)R, okay. If you want to check it, you check it very simply, suppose my data matrix is like this, I will take only three values x_{11}, x_{21}, x_{31} then here the second variable you take x_{12}, x_{22} and x_{32}.

So, in this case the data matrix is 3x2, where n is 3 and p is 2, okay then what you are creating here, you are creating x tilde. What is this, this is nothing but x_{11}- $\bar{x}_1$ by I am writing s_1 only, instead of square root of s_{11} that is s_1 I am writing. Then this will be what x_{21}- $\bar{x}_1$ by s_1 and x_{31}- $\bar{x}_1/s_1$ and this will be x_{12}- $\bar{x}_2/s_2$ then x_{22}- $\bar{x}_2/s_2$ then x_{32}- $\bar{x}_2/s_3$. What will happen if you now do like another, yes this is s_2, keep in mind the variable and accordingly write down this one, okay.

(Refer Slide Time: 28:03)

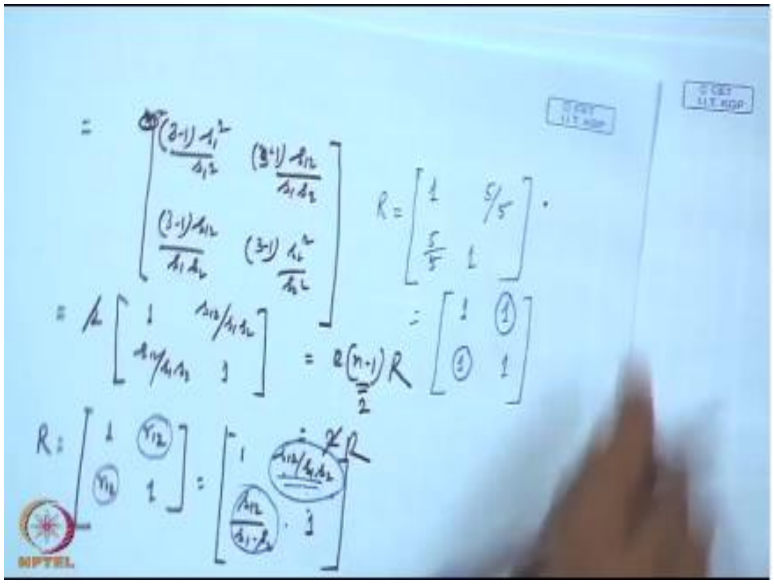


So, then if I do like this x tilde transpose x tilde, what will happen your x_{11}- $\bar{x}_1/s_1$, x_{21}- $\bar{x}_1/s_2$, x_{31}- $\bar{x}_1/s_1$ and then this one will be x_{12}- $\bar{x}_2/s_2$, x_{22}- $\bar{x}_2/s_2$, x_{32}- $\bar{x}_2/s_2$ this times you will be writing the same thing x_{11}- $\bar{x}_1/s_1$, x_{21}- $\bar{x}_1/s_1$, x_{31}- $\bar{x}_1/s_1$ the here x_{12}- $\bar{x}_2/s_2$, x_{22}- $\bar{x}_2/s_2$, x_{32}- $\bar{x}_2/s_2$ this one is 1 2 3. So, 2x3,this is 3x2 you see this into this plus this into this plus this into this, what is happening here x_{11}- $\bar{x}_1/s_1$, x_{11}- $\bar{x}_1/s_1$ you are getting a square. So, what you are doing then you are basically getting sum total if I write i=1to 3 because our observation is 1 2 3.

Now, second one stands for the variable so x_{i1}- $\bar{x}_1/s_1$ that square you are getting, okay then what will be this one, this cross this, now this cross this, what will we get x_{11}- $\bar{x}_1$, x_{12}- $\bar{x}_{2.}$ You see you what you are getting here, you are also getting i= 1 to 3 x_{i1}- $\bar{x}_1/s_1(x_{i2}$- $\bar{x}_2/s_2)$ same quantity will be getting here i=1 to 3, x_{i1}- $\bar{x}_1/s_1$ and xi_2- $\bar{x}_2/s_2$ here you will be getting a square term i=1 to 3 xi_2- $\bar{x}_2/s_2$ getting any similarity here, any clue are you getting.

You see we say what a sum total is of if I ask you what s_1^2 suppose n=3, then what is s_1^2, 1/n-1 sum total of i=1 to n, x_1 or x_{i1} you write - $\bar{x}_1^2$ s^1 you see what is happen x_{i1}- $\bar{x}_1^2$ 1/n-1 so, this quantity is what (n- 1)s_1^2 so and yes, because s_1/s_1 is coming.

(Refer Slide Time: 32:04)



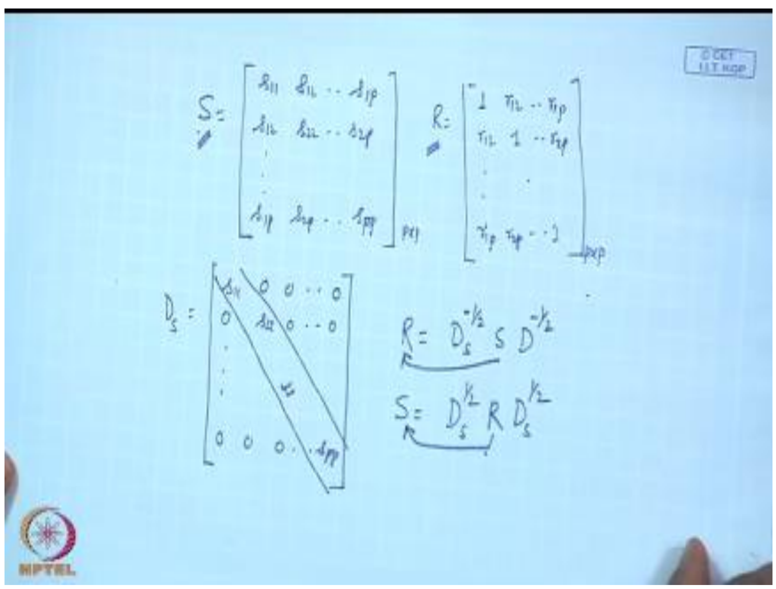
So, ultimately what will happen ultimately, you will get this quantity like this that we can write n-1 means here 3-1, $(3\text{-}1)s_1^2/ s_1^2$ then second one what will happen, this is the covariance yes or no, n-1 you will get s_{12}/s_1s_2 that is n is 3 here. So, 3-1 s_{12}/s_1s_2 and then this one also you get $3\text{-}1s_2^2/s_2^2$ Then if I just take out 3-1 which is basically 2 then what I will get, I will get 1 $s_{12}/s_1s_{2,}$ s_{12}/s_1s_2 and 1 and as a result what we writing this is (n-1)R, (n-1) is 2 that mean this is 2R. Now, 2 is 2 cancelled out so R is now 1, r12, r12, 1, which is now 1 s_{12}/s_1s_2 then s_{12}/s_1s_2 into 1, correct.

And you, we have seen earlier, I think we have earlier that the correlation and covariance, the relationship you were seen earlier the relationship what in population domain we say this one I have said to you, this as well as this. Now, instead of ρ in the sample domain, what you can write, you can write $s_{jk}=\sigma_{jk}/\sigma_j\sigma_k$. Now, if your j=1and k=2 and this $s_{12}=\sigma_{12}/\sigma_1\sigma_2$, what is exactly happened here. This is not s_{jk}, this is basically ρ_{jk}, then this will be your r_{jk}, ρ_{jk} is this is basically replaced by σ and then it will be like this. So, if I can write ρ_{jk} here, rh ρ_{12} like this one.

Now, in the sample domain when we go that will be your r_{12} will be your s_{12}/s_1s_2 that is why these r r_{12} is s_{12} covariance by their corresponding standard deviation, okay. So, for the same data set now can you compute this your r value? I think we have computed one place r value, you have computed, we have computed here. We have seen S equal to this, okay so, that means your standard deviation s_1 is given and s_2 is given for the same data. If I want to compute my r, then what will be your r value? Your r value will be first that blindly you can write like this, diagonally it will be 1 and r_{12} will be s_{12}, so s_{12} is 5.

So, you can write 5 divided by their standard deviation s_1 is 1, s_2 is 5 so that mean 5/5, 5/ 5, 1 1 1 1 1, now you are getting perfect correlation, 1 means perfect correlation. So, this is what your multivariate descriptive statistic, we will talk about that is why the mean vector correlation matrix and covariance matrix. Now, you can very easily convert this one.

(Refer Slide Time: 37:22)

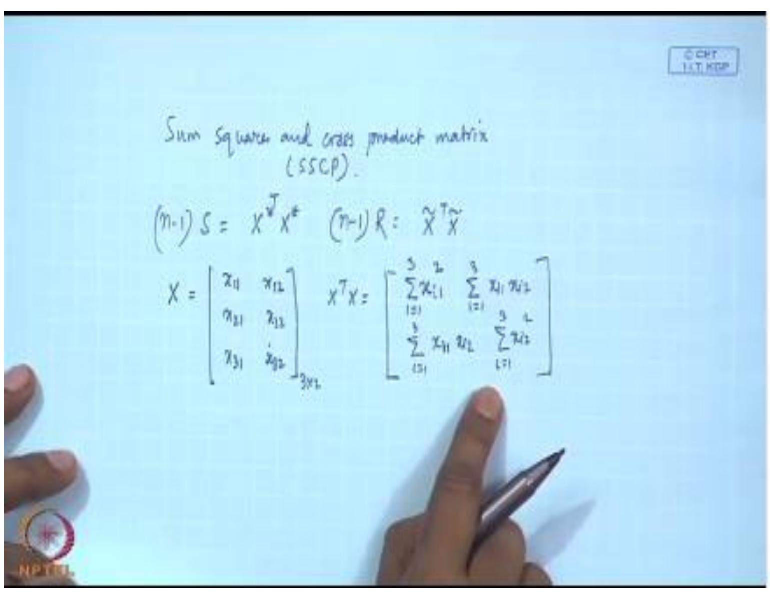


Suppose my covariance matrix is this, s_{11}, s_{12}, s_{1p}, s_{12}, s_{22}, s_{2p}, like s_{1p}, s_{2p}, s_{pp}. My correlation matrix is this 1, r_{12}, r_{1p}, r_{12} again 1 r_{2p} so like this r_{1p}, r_{2p}, 1. So, you create another diagonal matrix D yes, this one is same pxp matrix, this is pxp this also pxp, same pxp matrix only the diagonal elements will be the variance component of diagonal will be 0. So, this is s_{11} 0,0,0,0,s_{22} like0 0 0 0 s_{pp}. So only diagonal elements are the diagonal elements of Ds are the diagonal elements of the covariance matrix, up diagonal elements are 0.

It will create like this and suppose you know s, you can just you with one trick you can find out that R is $Ds^{-1/2}$ s $D^{-1/2}$ both case -1/2 $Ds^{-1/2}$. If you use must left, if use this mat lab now straight way, we will calculate all those things from the data, but suppose you want do the conversion, you mean in excel you can do this, getting me. What you are doing, you most of the time you may be knowing this one that variance component, once you know s you know the variance, covariance also, you want to calculate R.

Suppose, you know this variance component and correlation is known, correlation matrix is known, you want to go to co variance matrix from correlation matrix. What you have to do, you have to write like this plus R. So, this one this is basically from covariance to correlation and here correlation to covariance. Only thing you want to require in the second case, the variance component of all the variables considered, okay. Another important concept in multivariate data analysis is sum square and cross product matrix.

(Refer Slide Time: 40:33)

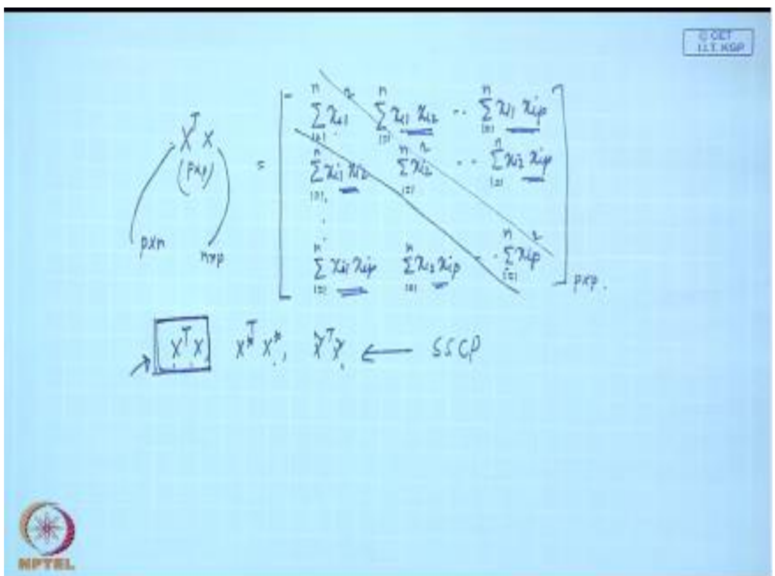


Which is known as SSCP sum square and cross product, okay. So, if you see, when you calculate the covariance matrix, correlation matrix, we are using the formula that one for $(n-1)S=x^{*T}$ x* we have used. We have used (n-1)R equal to X tilde transpose, x tilde we have used where, both

x^* and x tilde are basically transformed matrix from the original data which is x. So, suppose x I am writing like this, x_{11}, x_{21}, x_{31}, x_{12}, x_{22}, x_{32}, you do like this. Now, you calculate x^T x, what will happen here, you will be getting like this x_{i1}^2 i=1 to 3 here the i= 1 to 3 x_{i1}, x_{i2}, here i= 1 to 3 x_{i1}, x_{i2} here x_{i2}^2 i= 1 to 3.

I have taken 3x2, if you multiply we will be getting because earlier, we have seen subtracted by mean and for divided by standard deviation case, we have seen we got similar formula. Now, the same thing if you think from the pxp variable point of view.

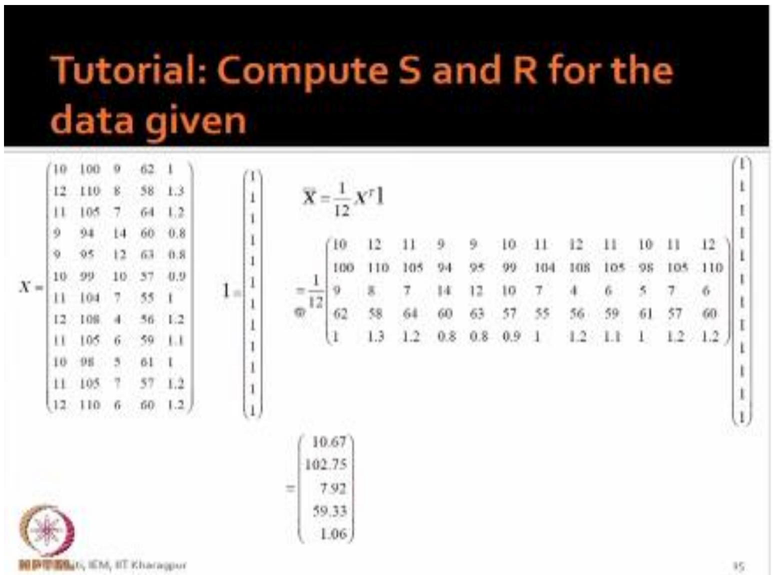
(Refer Slide Time: 42:44)



Then x^Tx will be a pxp because x^T this one is pxn and this one is nxp, you will be getting like this. So, your matrix will be like, this sum total x_{i1}^2 then x_{i1} x_{i2}, then your x_{i1} x_{ip}, here x_{i1} x_{i2} here x_{i2}^2, then x_{i2} x_{ip} in similarity, I am writing all those things x_{i1}' x_{ip} x_{i2} x_{ip} then x_{ip}^2 okay. So, it is a pxp matrix, correct then i equal to, I definitely equal to 1 to n all cases 1 to n. This 3 matrices like x^Tx $x^{*T}x^*x\sim^T$ x~ these are all sum square and cross product matrix, all SSCP why? You see now, these are the some square, all diagonal elements are sum square and up diagonal you see cross product all up diagonal are cross product.

So, sum squares for the variance cross product from the covariance that mean from this matrix also. Once you know these we can use these or these matrixes were ultimately, we can calculate the descriptive statistics like co variance and correlation matrix. This one is very, very important matrix, later on particularly in regression; you will be using this matrix.

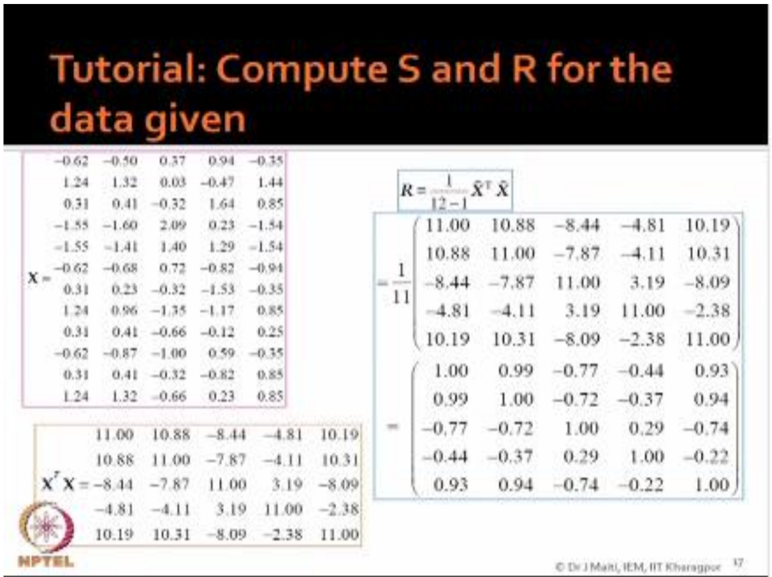
(Refer Slide Time: 45:17)



Now, let us see that how to calculate, suppose this is the problem given, okay that compute S and R for the data given, this data said you have seen earlier so I have used excel only. Using excel I have created so this is my data matrix, I want to compute the mean then as there are n data points so I created one unit vector with n data points. So, my aim is 12 here. So, it is 12x1 vector then $\bar{x}$ is 1/12 x^T1 when I multiply all those things, I got this values. So, profit mean is 10.67, then sense volume 102.75, 7.92 is your absent sing, then break down 59.33 and 1.06 is the M ratio case.

So, your first step will be this, find out $\bar{x}$ and you use this type of formulation then what you require to calculate, you require calculating s. You require converting this x value to x* that means each of the suppose10-10.67 that is why minus 0.67 coming here x*.

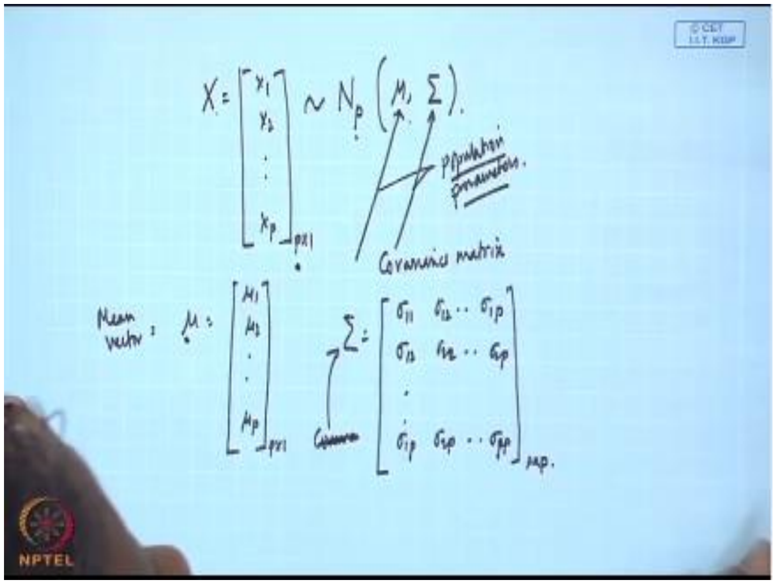
(Refer Slide Time: 47:01)



Once you get this, this is the formula S is 1/n-1 $x^{*T}x^*$ this will give you this value, okay if you and the left hand side the bottom portion, this is nothing but SSCP matrix $x^{*T}x^*$. You can do the same thing now we are interested to know that tilde. Here this is basically x tilde, this one x tilde transpose x tilde, this transpose part is missing x tilde and R is 1 /11 into X tilde transpose X tilde, you are getting like this. You see once you go by this way calculating, we will get all the diagonal elements 1, if you do not get that, there is a problem.

You have any questions so far, okay. Now we will go to although the next class I will be explaining in detail, when we will start that multivariate normal distribution. What we have assumed here?

(Refer Slide Time: 48:30)

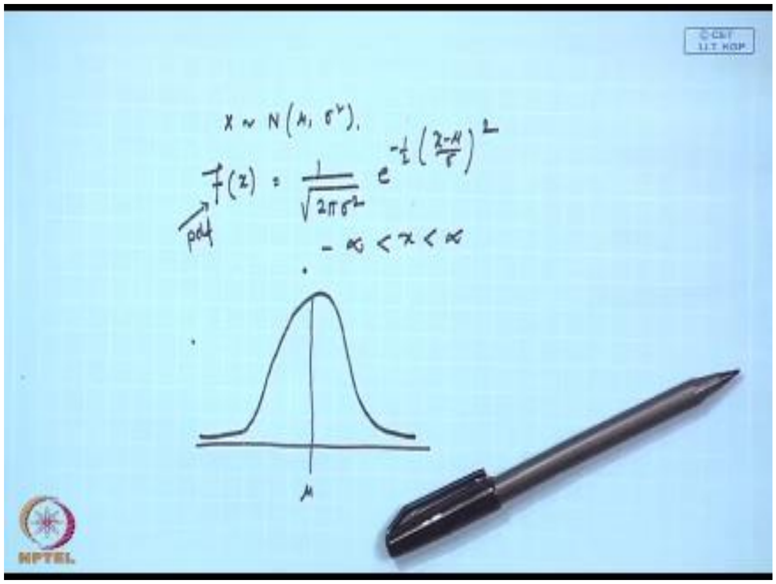


We have assumed here is x that is a variable vector x_1 x_2 dot dot x_p, that is px1. And we assume that this follows normal distribution that is multivariate normal, which will be denoted by like this np and μ and σ and, okay now, you are well accustomed with the nomenclature. Nomenclature in the sense you know that mean is mean vector, if there is px1 variable, then my mean is again px1 that is μ_1 μ_2 μ_p and you also know this one, this is nothing but covariance matrix so this covariance matrix is our pxp matrix σ_{11},σ_{12},σ_{1p} σ_{12},σ_{22},σ_{2p}, σ_{1p},σ_{2p},σ_{pp}.

Multivariate normal distribution is characterized by p variable with parameter that is μ that is a mean vector and covariance matrix, correct. So, please remember these are population parameters, these two are population parameters. So, far we have not discussed about that whether the data is coming from multivariate normal or not but ultimately, we will be going to multivariate normal distribution because most of the models that will be relied on these assumption multivariate normality, okay.

Now, when p=1 that is univariate normal, and now what is the probability density function of univariate normal.

(Refer Slide Time: 50:52)



Suppose x is a random variable, which is univariate normal with μ and σ^2 then if I want to know, what is your probability density function of x, this is your f (x)pdf, you will write $1/\sqrt{2\pi\sigma^2}\, e^{-1/2(x-\mu/\sigma)2}$ where $-\infty<x<+\infty$, correct. So, this is what you have seen earlier that this is what our normal distribution is. So, what will be the equivalent distribution when number of variable is more than 1 that will be our starting point in the next class. Okay, then thank you.

(Refer Slide Time: 52:00)

NPTEL Video Recording Team

NPTEL Web Editing Team

Technical Superintendents

Computer Technicians

A IIT Kharagpur Production

www.nptel.iitm.ac.in