

**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**

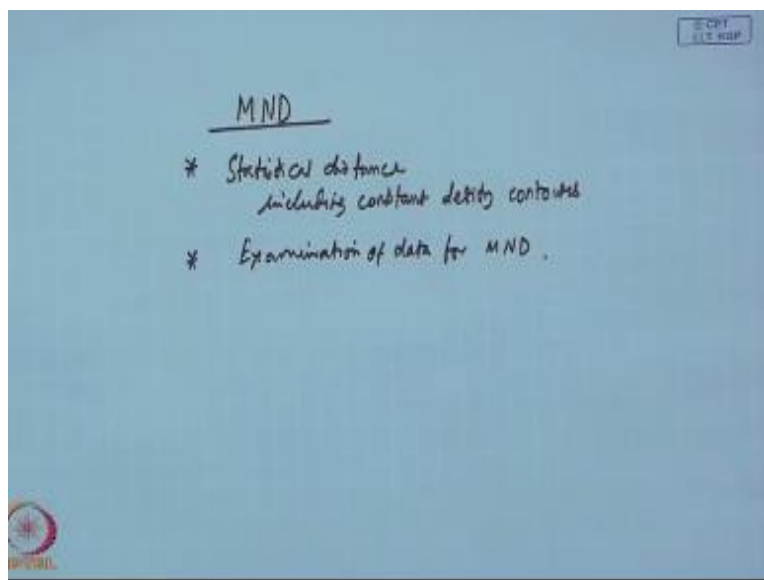
Lecture – 11

Topic

**Multivariate Normal Distribution
(Contd.)**

We will continue multivariate normal distribution.

(Refer Slide Time: 00:22)



Today our discussion will be on two issues; one is statistical distance, distance including constant density contours, and then we will see that how to determine that your data is multivariate normality. So examination of data for multivariate normality okay.

(Refer Slide Time: 01:15)

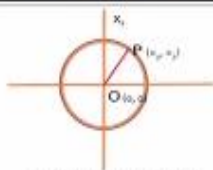
Properties of MND

- (i) If $X_{p \times 1} \sim N_p(\mu, \Sigma)$, then X_j is $N(\mu_j, \sigma_j^2)$ for all $X_j, j = 1, 2, \dots, p$.
- (ii) If $X_{p \times 1} \sim N_p(\mu, \Sigma)$, then the subset of $X_{p \times 1}$, i.e., $X_{q \times 1}$ is $N_q(\mu, \Sigma)$.
- (iii) If $X_{p \times 1} \sim N_p(\mu, \Sigma)$, then the linear combination of $X_j, j = 1, 2, \dots, p$, is univariate normal.
- (iv) If $X_{p \times 1} \sim N_p(\mu, \Sigma)$, then the q linear combination of $X_j, j = 1, 2, \dots, p$, is multivariate (q -dimension) normal.

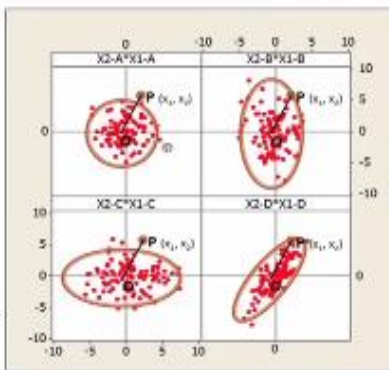


(Refer Slide Time: 01:15)

Statistical Distance



Euclidean distance

$$d(OP) = \sqrt{x_1^2 + x_2^2}$$
$$d(OP) = \sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$$
$$d(PQ) = \sqrt{(x_1 - q_1)^2 + (x_2 - q_2)^2 + \dots + (x_p - q_p)^2}$$


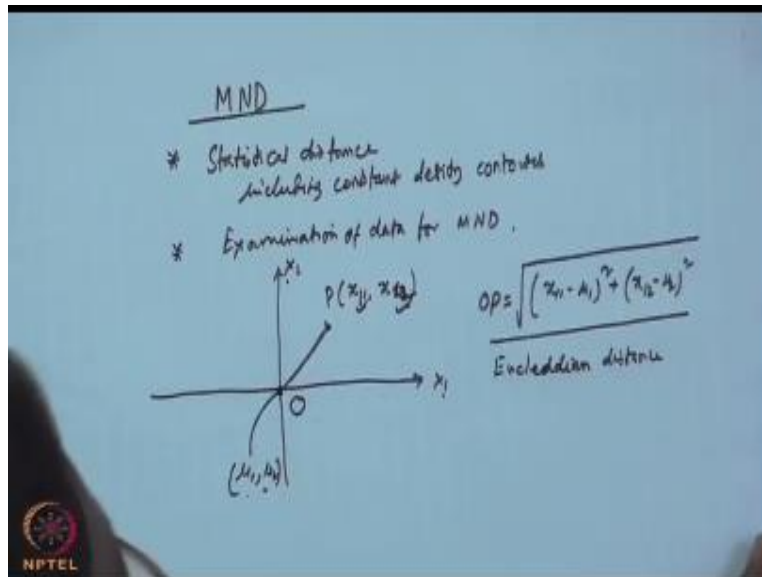
NPTEL

Dr. J. Mani, IIT Kharagpur

35

So let us see a statistical distance first.

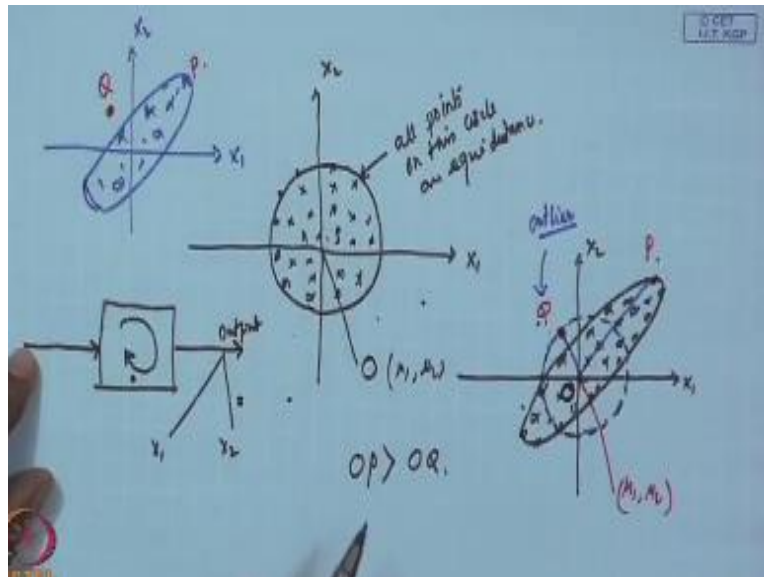
(Refer Slide Time: 01:20)



Suppose you consider two variable case, X_1 and X_2 , now this is let it be this is the origin with μ_1 and μ_2 , in that sense. So let us give this point name as O and you want to know the distance from O to P , P is another point what is X_{11} from 1 and your X_{21} . So, that when this point is from variable point of view X_{11} and X_{21} or other way you can write that X_{12} you can write that manner also.

So then the distance between these two points that is OP that every all of us know that this is $X_{11} - \mu_1^2 + X_{12} - \mu_2^2$ and this square root. This is a point with coordinate values for X_1 , this is X_{11} for X_2 , this is X_{12} and this is the reference point O with μ_1 and μ_2 as coordinate then the distance is this distance is known as Euclidean distance.

(Refer Slide Time: 03:14)



Now, let us see that we have several points scattered on a bi-variate plot X_1 and X_2 in such a manner that it basically resembles like this, the figure is like this or you can increase this. Now, if you find out all points equidistant from this O , that is μ_1 and μ_2 coordinate, all points equidistance from O that mean the equal Euclidean distance, and if you join them you will get a circle.

So, all points on this circle are equidistance okay. Now, we will go to the process level for example, I told you in one of the classes that we have a manufacturing process and that process take certain inputs and gives certain outputs. This output is measured in terms of its quality, let there are two variables, you are measuring here X_1 and X_2 related to the output, it may be related to the process also okay.

For the time being let us consider the output, then if you make a plot of X_1 and X_2 you may get a figure like this. Now, with little assumption in the sense that as if the figure resembles an ellipse okay, then I am keeping another observation on X_1 and X_2 which is here closer to this mean point, let it be here. So, as we have assumed earlier also this is our μ_1 and μ_2 this is one point, let this point is Q , let another point somewhere here which is P , correct?

Now, if we go by Euclidean distance OP the distance between O , the origin and P it will be greater than OQ from Euclidean distance point of view. Euclidean distance what we say, that all points which are equal distance that will make an ellipse, and this suppose I want to get Euclidean distance with respect to Q , OQ that is the distance then you will be getting something like this the circle when you make, this circle see definitely OQ , this distance is less than OP , correct?

Suppose, you do not know the distance concept, we are not interested in terms of defining the scattered observation from distance point of view, getting me? Then which one is closer to O , P or Q . So that if you see the scatter plot which one is closer to O , that μP , this P is closer to O , the reason is because you are the general mass, the behavior of the general mass is like this, like an ellipse and point P belong to this general mass whereas, point Q is does not belong to the general mass.

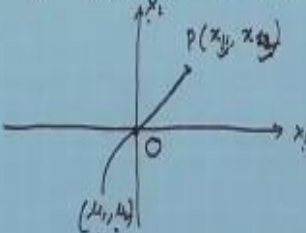
So, from the process point of view this is outlier, Q is outlier. Whereas, P belongs to the general mass because our figure is like this, it is like this, your point is here and all values are like this, as it your Q point is here, see Q and P . It simply indicates that if we go by Euclidean distance measure we cannot capture this behavior. What is the problem? Here the problem lies, see the variability across X_1 and X_2 is not captured.

(Refer Slide Time: 09:21)

© CPT
111 RUP

MND


- * Statistical distance including constant density contours
- * Examination of data for MND.



$P(x_{1j}, x_{2j})$

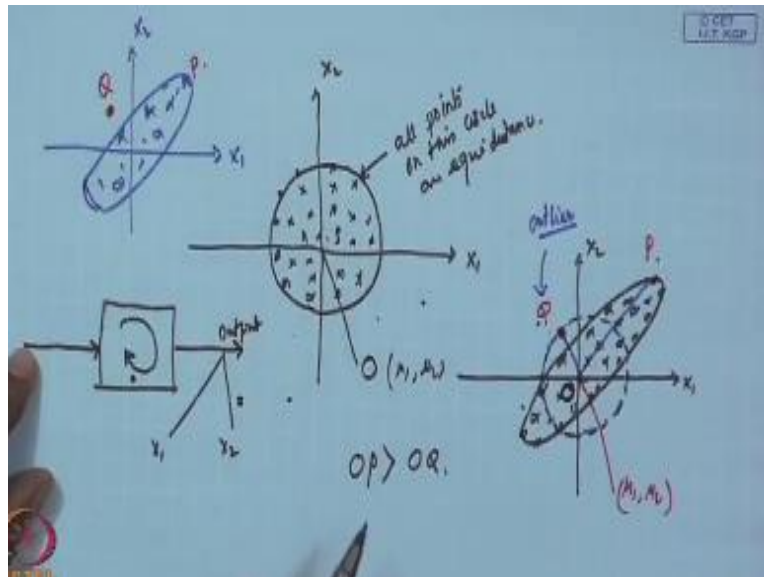
$$OP = \sqrt{\underbrace{(x_{1j} - A_1)^2} + \underbrace{(x_{2j} - A_2)^2}}$$

Euclidean distance



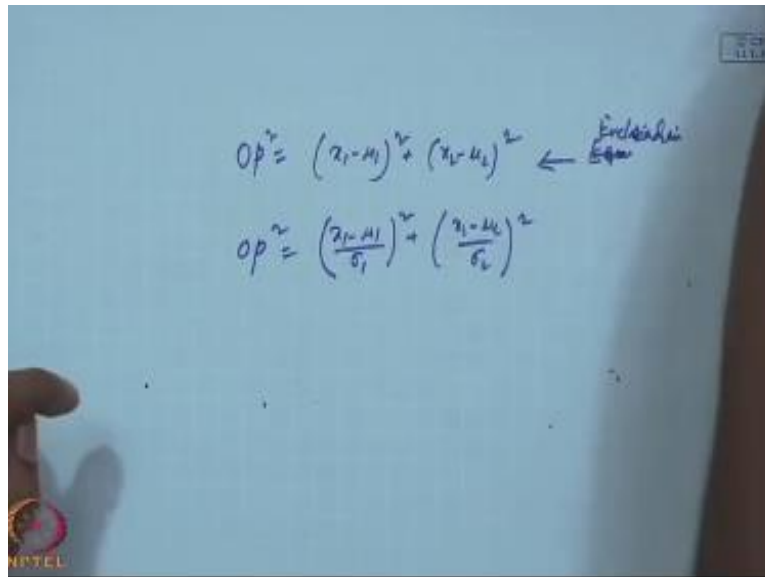
In this equation.

(Refer Slide Time: 09:27)



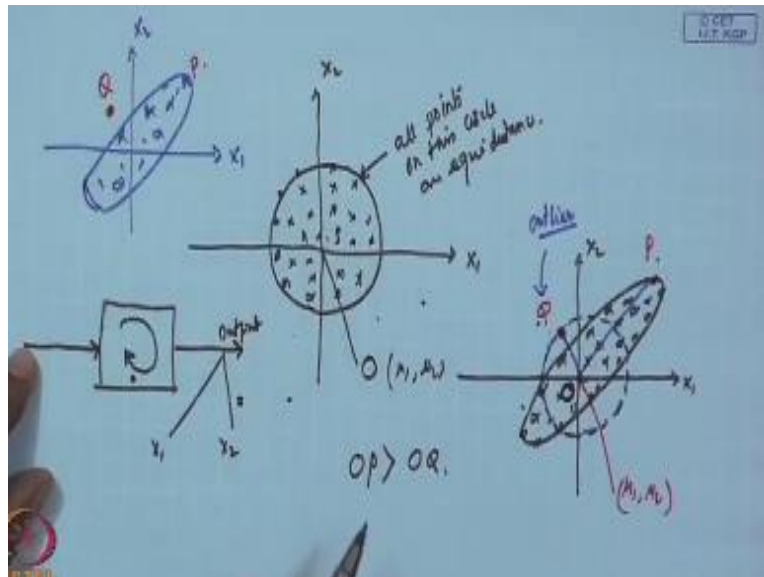
And it is because of the variability along the two dimension the difference in variability you are getting this structure. So, we want to include variability into the equation. Now, let us see if we give weightage to each observation by its variability. What will happen?

(Refer Slide Time: 09:54)


$$OP^2 = (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 \leftarrow \text{Euclidean Eqn}$$
$$OP^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2$$

Suppose, we will now first write this OP^2 , in the general sense $X_1 - \mu^2 + X_2 - \mu^2$ X_1 and X_2 variables this is your Euclidean distance. Instead of this if I give weightage to the observations by its variability I can write like this, $X_1 - \mu_1 / \sigma_1^2 + X_2 - \mu_2 / \sigma_2^2$ that means the mean subtracted observation is weighted by 1 by σ_1 as well as 1 by σ_2 depending on the variable. So, this one if I write this is OP^2 , is it not an equation of ellipse, definitely.

(Refer Slide Time: 11:03)



But not this ellipse that is equation of ellipse. But not this ellipse, why here what happened?
There are correlations.

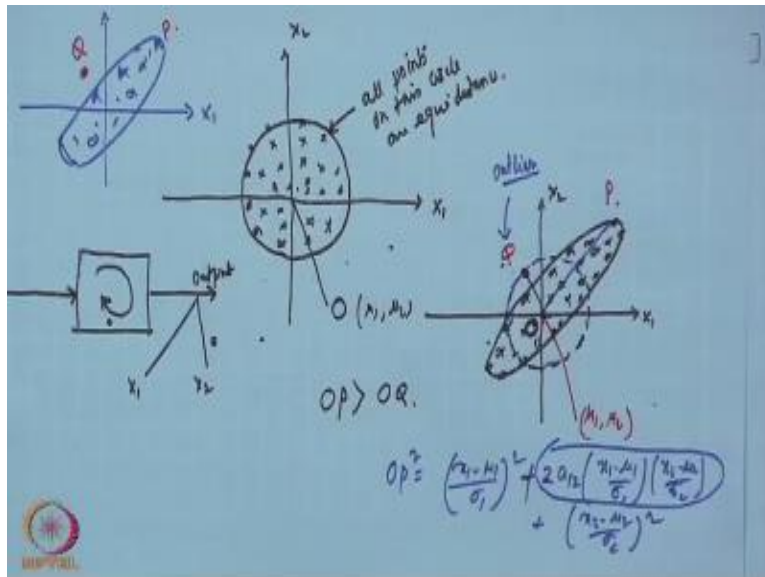
(Refer Slide Time: 11:12)

The image shows a handwritten derivation of the ellipse equation. At the top, the distance squared from the origin O to a point $P(x_1, x_2)$ is given as $OP^2 = (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2$, with a note "Euclidean Equ". Below this, the equation is rearranged into the standard form of an ellipse: $OP^2 = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2$. A diagram below shows an ellipse centered at $O(\mu_1, \mu_2)$ on a coordinate system with axes x_1 and x_2 . A point $P(x_1, x_2)$ is marked on the ellipse. A note at the bottom states: "All points on this ellipse are equidistant from O ."

So if we change, so this is an equation of ellipse and it represent two variables which are independent like this. So, if I say across X_1 variability high X_2 then this one is our equation of ellipse, and if you take a point here, suppose this is your P , this one is O , O is μ_1 and μ_2 and P if say X_1 and X_2 then this OP , this distance is this one, this is the formula for this distance. And as this one is the equation of ellipse all points on this ellipse, the distance will be computed using this function.

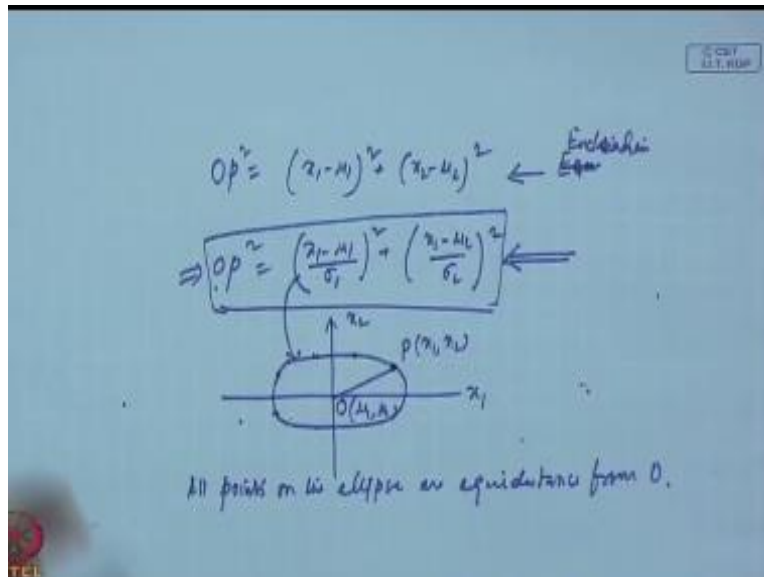
And what we say here? We say all points on the ellipse are equidistance, we say that all points on the ellipse are equidistance definitely from O , from a reference point.

(Refer Slide Time: 12:40)



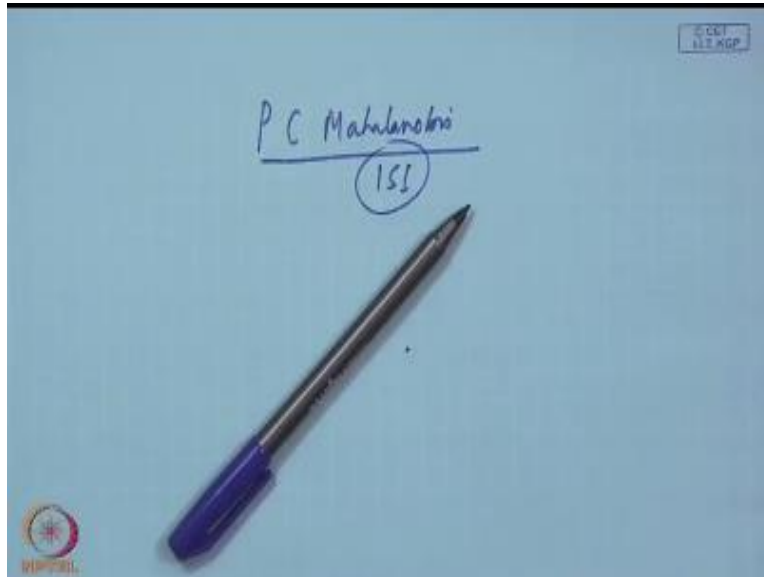
If you take this ellipse what will happen? Your OP^2 , you can write $X_1 - \mu_1$ or $+$ you write whatever may be there, you write plus some constant will be coming here, let it be a 1 2, I am giving $(X_1 - \mu_1 / \sigma_1) (X_2 - \mu_2 / \sigma_2) + (X_2 - \mu_2 / \sigma_2)^2$ this is the covariance part between the two variables okay.

(Refer Slide Time: 13:23)



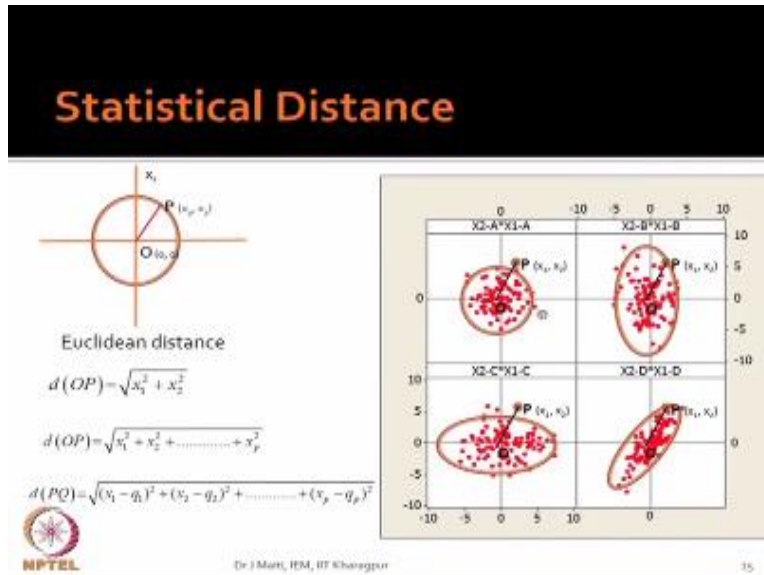
This is what is statistical distance okay, this is what either this elliptical distance, this is what is statistical distance. So in essence that the statistical distance, one of the pioneers in developing.

(Refer Slide Time: 13:46)



This is PC Mahalanobis okay he is an Indian scientist, he is founder of ISI, Indian Statistical Institute okay. Now, look at the slide.

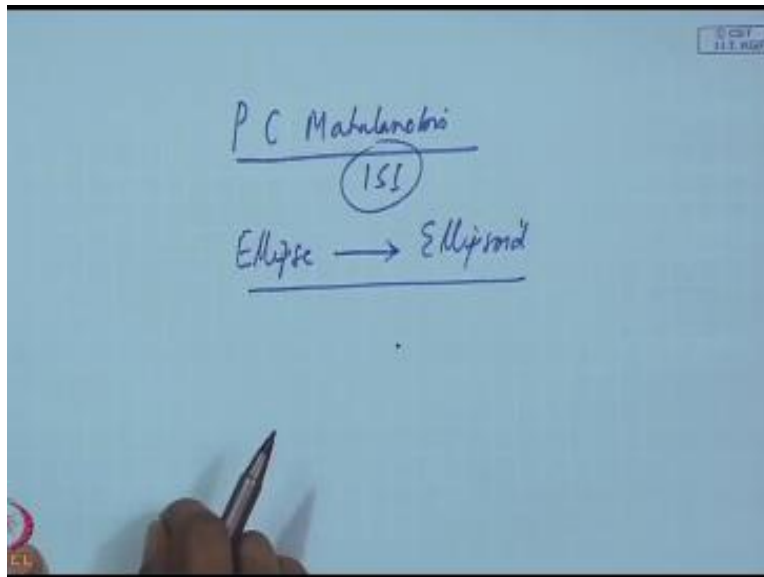
(Refer Slide Time: 14:08)



In this slide there are, in the right hand side there are four figures, first that left most one, top one, that one resembles circle, second one, it basically depicts one ellipse, third one is also ellipse, fourth one is also ellipse, but the second and third one, the variability along X2 is more in case of second one. And in case of third one the variability along X1 is more and in case of fourth one, it is there are change in variability as well as there is correlation between the two variables okay.

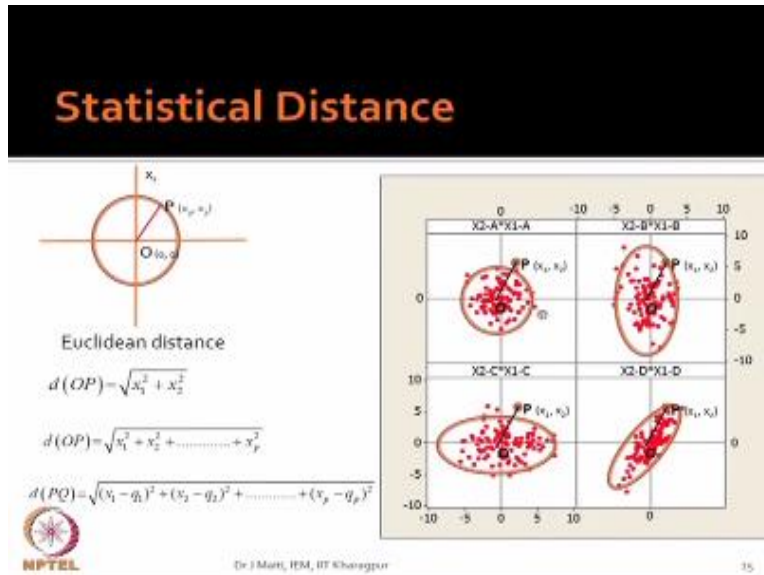
So, first one is random in the sense we are saying that the two variables are independent, second also two variables are independent, third one also two variables are independent, only in fourth one variables are dependent. Now, you think that this is the bi-variate case, you think from multivariate point of view. So, when number of variable will be more than 2 then it will be difficult to pictorially visualize, but the points are there.

(Refer Slide Time: 15:35)



So, ellipse will become ellipsoid okay.

(Refer Slide Time: 15:52)



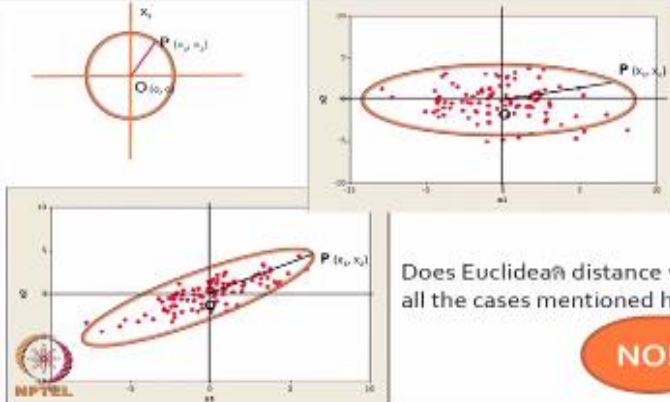
So, can you tell me that what is the difference between this figure, this or this figure. So, if I say this is my quadrant one, quadrant two, quadrant three, quadrant four. Quadrant one versus quadrant three one and quadrant three, quadrant three this is quadrant one this is quadrant three yes along Y axis or X1 and X2 we have given X1 and X2 uncorrelated okay. First quadrant, second quadrant, third quadrant variables are not correlated.

Now, what can you talk about the variability of X1 and X2 in the second quadrant, the circle is there, variables are uncorrelated and variance same along X1 and X2 variance is same. So, if X1, X2 variance is same then it will ultimately resemble a circle. If there is difference in one of the axis the length will be more that will be the major axis, other one will be the minor axis. If there is correlation ultimately the total ellipse will be shifted to the other direction okay.

This is very important concept, later on in principal component analysis we will again bring back to this figure there we will see that if they are highly correlated. What is the need of measuring so many variables? Can you not measure a smaller number of variables that we will discuss okay?

(Refer Slide Time: 17:52)

Statistical Distance



Does Euclidean distance work in all the cases mentioned here?

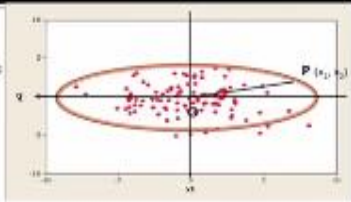
NO!!

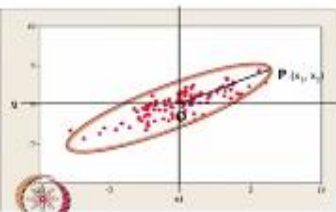
35

(Refer Slide Time: 17:54)

Statistical Distance

$$\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 = (OP)^2 = d^2$$





$$(OP)^2 = d^2 = \frac{1}{1 - \rho_{12}^2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho_{12} \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right]$$

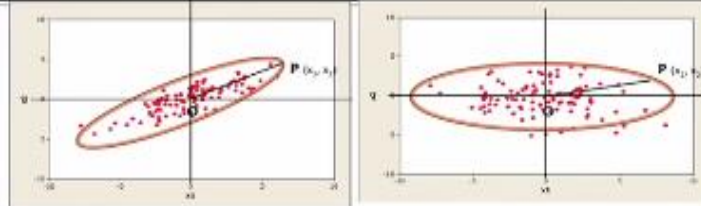
$$(OP)^2 = d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \dots (7)$$

NPTEL
Dr. J. Mani, IEM, ST. Xavier's College

I think if you look into this figure, this figure you have seen earlier also and this equation also you have seen earlier. What is the exponent part of a multivariate normal bi-variate normal density function? This is the exponent part and we have also discussed that exponent will resemble an ellipse so where is the distance? This is what is the statistical distance okay.

(Refer Slide Time: 18:37)

Mahalanobis Distance



$$(OP)^2 = d^2 = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad \dots(8)$$



S is sample covariance

Dr. I. Mohit, IITM, IIT Kharagpur

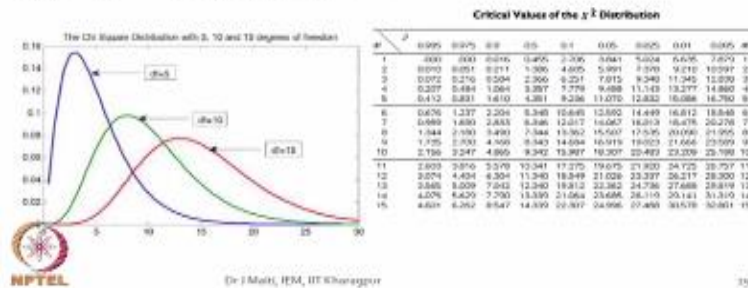
38

(Refer Slide Time: 18:39)

Distribution of d-square (Eq. 7 and 8)

d-square follows chi-square distribution with p degrees of freedom.

$$P\left[(x-\mu)^T \Sigma^{-1} (x-\mu) \leq \chi_p^2(\alpha)\right] = 1 - \alpha$$



What will be the distribution of this exponent part? Yes it is chi-square, why it is chi-square?

(Refer Slide Time: 18:56)

P C Mahalanobis
(151)
Ellipse \rightarrow Ellipsoid
$$P \left\{ (X-\mu)^T \Sigma^{-1} (X-\mu) \leq \chi_p^2(\alpha) \right\} = 1-\alpha.$$

population parameter.

$(X-\mu)^T$ then $\Sigma^{-1} (X-\mu)$ that is exponent part and minus up is there when causing their minus up, you see this $(X-\mu)^T$ and $(X-\mu)$ that is the square term of normal variable and this Σ in the variance part covariance part. So what your Σ is population Σ it is a constant value, so the square of normal is Z , and Z follows chi-square a linear combination of Z follow chi-square distribution.

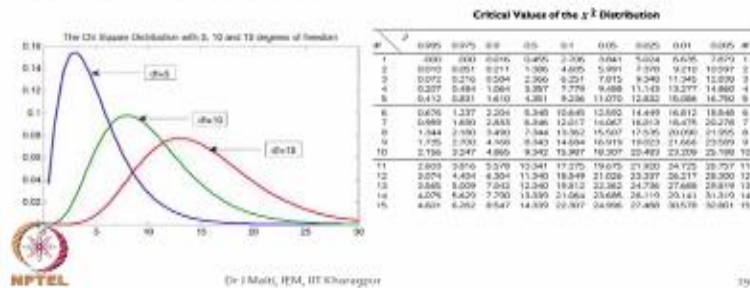
So, this will follow chi square distribution that what we are saying probability that this value will be less than equal to chi-square P , some α value, this will be $1 - \alpha$. Now, in this equation μ , Σ , Σ are population parameters and X is the random variable.

(Refer Slide Time: 20:27)

Distribution of d-square (Eq. 7 and 8)

d-square follows chi-square distribution with p degrees of freedom.

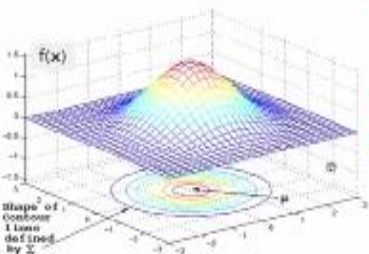
$$P\left[(x-\mu)^T \Sigma^{-1} (x-\mu) \leq \chi_p^2(\alpha)\right] = 1 - \alpha$$




And your chi-square distribution will be a different types, the shape of the chi-square distribution can be like this, can be like this, can be like this depending on the degrees of freedom and you all know how to see chi square distribution, chi square distribution degrees of freedom is there one side, another side the probability values.

(Refer Slide Time: 20:50)

Constant Density Contours



- A **contour line** of a **function** of two variables is a **curve** along which the function has a **constant value** (Wikipedia).
- In this case, the constant value is **constant density**.
- The curve is defined by the following equation:
$$(x - \mu)^T \Sigma^{-1} (x - \mu) = \chi_p^2(\alpha)$$

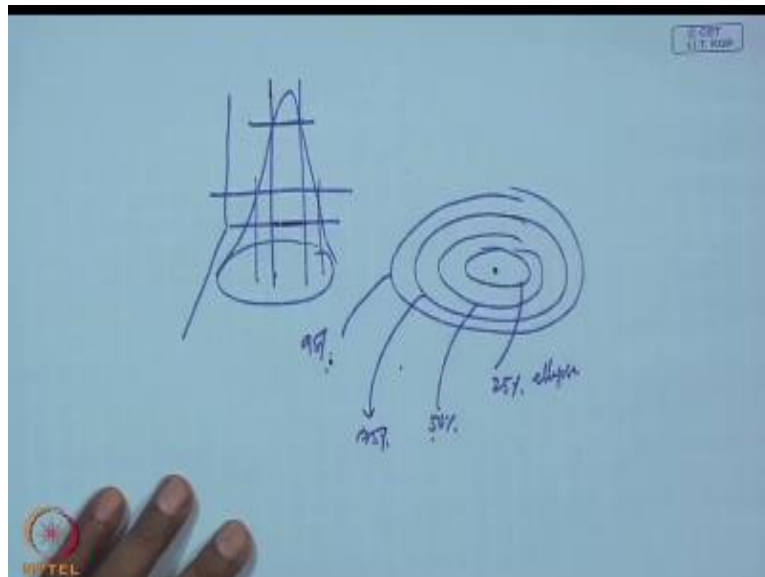
 Dr. J. Mohi, IISM, IIT Kharagpur 20

Then we will discuss contour, what is basically constant density contours? You see this figure carefully, this is bi-variate density function, this is the μ that is the bottom reference point we were seeing that is coming through, if you just see from the top if you see what you will see, you will first see the point for this top point, then if you come little distant lower and take a cross section then you will be seeing this ellipse.

If you come even little more you will be getting the second ellipse, like this you will be getting different ellipse and as we have already discussed that all points along this ellipse is equidistance, we have proved this one by statistical distance they are equidistance. So, here what is happening? The total dataset, the total data which is represented by this MND here it is basically bi-variate normal distribution and you are capturing.

Now, the more you want to include more number of observations your ellipse will be bigger. What does it mean?

(Refer Slide Time: 22:30)



What I am trying to say, suppose your case is like this, this is the case, now if you take this, you take like this, this is cross section here then what is the portion you are considering? You are considering a smaller portion, if you take a cross section, here you are considering a bigger portion. So, as a result what is happening your ellipse is becoming like this. So, I can say this may be 25 % ellipse, this is may be 50 % ellipse, this one may be your 75 % ellipse, this may be 95 % ellipse. What does it signify?

This signifies that 25 % of observation will be on or within this ellipse, 50 % will be within or on this ellipse, 75 % will be on or within this ellipse, then 95 % will be on the ellipse.

(Refer Slide Time: 23:44)

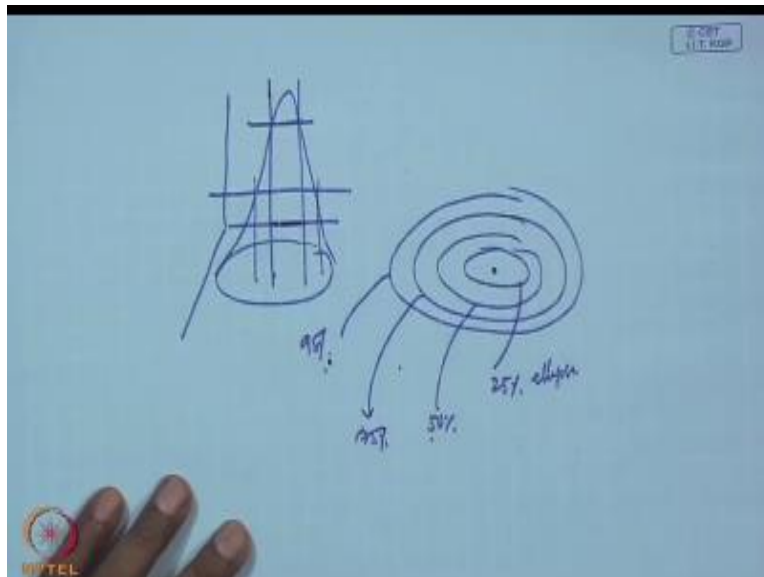
P C Mahalanobis
(151)
Ellipse \rightarrow Ellipsoid

$$P \left\{ (X-\mu)^T \Sigma^{-1} (X-\mu) \leq \chi_p^2(\alpha) \right\} = 1-\alpha = 0.95$$

population parameter. $\alpha = 0.05$

And that is what is this one, α if you choose $\alpha = 0.05$ then you are creating an ellipse in such a manner that this will be 95 %, 0.95 that means you are considering 95 % observations to be within or on the ellipse, rest 5 % will be in other, out of this region.

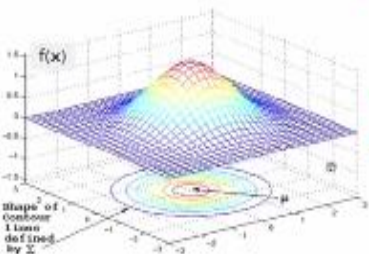
(Refer Slide Time: 24:19)




So, why we are saying this one is constant density contours?

(Refer Slide Time: 24:21)

Constant Density Contours



- A **contour line** of a **function** of two variables is a **curve** along which the function has a **constant value** (Wikipedia).
- In this case, the constant value is **constant density**.
- The curve is defined by the following equation:
$$(x-\mu)^T \Sigma^{-1} (x-\mu) = \chi_p^2(\alpha)$$

 Dr. J. Mohi, IISM, IIT Kharagpur 20

So why we are saying this is constant density contours? Why we are saying this is concentrated contours, what is contour? Elliptical region is a plan that is why I said if you see the plane view from the top you see in the floor itself the footprints will be generated. An equal distance points will be making one ellipse, why constant density contour? What is the contour? A contour line of a function of two variable is a curve along which the function has a constant value, this is Wikipedia definition.

So, our function is $f(x)$ X_1, X_2 that is a function then we are taking when we are making a cross section here that means, we are taking a constant distance, constant density. Suppose, you are taking this 0.5 as a density then along this the problem, 0.5 we are making a cross section and that cross section will bring certain ellipse here also. And as a result we are saying that this is constant density, that mean all the points on this ellipse is constant density, having constant density.

Yes, yes it can be, it is not that only two variable case for example, we want to draw a contour on the wall map for the cities having equal altitude, that mean from the mean sea level you will find out the cities altitude value. If you join a line that is also contour distance contour, similarly

temperature contour, same manner is just you think here what is happening you think from a hill point of view, this is a hill.

Now, from the floor of the hill bottom of the hill you are going up and after certain height make a cross section as this is the for example, this is my floor, this is the height. So, this is the density you are making a cross section, there need not be that it will be a two variable function of two variables, it can be function of many variables and ultimately our case is multiple variables only.

(Refer Slide Time: 27:29)

Example-2

Consider the data given in example-1. Obtain constant density contours for $\alpha = 0.05$ and $\alpha = 0.10$.

$$d^2 = \left(\frac{x_1 - 100}{\sqrt{10}} \right)^2 + \left(\frac{x_2 - 50}{\sqrt{5}} \right)^2 = \chi^2_2(\alpha)$$

and $\chi^2_2(0.05) = 5.99$, & $\chi^2_2(0.10) = 4.61$

Level of Significance	Critical Value of chi-square
0.05	5.99
0.025	7.38
0.01	9.21
0.005	10.59

Dr. J. Mohan, IEM, IIT Kharagpur 23

Here, one example consider the data given in example one obtain constant density contours for $\alpha=0.05$ and $\alpha=0.01$ and we have seen that the d^2 this is x_1-100 by square root of 10 square. This one we have given earlier considering this equal to chi-square alpha. Now, when $\alpha=0.05$ then this will be chi-square 0.05 variable is 2 so, degree of freedom is 2 and you have to see this value. What is this value? This value is 5.99.

So now, this is the ellipse, which is basically taken into consideration, chi-square 2.05, 5.99 this is the ellipse what is basically constant, all the points are in equal density value. Now, how we get this 5.99 value? This is what is the chi-square distribution and when your degrees of freedom

for that distribution is 2 and our value is 5.99, this is 0.055.99 this value if it is 0.10 then your value will be 4.61 so, ellipse will be little smaller okay any question from your side?

No, here that degrees of freedom case is purely dependent on the number of variables considered, whether they are correlated or not correlated it is a material. No, the ultimate exponent value will be different because distribution will not change. No that will not be that is not correct.

(Refer Slide Time: 30:15)

Why if there are P variables then our case is to the exponent is this one $e^{-1/2T}$ this one you are talking about this part. Whether there is, because this is covariance matrix Σ so definitely there is correlation or covariance, we are considering this correlation component or covariance component this here basically, if you say this we are not going to statistic, it is a population domain. So, this expression is chi square distributed. Degrees of freedom always okay that mean what we are saying that $X-\mu^T$ this $X-\mu$ it is chi square P.

Now, I think your question is suppose if you go for some transformation of the variable then your number of variable will reduce because of correlations. Suppose, if the data is correlated

when you go for principal component analysis, like analysis what will happen your number of dimension will reduce. No, here it is not like this, here the covariance component is well accounted for the derivation part okay.

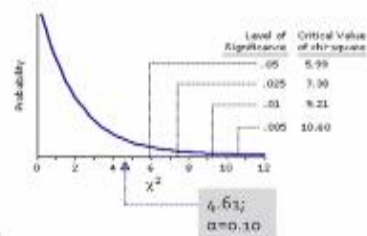
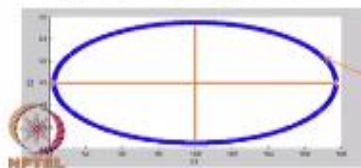
(Refer Slide Time: 32:08)

Example-2

Consider the data given in example-1. Obtain constant density contours for $\alpha = 0.05$ and $\alpha = 0.10$.

$$d^2 = \left(\frac{x_1 - 100}{\sqrt{10}} \right)^2 + \left(\frac{x_2 - 50}{\sqrt{5}} \right)^2 = \chi_2^2(\alpha)$$

and $\chi_2^2(0.05) = 5.99$, & $\chi_2^2(0.10) = 4.61$




$\chi_2^2(0.05) = 5.99$

(Refer Slide Time: 32:09)

Examining data for MND

- Testing multivariate normality is crucial
- Techniques
 - Probability plots
 - Q-Q plot
- Steps for Chi-square Q-Q plot
 - Step-1: Compute $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, n$
 - Step-2: Order d_i^2 as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
 - Step-3: Graph the pairs $\left(\chi_p^2((n-l+\frac{1}{2})/n), d_{(l)}^2 \right)$

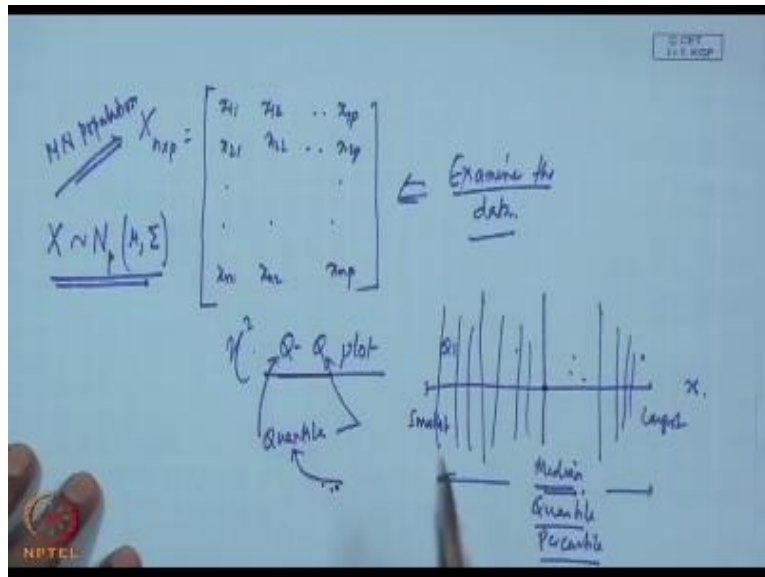


Dr. I. Mathi, IIM, Hyderabad

33

Now, we will discuss another important aspect that is very, very important for all of us that when we collect data.

(Refer Slide Time: 32:16)



Suppose, you collect a data $n \times p$ and our data points are like this. We are saying that we have collected the data from multivariate normal population, MN population, mean multivariate normal population and we are defining that our X is multivariate normal with p variables, μ and Σ the population parameters. What is the guarantee that your data is multivariate normal? It is true that if your population is multivariate normal the data we generate that will be multivariate normal also.

But you do not know you are you do not know whether that population is really multivariate normal or not. So, what we do up here, we examine the data to understand that whether data comes from multivariate population or not. So, in this case in multivariate domain we will use chi square quantile, quantile plot Q Q plot, chi square quantile, quantile plot, Q stand for quantile. What is quantile? Any idea?

You know median so how do we get the median? Your data point you first order the data from smallest to largest ascending and then you find out the middle location and you say if this is my variable X and what is the middle value of X , that is median. So, when you are making median,

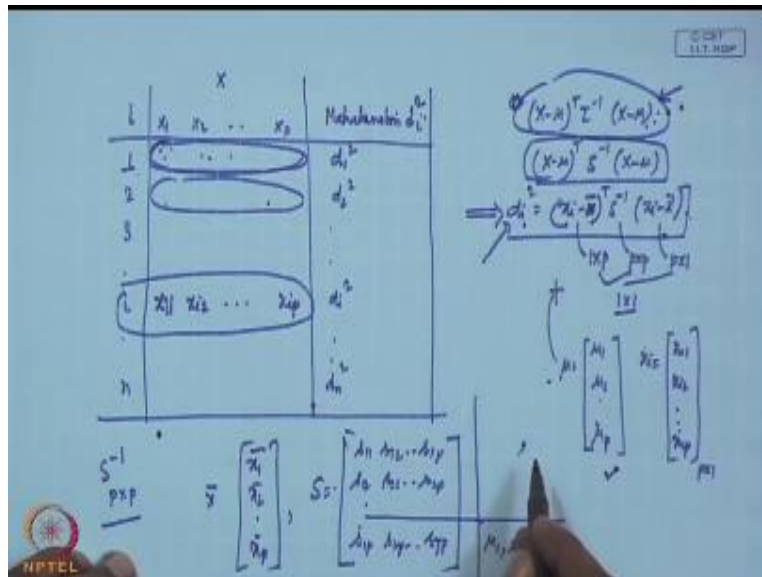
how many parts you are making for the data? Two parts okay 50, 50 where this is one side, this is another side.

You know quantile, quantile will make partition, the data total data into four parts. This is my quartile one, quartile two, quartile three, quartile four and we know that Q 1 and Q 3 the inter quartile range that we have discussed in descriptive statistics. So, while making quantile you are making partitioning the data into four parts. Now, you do one thing, you partition into hundred parts. What will happen?

Sorry, that is percentile, so median then you are saying quantile then you are saying percentile. Suppose, I do not want like median quantile, I want to partition the total set into n parts then what name you will give? You cannot give you have you have to have some name that is quantile. So, that mean if you partition the data into two parts, the median that is also quantile, general name quartile is also quantile percentile is also quantile and if you partition even more or less number than the hundred parts then that is also quantile.

So, by quantile, quantile plot, we want to see that how multivariate normality will be assessed okay.

(Refer Slide Time: 36:30)



For example, suppose you have collected n data points 1, 2, 3, n data points you have collected. Let the data is X_1, X_2 on x_p variables and I am sure all of you are in a position to fill up this. Now, we will calculate the d^2 Mahalanobis d^2 . I am writing this one as d_i^2 where i stands from 1 to n correct. So, what is this? So, you have some reference point., somewhere that reference point definitely if it is a p variable case then it is the μ vector μ_1, μ_2 like μ_p you are getting on p variable you are getting this is your first observation, this is second observations.

So, like this if there is i^{th} point you will be getting x_{i1}, x_{i2}, x_{ip} . So, this is your i^{th} observation. Now, you want to find out where does that i^{th} observation lie, when if we consider a two dimensional case. This is my μ_1 and μ_2 then somewhere here this is my i^{th} observation is falling. So, I want to get this distance, I think we have discussed the statistical distance part. What is the formula for statistical distance? $(X - \mu)^T S^{-1} (x - \mu)$.

Now, in Mahalanobis distance the formula is $(X - \mu)^T S^{-1} (x - \mu)$ getting me? So, when I say d_i^2 I am saying $x_i - \mu$ but I do not know the μ . So, I will write $x_i - \bar{x}$ S^{-1} $x_i - \bar{x}$. So, if you write in this format $X - \mu^T$ and this $S^{-1} x - \mu$ this is the exponent of the multivariate normal distribution that also follows

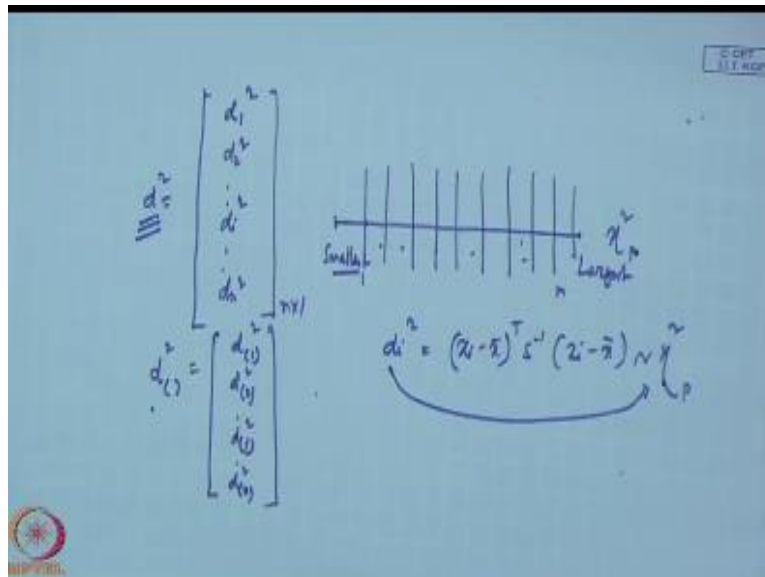
chi square distribution that is the general statistical distance. But when I say this is S^{-1} , this is we show that Mahalanobis has developed this.

So far every observation you can get this distance. Now, x_i is what, x_i is $x_{i1}, x_{i2}, \dots, x_{ip}$, $p \times 1$. So, $(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$ what will happen? This is $1 \times p$, S^{-1} will be $p \times p$ and this one will be $p \times 1$. So, resultant will be 1×1 . So, you will be getting a value I am writing here, suppose this is d_1^2 . For the second one you calculate d_2^2 , same formula like this you will be getting d_i^2 then d_n^2 , correct?

So, what are the steps then you have collected data there are n multivariate observations, we want to compute the distance of each observations from the mean vector. Here, we will be concentrating on the sample mean that is \bar{X}_1, \bar{X}_2 like \bar{x}_p , that mean so you require to calculate this mean vector, you also require to calculate the covariance matrix. Once you know this, you require to calculate S^{-1} , this also will be $p \times p$ matrix.

Once you do all those things you have computed then you go for d_i^2 , for $i=1$ take this row first, put here every observation will be subtracted by the corresponding mean okay then you go for the second one, third one, like this and using this equation you have n number of distance values. Now, depending on the observation value it will be different types.

(Refer Slide Time: 42:19)



For example, what I mean to say now from the set of multivariate observations you calculated the distance and finally, you got data matrix like this d_1^2, d_2^2, d_i^2 then d_n^2 , then the n^2 that is what data you got and this one I am saying, this is my d^2 matrix. What you require to do? Now, you have to find out the quantiles, please remember we say quantile quantile plot. So, if you require to know the quantile the one of the issue is you have to arrange these values observed values in ascending order.

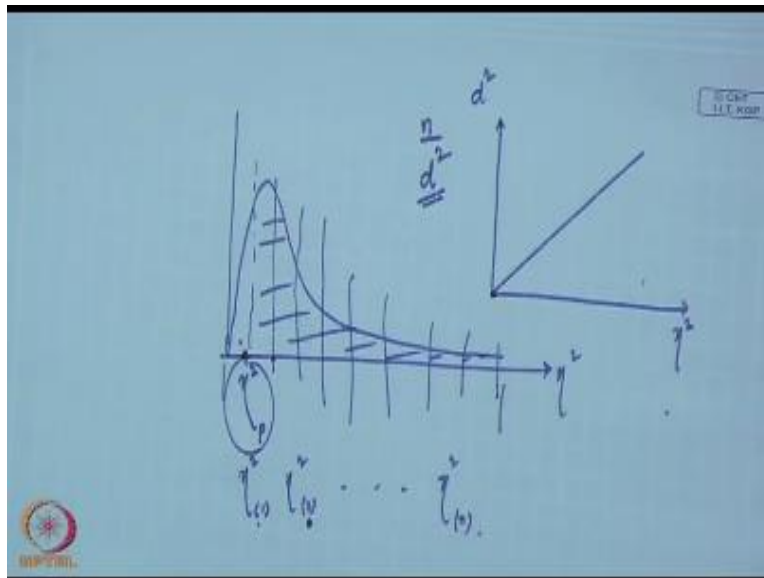
So, if I say my ascending data for the distance values are like this, d within bracket, 1^2 within bracket, 2^2 , like this d within bracket i^2 like, within bracket n^2 , this is my ordered data. So you will be having smallest to largest data set, fine? Now, how many data points you have? n data points, you create n quantiles here. So your 1, 2 now partitioning the entire data, this data transformed data, distance data partitioning this distance data into n quantiles.

What will be the distribution of this d_i^2 ? What is d_i^2 ? $(x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$ chi-square. So, this follows chi square as p variables so they are p then as I know that d^2 is chi square p d_i^2 and what you have created, you have created a data matrix now different data that is where $n \times 1$ transformed

data d and which is chi square distributed. So, that means all those you, this can be a chi square axis, I can get for quantile all points, chi square value I will be getting.

Now, d_i^2 is chi square distributed so from d_i^2 if I compare with the chi square quantile values I must get a relationship understood or not? This is nothing difficult.

(Refer Slide Time: 45:24)

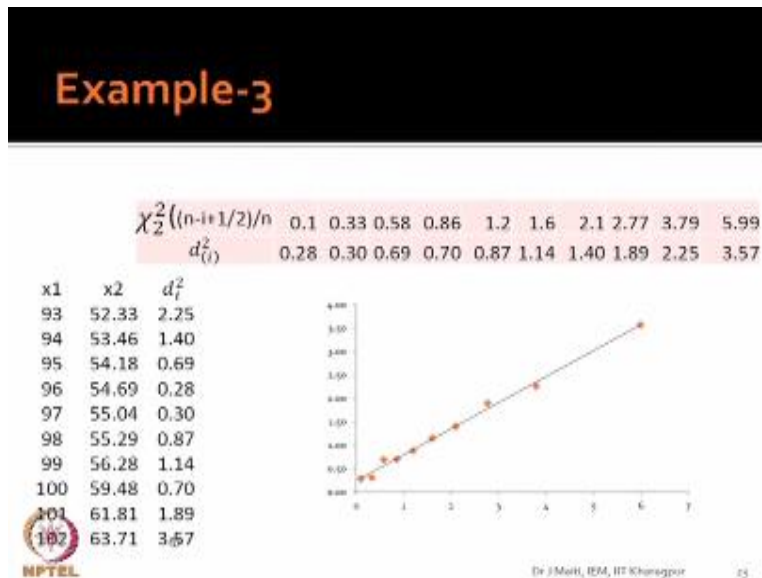


Suppose, this is my chi square distribution for example, let like this is your chi square, you know the lowest value and largest value from the data set, it is known what is this chi square value, if I say I want to know this chi square p what will be this value? We require to know, then what is the probability right hand side getting me? So, if I know that what is the probability of this? And if I subtract by 1, I will be knowing the what is the probability of this.

So, with respect to this probability from chi square table we will be getting this value. Similarly, for the second quantile you will also be getting, third quantile value you will be getting because you have n data set, this chi square X is each partitioned into n parts. So, everywhere I am saying that you have chi square 1, chi square 2, this one and two not degrees of freedom these are the that is the quantiles.

So, like this you will be getting chi square n values. Now, as we have said that d_i^2 is chi square distributed, so there will be relationship between d^2 and chi square. If you plot d^2 in this side or d_i^2 and chi square in this side you will be getting a straight line like this okay, you may get here, may get, no it will be a straight line linear relationship. Now, see the problem given here.

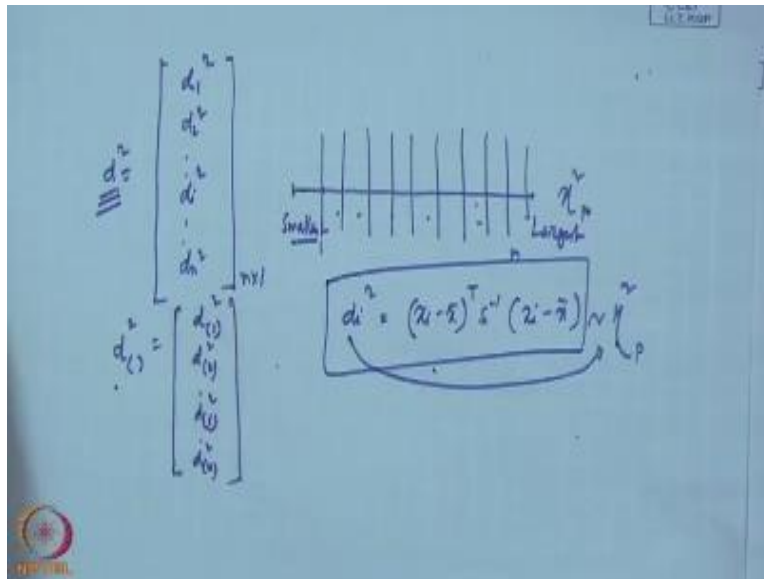
(Refer Slide Time: 47:24)



See this is my data set bi-variate case, we have considered 10 points, 1 to 10 and this is my first observation 93, 52.33. My second observation is 94, 53.46 like this what we have done? You have calculated the distance for this observation. First observation from the mean value, X_1 and X_2 that mean value we have to find out and from that mean value and then using the distance formula $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$ we computed this value 2.25.

So, my first observation is 2.25 distance apart from the mean, my second observation is 1.40, this distance unit distance apart from the mean, like this 102 and 63.71 this is the tenth observation is 3.57 units apart from the mean okay understood? Now, this is the calculation part. Only you have to utilize this formula, correct?

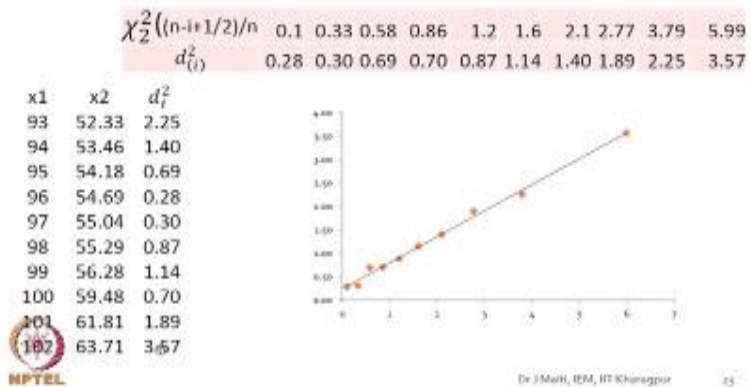
(Refer Slide Time: 48:50)



Carefully if you use this formula.

(Refer Slide Time: 48:57)

Example-3



You are in a position to get this, this is your first step.

(Refer Slide Time: 49:02)

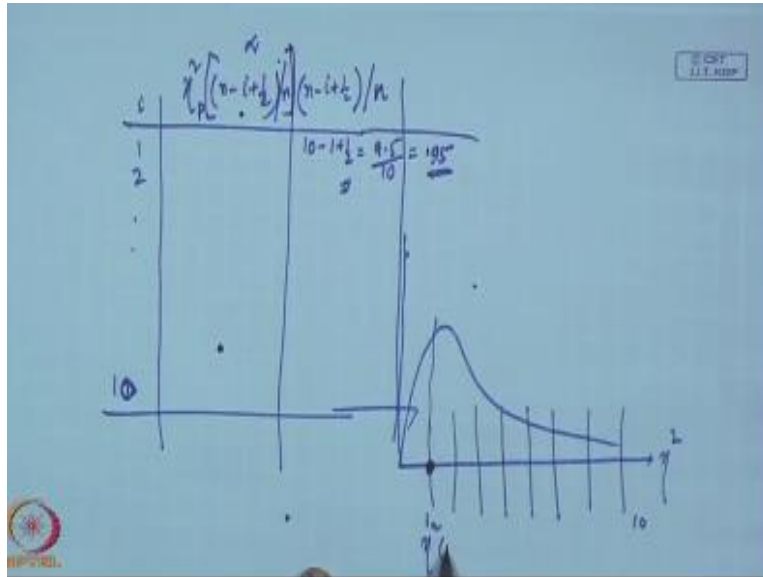
Examining data for MND

- Testing multivariate normality is crucial
- Techniques
 - Probability plots
 - Q-Q plot
- Steps for Chi-square Q-Q plot
 - Step-1: Compute $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, n$
 - Step-2: Order d_i^2 as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
 - Step-3: Graph the pairs $\left(\chi_p^2((n-1)/n), d_{(i)}^2 \right)$



What is your second step? Second step is order because you want to find the quantile values. So you order that is ascending order, third one is you find out the chi square quantile value. What is this chi square? $\chi_p^2((n-1)/n)$ this is the chi square quantile value. What is happening here you see, you see that.

(Refer Slide Time: 49:34)



You have 10 data points, $i = 1$ to 10 data points. So, these 10 data points order data points so that means your this is the chi square axis. Let it be this is the chi square variable axis so you have divided into 10 parts, 1 to 10 parts, you have divided and what will happen, your chi square will be something like this we have discussed. Now, you want to get the chi square value here. So, first you find out the probability value here right hand side you use because you will be using table.

Then for that probability value you find out the chi square value, that is why what you are doing here, you are writing chi square $p n - i + \frac{1}{2}$. So, if I want to do little more manipulation here what is $n - i + \frac{1}{2}$ n is 10. We have taken 10 data points, so this $10 - 1 + \frac{1}{2}$, 9.5. So, we want to find out the probability, so how many data points are there? 10 data points are there. So, if I divide by n again, so that by 10, so this that mean this will be 0.95 and this one also you write this by n because this probability value you want to, this is the probability value, α value. So, your first one first chi square, this value is chi square 0.95. What is this value?

(Refer Slide Time: 51:29)

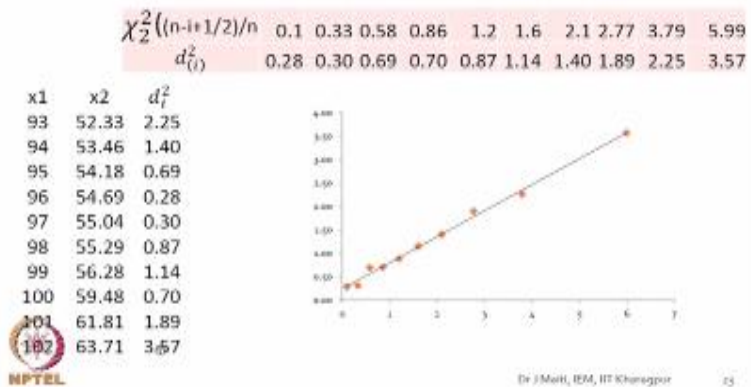
Examining data for MND

- Testing multivariate normality is crucial
- Techniques
 - Probability plots
 - Q-Q plot
- Steps for Chi-square Q-Q plot
 - Step-1: Compute $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, n$
 - Step-2: Order d_i^2 as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
 - Step-3: Graph the pairs $\left(\chi_p^2((n-t + \frac{1}{2})/n), d_{(t)}^2 \right)$



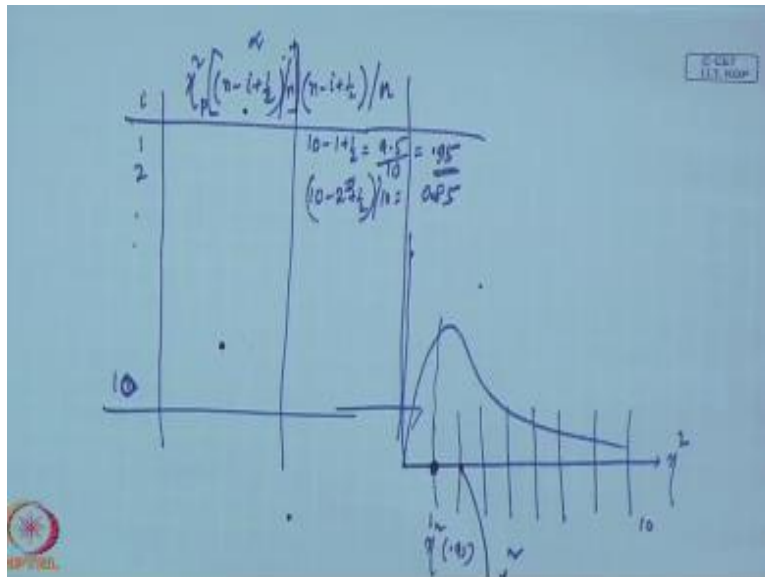
(Refer Slide Time: 51:31)

Example-3



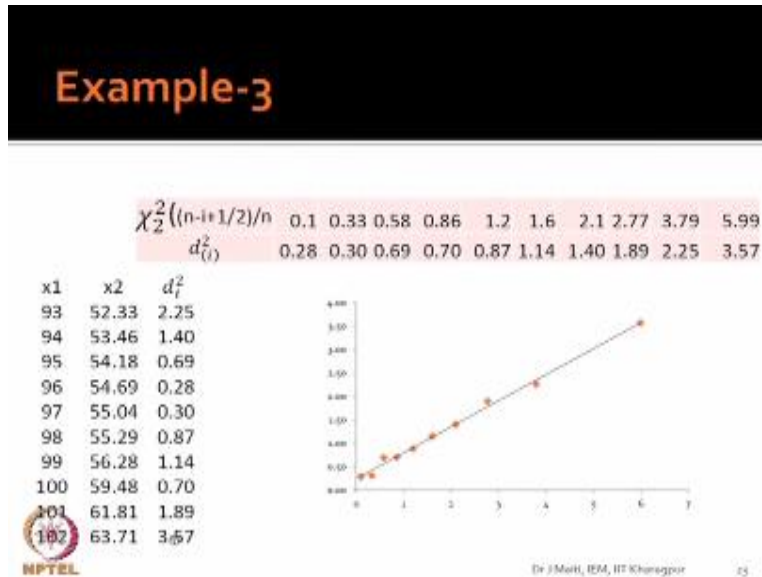
You see here, 0.1 then what will be the second one?

(Refer Slide Time: 51:42)



$10 - 2 + \frac{1}{2} / 10$ this will be 8.5 means 0.85. So, your second value here you find out for chi square 0.85 chi square, 0.85 is 0.33.

(Refer Slide Time: 52:10)



So, third one will be 0.75 0.58 0.65 chi square 0.65 0.56. So, like this then what we have done here? Then we have this side d_i^2 Y axis, X axis is chi square value and when you plot this, you will be getting this type of straight line, getting me? So, clear that your total data set is this one.

(Refer Slide Time: 52:59)

The whiteboard contains the following handwritten content:

i	x_i	x_2	\dots	x_p	Mahalanobis d_i^2
1	\dots	\dots	\dots	\dots	d_1^2
2	\dots	\dots	\dots	\dots	d_2^2
3	\dots	\dots	\dots	\dots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	x_{i1}	x_{i2}	\dots	x_{ip}	d_i^2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	\vdots	\vdots	\vdots	\vdots	d_n^2

Formulas and derivations:

- $(X-\mu)^T S^{-1} (X-\mu)$
- $(X-\mu)^T S^{-1} (X-\mu)$
- $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$
- $S^{-1} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{bmatrix}$
- $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$
- $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$
- $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$

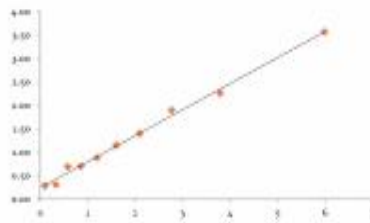
Every observations is transformed into one value here, we get value that is the Mahalanobis d^2 . And you are now finding out the distribution for this and you all know this is chi square distributed using this property.

(Refer Slide Time: 53:19)

Example-3

$\chi^2_{(n-i+1/2)/n}$	0.1	0.33	0.58	0.86	1.2	1.6	2.1	2.77	3.79	5.99
$d_{(i)}^2$	0.28	0.30	0.69	0.70	0.87	1.14	1.40	1.89	2.25	3.57

x1	x2	d_i^2
93	52.33	2.25
94	53.46	1.40
95	54.18	0.69
96	54.69	0.28
97	55.04	0.30
98	55.29	0.87
99	56.28	1.14
100	59.48	0.70
101	61.81	1.89
102	63.71	3.57



(Refer Slide Time: 53:20)

Examining data for MND

- Testing multivariate normality is crucial
- Techniques
 - Probability plots
 - Q-Q plot
- Steps for Chi-square Q-Q plot
 - Step-1: Compute $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, n$
 - Step-2: Order d_i^2 as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
 - Step-3: Graph the pairs $\left(\chi_p^2((n-t + \frac{1}{2})/n), d_{(t)}^2 \right)$



And the steps like this plot, if you get a straight line.

(Refer Slide Time: 53:27)

Handwritten mathematical derivations on a blue background:

Top left: Matrix X with columns x_1, x_2, \dots, x_p and rows $1, 2, 3, \dots, n$. To its right is a column of eigenvalues $d_1^2, d_2^2, \dots, d_n^2$.

Top right: Quadratic form $(X-u)^T S^{-1} (X-u)$ and its expansion $\Rightarrow d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$.

Bottom left: Inverse covariance matrix S^{-1} and covariance matrix $S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix}$.

Bottom right: Mean vector $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$ and data vector $x/c = \begin{bmatrix} x_{1c} \\ x_{2c} \\ \vdots \\ x_{pc} \end{bmatrix}$.

That is multivariate normal. If we depart substantially then definitely that is not multivariate normal, but there will always be some amount of departure. So, what is the departure acceptable that also you want to check.

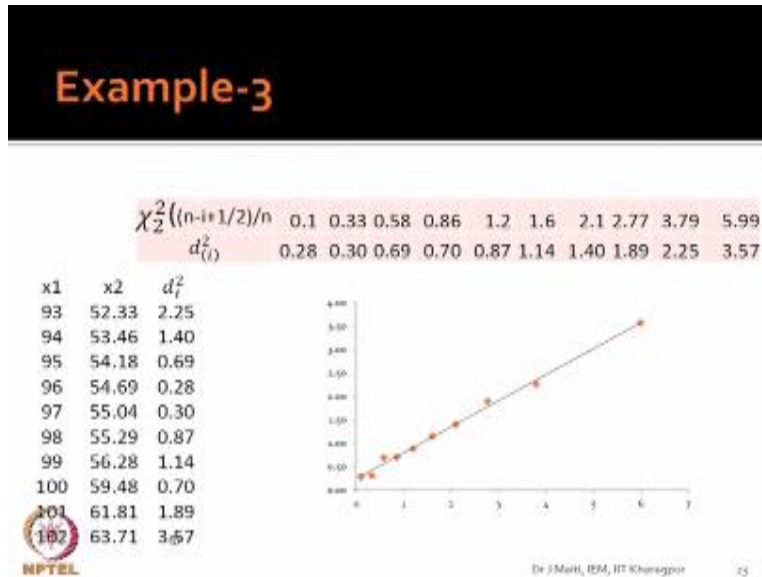
(Refer Slide Time: 53:46)

Examining data for MND

- Testing multivariate normality is crucial
- Techniques
 - Probability plots
 - Q-Q plot
- Steps for Chi-square Q-Q plot
 - Step-1: Compute $d_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}), i = 1, 2, \dots, n$
 - Step-2: Order d_i^2 as $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
 - Step-3: Graph the pairs $\left(\chi_p^2((n-t + \frac{1}{2})/n), d_{(t)}^2 \right)$



(Refer Slide Time: 53:46)



So, there are some other methods for checking also, like case testis there, Kolmogorov Smirnov test, d max. So you will be able to check all those things okay this is what is our in totality the multivariate normal distribution, we have discussed along with the properties so my, as it is very important one.

(Refer Slide Time: 54:19)

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} [(x-\mu)^T \Sigma^{-1} (x-\mu)]}$$

$$-\infty < x_j < \infty, j=1, 2, \dots, p.$$

- * Constant
- * Exponent \leftarrow
- * $(x-\mu)^T \Sigma^{-1} (x-\mu) \sim \chi^2_p$
- * Constant density
- * Resembles an ellipse when $p=2$
- * Statistical distance \leftarrow
- * Constant density contours
- * Properties of MND

$X \sim N_p(\mu, \Sigma)$
 $x_j \rightarrow N(\mu_j, \sigma_j^2)$
 $j=1, 2, \dots, p.$
 $x \sim N_p(\mu, \Sigma)$
 $a^T x \sim N(a^T \mu, a^T \Sigma a)$
 $A^T x \sim N(A^T \mu, A^T \Sigma A)$

So, first of all you must know that your multivariate normal density is $1 / (2\pi)^{p/2}$ then your covariance to the power $1/2 e^{-1/2 (x-\mu)^T \Sigma^{-1} (x-\mu)}$ okay and definitely your all x_j less than equal to infinite $j = 1$ to p . So, there are two parts, one is constant and another one is exponent and it is exponent which is very, very important. So, the exponent part $(x-\mu)^T \Sigma^{-1} (x-\mu)$ it follows chi square p , this distribution is chi square p . And this is also known as constant, this will give you the constant density.

Now, this formula is chi square, this is chi square, but this quantity resembles an ellipse okay this will be an ellipse when $p = 2$ for more than two variables it will be ellipsoid okay. Then there is another concept called statistical distance keep in mind this is very, very important concept and constant density contours the properties of multivariate normal distribution is important properties of MND that if x is what we say properties X is multivariate normal then x_j will be univariate normal. That we have seen $j = 1$ to p .

Now, if X is multivariate normal subset will also be multivariate normal, so that mean if I create a subset $x_{q \times 1}$ that will be $q \times \mu_q$ and Σ_q then linear transpose of x this will be your univariate normal with $A^T \mu$ $A^T \Sigma A$. If you find out q linear transpose then that will be your multivariate

normal that is $A^T\mu$ only the matrix multiplicability, compatibility part from multiplication point of view we have to check okay.

And then finally, when you collect data that data must be examined and you have to check that whether, the data is coming from multivariate normal or not, and that is possible through chi square quantile, quantile plot. There are other plots also, but we will be looking into this okay. So, I have given you one data set, that five variable data set if you can find out whether the data coming from multivariate normal or not using this statistical distance concept or you take your own data set just test use excel or mat lab okay, thank you.

NPTEL Video Recording Team

NPTEL Web Editing Team

Technical Superintendents

Computer Technicians

A IIT Kharagpur Production

www.nptel.iitm.ac.in

Copyrights Reserved