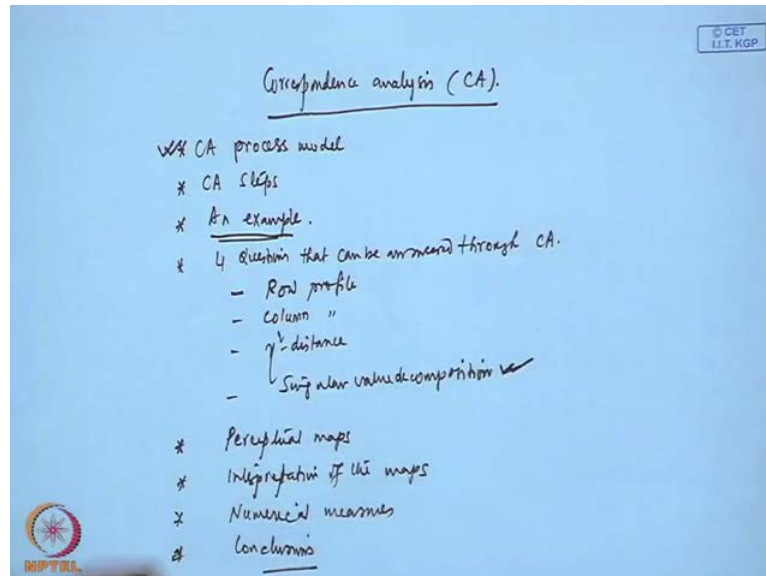


Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 42
Correspondence Analysis (Contd.)

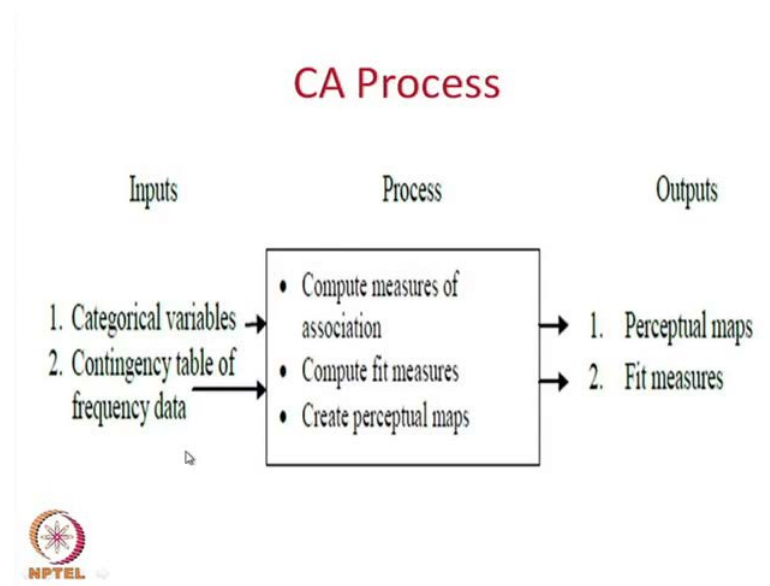
(Refer Slide Time: 00:22)



Good morning. Today, we will continue correspondence analysis. Correspondence analysis CA. Last class, we have seen the CA process model, then CA steps, then with an example. We discussed about four questions, four questions that can be answered through CA, through CA. Then here we have defined row profile, column profile, the chi squared distance and we were about to start the singular value decomposition singular value decomposition.

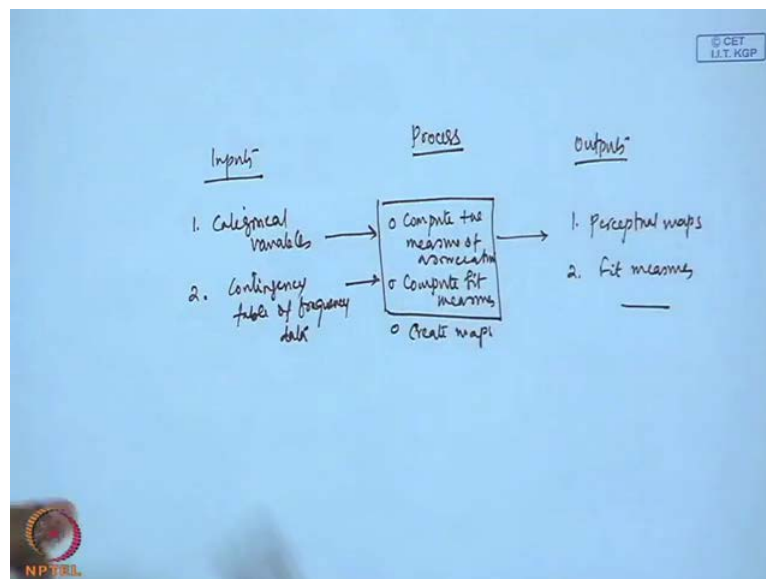
So, today we will discuss from singular value decomposition and the example we will continue. Then we will discuss how perceptual maps can be created perceptual maps, then interpretation of the maps, interpretation of the maps, then there are certain numerical measures, which ultimately help us to quantify the fit of correspondence analysis that also we will explain. Then finally we will conclude from the case study. So, what I show you now initially I will again show you the CA process model. Then we will straight away go to the singular value decomposition.

(Refer Slide Time: 02:59)



See the CA process model what I said earlier that it has inputs, inputs.

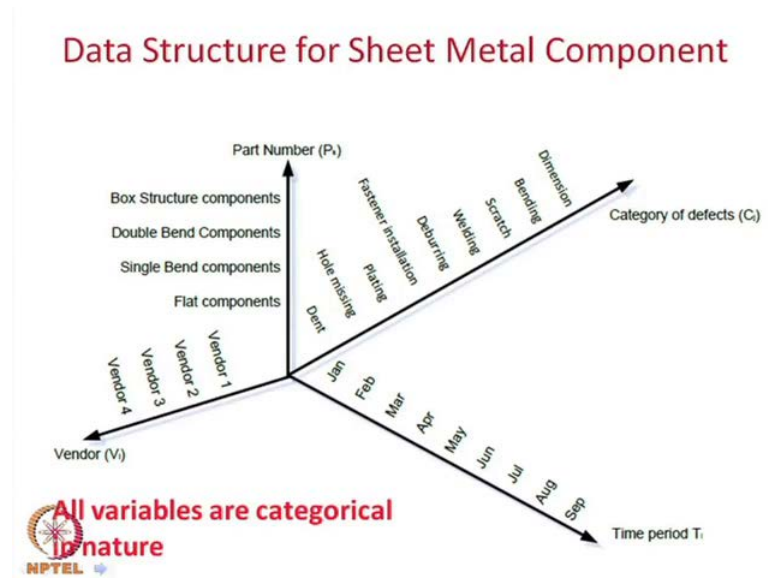
(Refer Slide Time: 03:09)



So, under inputs coming categorical variable, then you are coming to contingency table of frequency data. These are input to the process input to the process. Here what you do under process? You compute the measure of association, then compute fit measures, then you develop or create perceptual maps, create maps. What is your output? Outputs are one perceptual map and two fit measures. This is the model. So, here in this particular,

the first case, compute the measure of association. We have shown row profile column profile chi squared distance.

(Refer Slide Time: 04:49)




Let us go to now singular value decomposition here. Let me tell you one more thing that we are talking about the defect in sheet metal component of a medical diagnostic equipment and which are supplied by four vendors. The components are supplied by four vendors. There are different part numbers, different time period and different categories of defects and in when you compare any two of these categorical variables with the defects data.

(Refer Slide Time: 05:23)

Contingency Tables

Categorical data are collected in terms of frequencies and the data table is known as contingency table

	BD	D	DM	FI	HM	P	S	W	Total
Vendor 1	150	137	207	91	76	210	185	20	1076
Vendor 2	142	139	200	120	105	221	185	29	1141
Vendor 3	146	130	193	114	87	205	148	20	1043
Vendor 4	57	68	269	260	87	159	239	42	1181
Total	495	474	869	585	355	795	757	111	4441


 BD – Bend/dent; D – Deburring; DM – Dimension related; FI – Fastener installation; HM – Hole missing; P – Plating; S – scratch; W – Welding

We will be able to find out this type of contingency tables as there are four categorical variables. So, 4 C 2, 6 contingency tables can be created. In this lecture, we are basically talking about analysis of each of the contingency table with respect to correspondence analysis.

(Refer Slide Time: 05:48)

Answers using CA

- Q1 – Row profiles (R)
- Q2 – Column profiles (C)
- Q3 – Weighted χ^2 -distances (D)
- Q4 – Singular value decomposition (SVD) and perceptual map



So, there are many questions, actually four questions and four answers.

(Refer Slide Time: 05:55)

Key Questions

- Q1: What are the similarities and differences among the 4-vendors with respect to the 8-category of defects?
- Q2: What are the similarities and differences amongst the 8-category of defects with respect to the 4-vendors?
- Q3: What is the relationship between vendors and category of defects?
- Q4: Can these relationships be represented graphically in a joint low-dimensional space?



The last question was that can these relationships that means the question three was what is the relationship between vendors and category of defects? Now, question four is can these relationships be represented graphically in a joint low dimensional space? The question three will be answered through chi squared distance, but question four requires some advanced matrix manipulation that is known as singular value decomposition.

This is a singular value decomposition is similar to Eigen value Eigen vector decomposition of a matrix, but Eigen value and Eigen vector decomposition is possible only when the matrix is a square matrix. When the matrix is rectangular matrix, that decomposition is not possible. So, we will go for singular value decomposition which makes rectangular matrix to be decomposed into different vectors and scalar quantities.

(Refer Slide Time: 07:00)

Singular Value Decomposition (SVD)

- SVD is applied to partition the **D** matrix into three matrices **U**, **V** and **S**, where **U** is a $p \times k$ matrix, **V** is a $q \times k$ matrix, **S** is a $k \times k$ diagonal matrix with diagonal elements in the form $s_1 \geq s_2 \geq \dots \geq s_k > 0$, and k is the reduced dimensions.
- s_k is the singular value for the k^{th} PC. The square of s_k , i.e. s_k^2 is the eigenvalue (λ_k) of the k^{th} PC.
- The eigen value λ_k , represents the weighted variance explained by the k^{th} PC extracted.

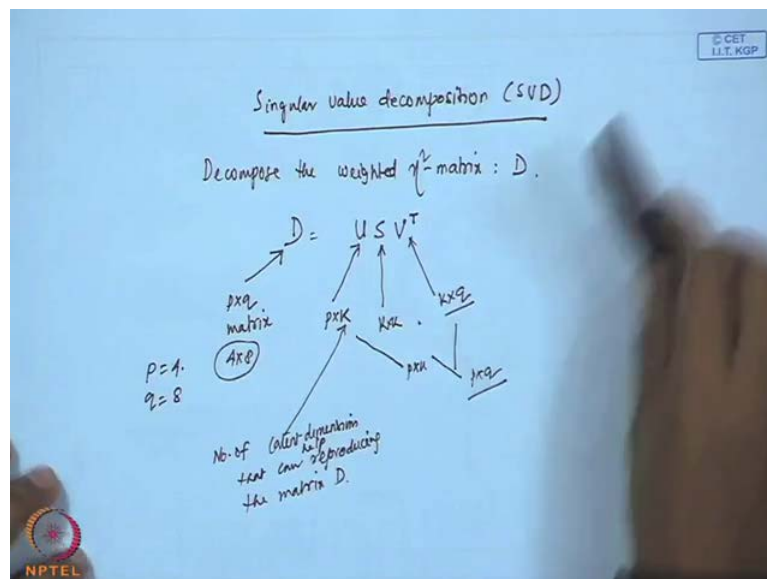


$$k = \min\{p-1, q-1\}$$

$P = 4$ and $q = 8$ for the example shown. So, $k = 3$. But $k = 2$ often serves the purpose.

Now, this is what is singular value decomposition?

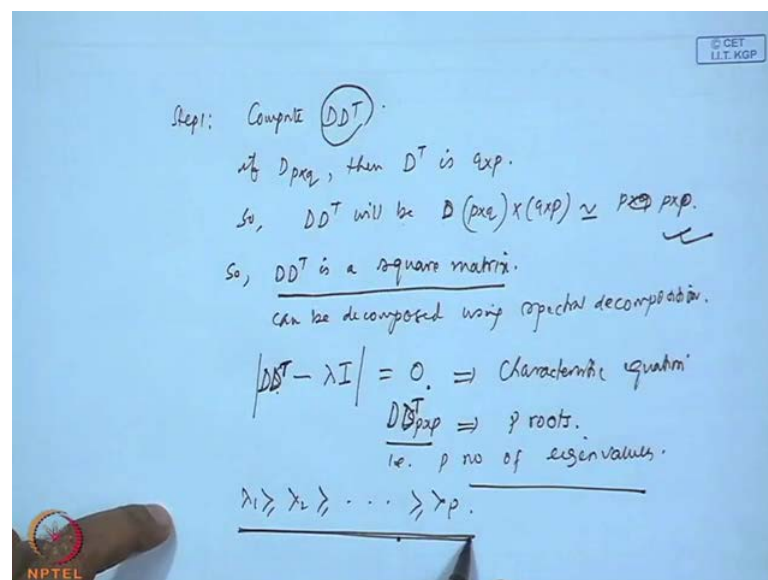
(Refer Slide Time: 07:11)



Singular value decomposition SVD, here we want to decompose the weighted chi squared matrix. This is basically chi squared distance matrix D . So, we want to decompose in such a manner that D will be $U S V^T$. Now, what is U ? Now, our D is let D is our p plus p cross q matrix, this p cross q matrix. For example, if there are four vendors, then p equal to 4 and eight categories of defect q equal to eight, then it will be 4 cross 8 matrix.

Now, we will create a matrix here, p cross k where k is the hidden dimensions latent dimensions that is number of latent dimensions. We can say number of latent or hidden dimensions that can that can reproduce or that can help reproducing the matrix D , matrix D when multiplied through U S and V transpose. So, what I mean to say this k is the dimensions, reduced dimensions, k is the reduced dimension. What we want to know how? How this reduced dimension is coming this S ? S will be your k cross k matrix. Then this will be your V transpose, so k cross q matrix, so p cross k , k cross k into k cross q , so this two will this two will lead to p cross q . Now, how do you get this U , S and V ? What is this method? So, I will discuss little some steps.

(Refer Slide Time: 10:12)

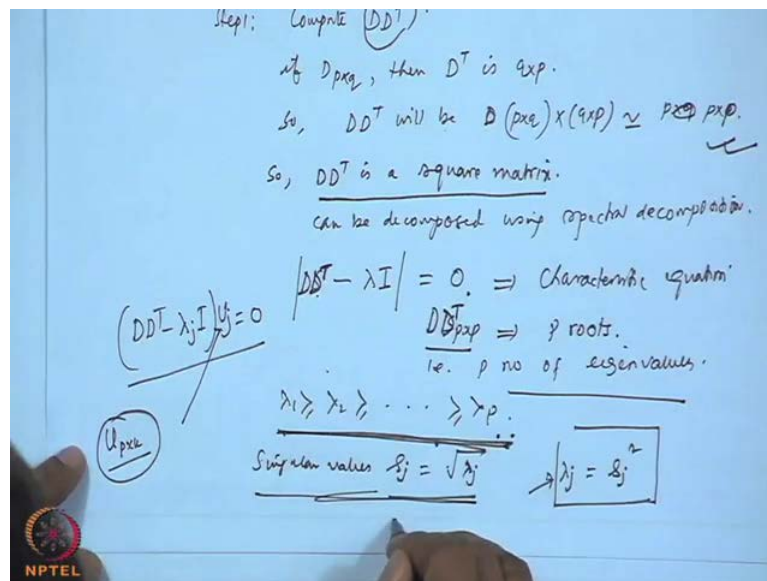


Step one is compute DD transpose. Now, if D is p cross q , then D transpose is q cross p . So, DD transpose will be, it will be p cross q into q cross p that is p cross p cross p . So, this is a square matrix. So, DD transpose is a square matrix. Now, this can be decomposed using Eigen value Eigen vector decomposition. So, it can be decomposed using spectral decomposition or Eigen value Eigen vector we will be able to find out.

I have shown you in principal component analysis that when a matrix S is a square matrix, then it is decomposed for Eigen values, then the determinant of this minus this is put to 0. Then we will get a characteristic equation. This is the characteristic equation, characteristic equation and the root of this equation there suppose S is p cross p , then there will be p roots of this equation.

So, that means that mean p number of Eigen values you will get in such a manner that if Eigen values are lambda 1, lambda 2 to lambda p, the lambda 1 greater than lambda 2 greater than like this greater than lambda p; it is the Eigen values. So, first you do this find out the Eigen values and then putting, here in our case here, it is DD transpose minus lambda I and you are making this case. Now, DD transpose is p cross p we have seen. So, you will be getting p roots, this p Eigen values.

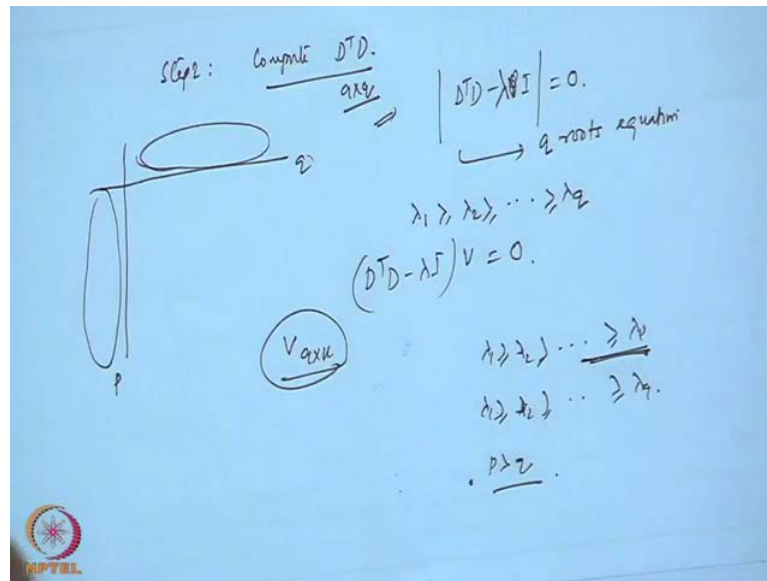
(Refer Slide Time: 13:02)



Now, singular values are supposed if I want I want to know the singular values singular values. If I say s_j , then singular values are basically the square root of lambda j. Other way, what we can say that each Eigen value is square of its singular value in this equation. You will be getting what? So, you will be getting like this DD transpose minus suppose if you say lambda j I into U_j , this will be 0.

Once you get particular lambda j, you will put this one is what is U in our case. So, if there are, if there are k hidden dimensions which can capture the variability in D transpose. Then ultimately our U total that matrix what will be getting that we will be getting U p cross q. So, from here, you are getting the Eigen values, this is the Eigen vector. So, you will be getting these singular values and Eigen, these values will be converted to singular values and Eigen vectors to Eigen values.

(Refer Slide Time: 14:30)



Step two is you compute D transpose D. The other way now it will be a q cross q matrix. So, if I see my contingency table, this is p and this is q. When I am getting a p cross p matrix, I am getting in terms of row. When you are getting q cross q matrix, you will be getting in terms of column. Now, again you do the same thing that D transpose D minus suppose gamma I, actually you write lambda I, no problem that equal to 0 here. You will be getting a q roots equation.

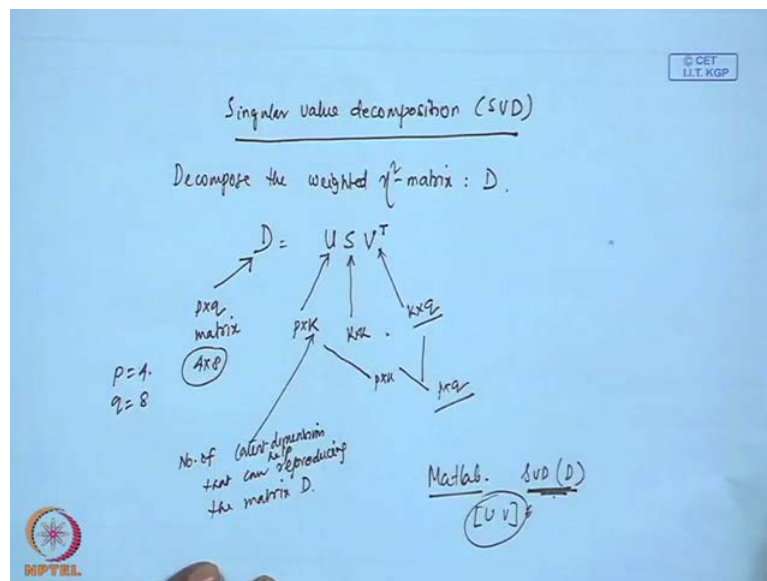
So, ultimately your case will be like this, all the Eigen values this one and using this, DTD minus lambda I, then this one is V is equal to 0. So, you will be getting a V, this is the Eigen vector that is q cross k you will be able to get. Interestingly, what you will find in, earlier you found that lambda 1 greater than equal to lambda 2 greater than equal to lambda p. Here, lambda 1 greater than equal to lambda 2 greater than equal to lambda q. So, ultimately p greater than q. So, what will happen? Ultimately that p minus q lambda values will be 0 here, but because of the special structure of the correspondence contingency table, you will find out here that that k is minimum of p minus 1 and q minus 1.

So, in this in our case, p is 4 and q is 8. So, k equal to minimum of 4 minus 1 and 8 minus 1 that is minimum of 3 and 7. This is 3, so three dimensions is enough for our work because for the perceptual map to be generated, what for the lower dimensions, we are basically trying to reduce the dimensions here. So, k is enough that is the dimensions

number of dimensions, which will ultimately explain the variability involved in the contingency table data.

So, this is basically that means what happened is you can go for the steps provided here for the singular values. As you will find out that even though we are basically going for Eigen value Eigen vector decomposition for two different matrix DD^T and $D^T D$, but as this square matrix, both are square matrix. You know that that the Eigen values are same for a matrix and its transpose. So, both the cases, you will be getting the same Eigen values. Eigen vector will differ because Eigen vector differ, the reason is one is in the $p \times p$ case that is for the row, other one is the column case though Eigen vectors V differ. So, this decomposition is known as singular value decomposition and singular value are the square root of Eigen values.

(Refer Slide Time: 18:20)




Alternatively, what happens ultimately if you use matlab, if you use matlab, if you give that SVD, suppose SVD D writing like U and V and something, this equal to this is SVD D. There is some changes may be, but SVD D in matlab will give you all those things U, S, V. So, once U, S, V is known, U, S as well as or as well as V is known.

(Refer Slide Time: 18:59)

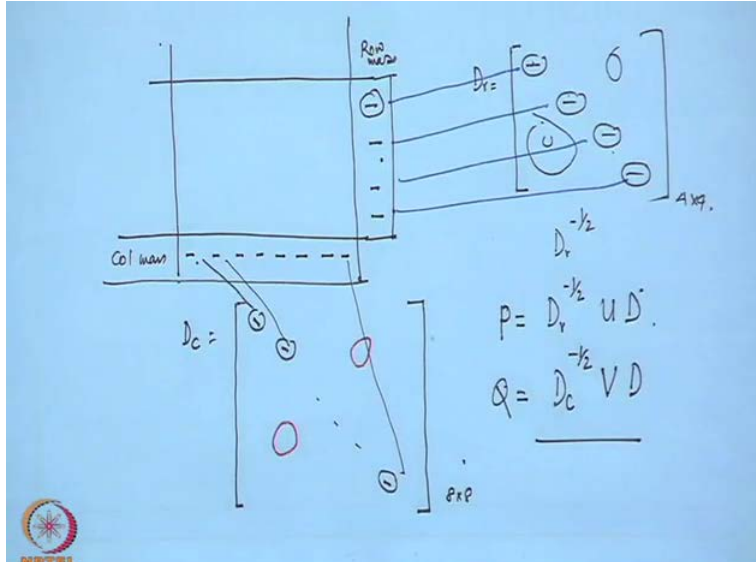
Obtain Perceptual Map

- Obtain row (vendor) PCs
- Obtain column (categories of defects) PCs


$$P = D_r^{-1/2} U D$$
$$Q = D_c^{-1/2} V D$$


Then, you can go for obtain row PCs and obtain column PCs. PC means principal component. Now, P is D_r to the power half U and D. I think now you know what is your D_r ?

(Refer Slide Time: 19:26)



The diagram shows a contingency table with 4 rows and 8 columns. The row masses are labeled as D_r and the column masses as D_c . The matrices D_r and D_c are shown as diagonal matrices with 0s off-diagonal. The formulas $P = D_r^{-1/2} U D$ and $Q = D_c^{-1/2} V D$ are written below the diagram.



D_r is basically when you are making a contingency table, this side is row mass, this is column mass for our case here, there are eight and four. So, four values here eight values here, for row mass, there are eight values. So, our D_r if we, we are creating like this, all these values, these are all diagonal values and of diagonal are put 0. So, D_r is 4 cross 4

matrix. Where from you get these values? These values you will be getting from contingency table when it is ultimately made to converted to correspondence table or correspondence matrix. Similarly, Dc will be 8 cross 8 matrix where the diagonal element will be this, this value and of diagonal element will be 0 again.


So, so you know now, now know that what is Dr half, Dr clear, Dr to the power minus half. So, you are making inverse of Dr and then square root of each of the elements, if you do this, then the principal component for the for the rows for the rows, this will be Dr half U D. Already, you have computed P. You know this is that. Similarly, the principal component for the column is, if you see Q that is Dc half V D. So, you are in a position to compute the principal components. See the duty of this correspondence analysis it started with nominal data, which are cross classified.

You are getting a frequency table and you do not know that from frequency data, how to get the distance similarity, all those things, difficult one and then what is the, intelligently what it is that the coordinates are created? Using weighted chi square distance, distance and singular value decomposition method, you are able to extract different PCs, principal component. This is a wonderful development in that count.

(Refer Slide Time: 22:46)

Principal Coordinates

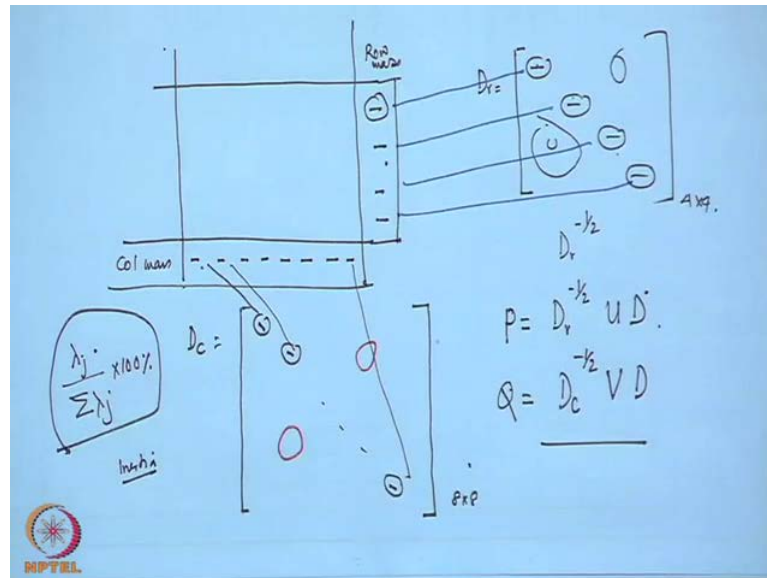
	PC1	PC2
V1	-.348	.324
V2	-.225	-.183
V3	-.293	-.163
V4	.794	.026
BD	-.718	.012
DB	-.579	-.028
DI	.199	.168
FI	.849	-.184
HM	-.080	-.456
PL	-.310	-.055
SC	.231	.264
WL	.547	-.225
Inertia (%)	95.600	3.000



Now, once PC is developed, for example, in the table, we have given you the contingency table for vendor versus category of defect. You see we have extracted two PCs here, but the k is minimum of p minus 1 and q minus 1, which is 3. So, one more PC

can be created here, but we are not created, we have created, only two PCs and you see the inertia, inertia. What is inertia here? The inertia is similar to variance explained in principal component analysis. What we have talked about? We have talked about the lambda value.

(Refer Slide Time: 23:31)

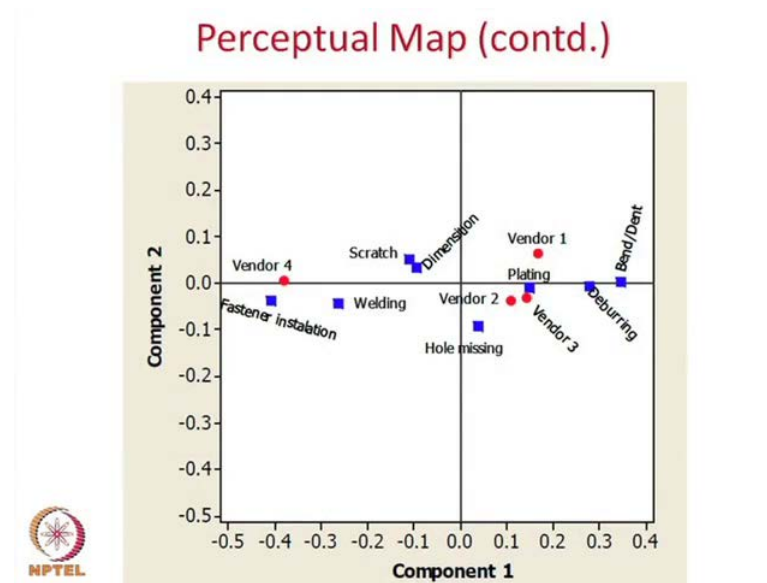


Lambda j divided by sum total of lambda j, this is into 100 percent, and this is the percentage of variance of the x variable explained by that jth PC here. We are using it here. We use the word inertia. Inertia talks about that what is the percentage of variability in the data set is explained by principal component one. So, inertia for PC1 is 95.6 percentage, for PC2, it is 3 percentage. The rest will be that total will be 100 percent. So, rest will be for PC3.

Now, if I go by that cumulative percentage explained by PC1 that is 95.6 percent and if I include PC2, this will be 98.6 percent, but 95.6 percent itself is big enough. That means we require only PC1 to explain the, explain the coordinates related to V1, WL that is basically the variability what is present in the contingency table. We do not require PC2, PC3.

We have been essentially, what happened here essentially, it has happened that although there are four vendors and eight different categories, different dimensions ,now ultimately using one dimension that is PC principal component 1, we are able to explain the variability present there in, it is a beauty.

(Refer Slide Time: 25:30)



Now, let us see in some more slides where we are talking about the perceptual map. You see the position of each of the vendor and category of defect. You are able to see that this horizontal line, this is the principal component 1 and this line is the principal component 1 and this is component 1, that is why component 2. So, although we are plotting using this bi variate type of hybrid plots, but if you see the position of each of the categories, the row as well as column variable, you see that all of them are very very close to principal component 1.

So, if I if I go by that additional way of understanding conventional way of understanding that mean along the principal component 1, all the points lie and the variability from here to here, so this is the that total variability. But, if we consider the component 2, then you see that there is hardly any variability along this component 2. If I consider these are all data points, so that is what is justified by this 2 percent inertia. That means 3 percent variability of the total data set in this case vendor versus that categories of defect table. So, that 3 percent variability is explained by PC2. We do not require. We we can only plot component 1 and series.


So, PC1 is able to explain the totality of this particular contingency table. Now, how do you interpret this, that what is that PC1, what is the name of this PC1. So, by interpretation, we want to mean that it is similar to factor analysis. You are getting a factor and you have to have the ability to name the factor. Unless you name the factor,

you cannot think that you have done the job correctly because you have to take decisions for improvement for betterment. So, I have a hidden dimension. If I do not know what is this hidden dimension, then I cannot improve on the basis of this result or on the basis of hidden dimensions, characteristics, but I do not know what is that dimensions.

(Refer Slide Time: 28:31)

Interpretation of the map

- Vendor 4 position is far away from other vendors
- Bend/dent and deburring defects are far away from fastener installation and welding
- Vendor 4 has more problems with fastener installation and welding defects
- Vendors 1, 2 and 3 have relatively more defects related to bend/dent and deburring.
- The points close to the centre of the display have no/negligible differences.




Let us see how we can do this.

(Refer Slide Time: 28:36)

Overall Fit Measures (contd.)

Name	Quality	Mass	Inertia	Component 1 ($\lambda_1 = 0.053$; 95.56%)		
				Coordinate	Correlation	Contribution
Vendor 1	1.000	0.242	0.140	0.167	0.868	0.127
Vendor 2	0.895	0.257	0.068	0.108	0.802	0.057
Vendor 3	0.919	0.235	0.096	0.141	0.872	0.088
Vendor 4	1.000	0.266	0.696	-0.381	1.000	0.728
Bend/dent	0.992	0.111	0.240	0.344	0.992	0.250
Deburring	1.000	0.107	0.149	0.278	0.999	0.155
Dimension	0.959	0.196	0.038	-0.096	0.852	0.034
Fastener installation	0.994	0.132	0.400	-0.407	0.986	0.412
Hole missing	0.881	0.080	0.016	0.039	0.132	0.002
Platting	1.000	0.179	0.072	0.149	0.994	0.075
Scratch	0.913	0.170	0.051	-0.111	0.742	0.040
Welding	0.930	0.025	0.034	-0.262	0.903	0.033



So, I will straight away go to some of the fit measures. Then I will come back again for the interpretation. See in correspondence analysis, inertia is an important measure. Now,

what we say inertia means that is the percentage of variance explained. When we talk about inertia of a component that means the component extracted, what is the percentage of variance the component is able to explain. When we talk about inertia for each of the categories that means it is basically the variance or a yes variance part that contributed by each of the categories to the total, your variance.

So, in that count, this is what and mass you know that is the total that row mass, column mass you have seen and quality is basically the quality of presentation in the perceptual map. But, more important thing is this correlation value, this correlation value because this correlation value talks about how good this component 1, how good this component 1 to explain, explain the each of the variable categories, categories, contribution towards explaining the variability or the variance. So, you see that it is basically the component 1 is able to able to explain the variability along these because 0.8860, 0.802, 0.872, 1 that means component 1 is enough to capture all the vendor categories variability.

Similarly, if you see these column categories except hole missing, all other cases you see the value is quite high. You go to contribution, these contributions also, then another one is the contribution, first one is basically component, how it is able to explain these things. By contribution, what we are saying how this individual categories are, what is the contribution of individual categories in, in building this component 1? Correlation means how this component 1 is able to able to explain or correlate each of the categories with its, with this dimensions. Coordinate is nothing but that these coordinate values what you have gone for, principal axis, that principal component, principal component, that coordinate positions.

So, using this contribution and correlation, you will be able to explain the variable categories with respect to the particular principal component. With this, let me come to this some interpretation. Vendor 4 position is far away from other vendors. You see vendor 4 is here, other vendors are here. So, that means I can say vendor 4 is distinctly different from other vendors. There is something different in vendor 4. Other vendors are very close with respect to categories, observations of defects. Bend dent and deburring defects are far away from fastener installation and welding. Bend dent, fastener installation is here, whereas bend dent is here. They are distance apart.

Vendor 4 has more problems with fastener installation and welding defects. How we can say because vendor 4, they are close and welding vendor 4 so on. But, vendor 4 versus these is far away. So, relatively we can say vendor 4 has this type of defects problem relatively more, not absolutely. Vendor 1, 2 and 3 have relatively more defects related to bend dent and deburring. The points close to the center of the distant no and significant differences, those which are coming to this, they are difference because they are; they are overlapping almost on the average profile. So, there are little differences. Now, if we see the correlation, we are able to find out that each of the, each of the categories are being explained by principal component 1 except one category.

(Refer Slide Time: 33:49)

Overall Fit Measures

- Total inertia explained by each of the components extracted
- Total inertia is defined as the “weighted sum of squared distances from the points to their respective centroids”
- Inertia explained is similar to explaining the percentage variation by R^2 in regression
- The number of components to be extracted is decided based on cumulative percentage of inertia explained (e.g., $\geq 90\%$).



For the case study, 98% of the total inertia is explained two PCs, while PC1 alone explains 95.56% of the total inertia

Now, let us see that as I told you that overall fit measures, total inertia explained by each of the components extracted, this is one measure. Total inertia is defined as the weighted sum of squared distances from the points to their respective centroids. So, you have seen suppose one point here, then that this is the point and then the centroid is this one. Now, what we are saying that weighted sum of the squared distance from the points to their respective centroids. Inertia explained is similar to explaining the percentage variation by R square in regression.

The number of components to be extracted is decided based on cumulative percentage of inertia explained similar to principal component analysis. 90 percent for the case study, 98 percent of the total inertia is explained by two PCs, while PC1 alone explains 95.56


percent of the total inertia. Inertia accounted for by each of the vendors as well as categories of defect that also I have explained contribution of vendors and categories of defect to the PC extracted. That is what is contribution. Now, inertia of a point each of the vendors or categories is explained by PC. That is what is correlation.

So, this correlation is very, very important measure. What is correlation inertia of a point explained by PC? Suppose this is vendor 1. Its inertia is 0.140. Now, what is the ability of component 1 to explain? That is 86.8 percent of this inertia is explained by principal component 1. So, that correlation is very big issue for naming the component and then what is a contribution? Contribution of vendors and categories of defect PCs extracted, that means what you are saying this is, what is talking about when you extract PC1, how this categories are contributing in extracting this? This is fine, but another thing is this one the categories, one the principal components are extracted. So, after that, how this principal component is able to explain the variability of each of the categories; that is correlation. Got it? So, we have to, I am coming to the interpretation, interpretation of the principal component.

(Refer Slide Time: 36:42)

Naming the components

- Component 1 (PC1) explains 95.56% of the total variance accounted for vendors and categories of defects
- PC1 is defined by vendor 4 (contribution = 0.728), fastener installation (contribution = 0.412) and bend/dent (contribution = 0.250) types of defects
- PC1 explains 100% of the variation of vendor 4 and more than 80% for each of the vendors 1, 2 and 3
- Except 'hole missing', PC1 explains other categories of defects reasonably well



So, let us see component 1 explain 95.56 percent of the total variances accounted for the vendors and categories of defects. Here, we have used the two things, contribution as well as your correlation into consideration. PC1 is defined by vendor 4, fastener installation bend and dent based on these are the contribution, but if you see the

correlation, see vendor 4, correlation is 100, bend dent 0.99, deburring 0.99, and fastener installation 0.998, again plating 0.99. So, these are the, these are the variable, variable categories which are ultimately explaining this, naming this.

(Refer Slide Time: 37:41)

Naming the components (Contd.)

- Naming a PC extracted calls for the analysts' knowledge about the functioning of the system under consideration as well as variables interaction effects
- Discussion with plant personnel reveals that
 - Vendor 4 uses all manual transfer installation machines whereas other vendors (1, 2 and 3) use a mix of automatic and manual machines
- **PC1 can be named as 'level of mechanization'**



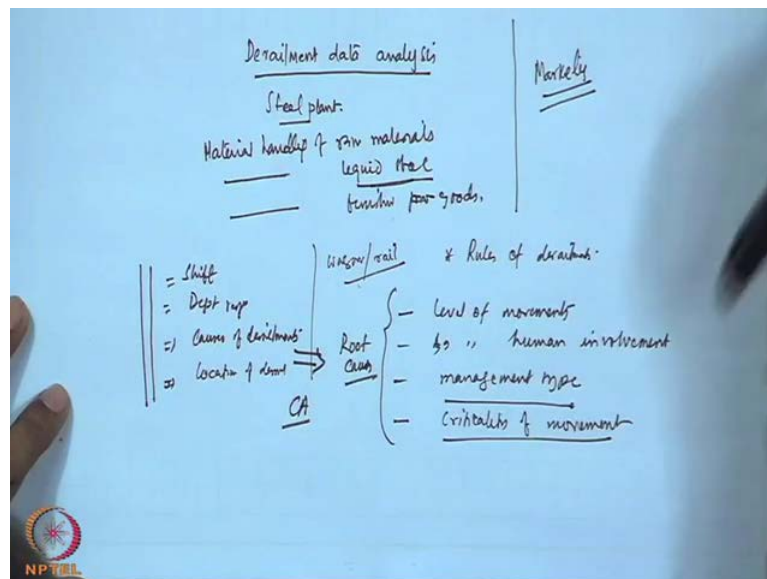
So, with this, what I can say? Ultimately, we have gone to the, we have gone to the plant where from we have collected the data and we have discussed that what could be this principal component 1. See principal component 1 is fine here, but naming the principal component 1 or as such component 2 or 3, it requires domain knowledge. So, that domain knowledge, the practitioners, the expert in this particular subject here will help you to to understand what this component is talking about.

So, naming a PC extracted calls for the analysts' knowledge about the functioning of the system under consideration as well as variables interaction effects. Discussion with plant personnel reveals that vendor 4 uses all manual transfer installation machines, whereas other vendors use a mix of automatic and manual machines. When we talk about the defects, the experts say that it is this difference, which is making this type of realization. As a result what happened? We have given the name; PC1 can be named as level of mechanization.

So, if this is this axis is level of mechanization, so from mechanization point of view, vendor 4 is manual, whereas these are semi mechanized or to some extent, some are semi mechanized. So, as a result, this distance is happening here, this when as it is manual,

that fastener installation, welding, scratch, all those things are also coming into consideration. But, you require exploring it further that whether some other explanation possible or not, but to some extent, it is basically the same thing that a level of mechanization can be given a name that to the PC extracted here, here, it is it is certified by the company people. Getting me? So, this is one use of correspondence analysis we have shown you. I will show you another use, another use of correspondence analysis what I have done; that one is a derailment data analysis.

(Refer Slide Time: 40:26)



So, we have considered a case of steel plant, steel plant and the steel plant has implant, railway system implant within plant for their material handling of the raw material, material handling of raw material, then material handling of liquid steel material, handling of finished products, finished goods. So, what we have done ultimately? We have considered that shift, then the department responsible, then the causes of derailment and one more variable we have considered, four variables department, shift, causes of derailment and line where the derailment taking location of derailment, location of derailment. Now, derailment is basically, these are wagon or rail derailment, wagon derailment that are derail that in house implant railway track systems.

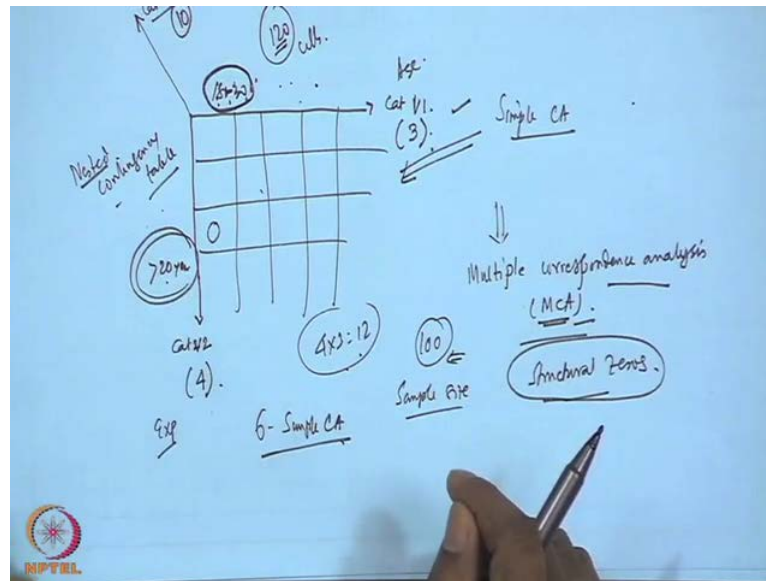
So, what will happen, even again, there are four different variables or six different contingency tables created. Then finally, we have done the same way, similar way analysis. Ultimately, what we have found out that we have created four rules, several

rules, rules of derailment. After extracting the PCs, then when we have given the name of PCs, then we found out from by naming the PCs, several things. One is level of movement that is very, very important, but which was not nowhere it is there because they are not also capturing the level of movement data that how frequently, for duration, how long this movement is taking place.

This type of data, they are not capturing, but from CA, we found out this one. So, level of human involvement. Then we found out that that basically the management, management types, another dimension, that mean whether it is actually in the mainly the raw material and liquid steel material handling is managed by the implant that company, but finished goods that railway tracking all systems are managed by the external company.

So, externally managed wagons and internally managed wagons that derailment pattern is different, then that criticality of movement, criticality of movement. So, we have started with this four categorical variables. Ultimately, will ended up to this root causes. So, this transformation was possible only through CA, but to come to this stage, we require lot of domain knowledge, otherwise you cannot. Getting me? So, you will be, if you go through marketing research, we will be getting a new innumerable, innumerable examples related to use of correspondence or application of correspondence analysis and its use in engineering studies is also available, but not that in that much. In social science, its huge use application is available.

(Refer Slide Time: 44:58)



So far, we have talked about cross table or contingency table of two variables, obviously categorical variables, so categorical variable 1, categorical variable 2, v_1 , v_2 . What will happen if there is one more categorical variable? Actually, that is all the cases in our in our, my example, all the cases, there are more than two variables, the either derailment or the defate that is occurring with based on the influence of all of the categorical variables, not necessarily that two at a time. So, if this is the case, then just going through only bi variate case, two that two categorical variable taking two categorical variables and doing simple correspondence analysis, simple CA will not serve the purpose fully intended for.

In this state, simple CA is good. It will give you some good idea. Most of the time what will happen because this CA is used to find out the root causes, so if you if you develop several contingency tables and do this, you will be finding out wall line root causes or overlapping root causes. So, then the final set will be distinct causes like this, but why should I go for this? Why should not I consider all the categories, categorical variables simultaneously? Then what will happen? You will be creating nested contingency tables, correct nested contingency table.

So, if you include more than two categorical variables in correspondence analysis, simple CA will not work. We have to go for multiple correspondence analyses. What will happen then? This multiple correspondence analysis, consider all the categorical variables with their categories and there will be nested, this table and that will have large

number of cells. For example, if category v1 has three categories, categorical v1 variable has three categories, v2 variable have four categories, and then it is 4 into 3, so 12 cells. But, again if v3 is having 10 categories, then 12 into 10, 120 we will get.

If your data size sample size is small, suppose you you have observed hundred cases, then you do not get 120 cells having cell counts having frequency. So, the more variable you add, more number of cells will be generated. So, more observations are needed. This is one aspect which you have to look into while going for multiple correspondence analyses, otherwise what will happen? You will report values to software, software will give something, but there will be erroneous results.

Another difficulty while going through multiple correspondence analysis is suppose the categories are such that there will be some structural zeros like if variable category one is age and this one is experience. Suppose if I say that 18 to 30 is the first category of age and like this and experience. Let this is greater than 20 years. So, the person having this much of age will not have this much of experience.

So, as a result, the the cell 18, 30 and greater than years of 18, 30 years of age and greater than 20 years experience will always have zero counts. This is what is known as structural zeros. So, these types of things are there. What will happen ultimately? Ultimately, if it is there, you have to look into these things very carefully structural zeros.


Another issue related to your correspondence analysis, which you have to keep in mind also that correspondence analysis tries to measure not only the difference or similarity between the row or column categories, but also association between them. So, if row and column categories are independent, row variable is not dependent on column variables, so association we will not find out. So, that when if there are four variables, 6 simple CA possible, but what is the need of doing 6 simple CA, if some of the CA that contingency table has no relationships between row and column variables, you can avoid this.

(Refer Slide Time: 51:14)

Should we go for CA?

Test of independence

Model No	Variables considered	χ^2 computed	χ^2 tabulated(d.o.f) $\alpha = 0.05$	Remarks
M1	Vendor and category of defects	246.14	32.67 (21)	Row and column categories are dependent
M2	Vendor and time periods	33.84	36.42 (24)	Independent
M3	Vendor and part number	6.76	16.92 (9)	Independent
M4	Part number and category of defects	867.00	32.67 (21)	Row and column categories are dependent
M5	Part number and time periods	89.70	36.40 (24)	Row and column categories are dependent
M6	Time period and category of defects	129.01	74.46 (56)	Row and column categories are dependent

 **M1 indicates that CA will give better insights**

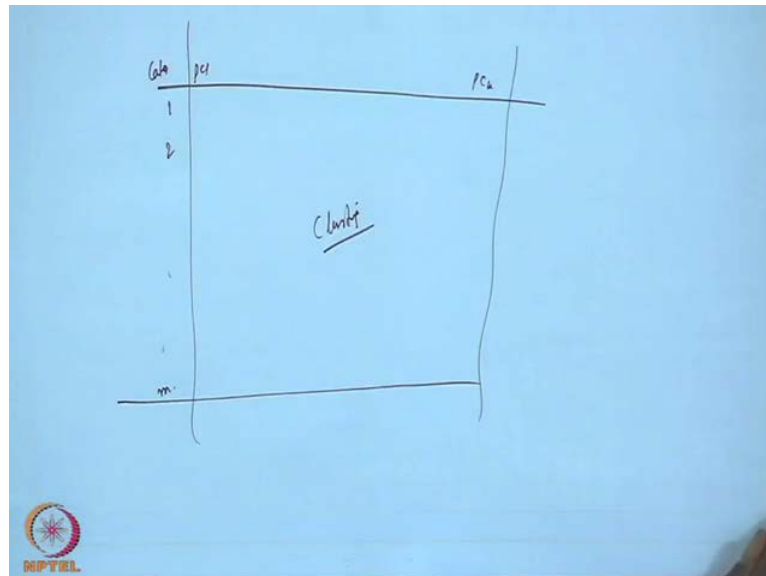
For example, we have done here like this. There are six models possible, but using chi squared test, what we have found out that row and first one vendor and category of defect row and column categories are dependent. Vendor time period, vendor part number, they are independent. So, no need of going for M2 and M3, no correspondence analysis is required where there is dependency. Then you go for correspondence analysis.

That is why we said M1 indicate that CA will give better insights, not M2, M3, this chi squared value they are very low, so compared to M2 and M1. M1 definitely gives better insights. So, when there is there is that dependency is available then only you go for M1. So, another issue, suppose if you use mat-labs many a times that straight away this SVD function is applied, you will get that this there will be a difficulty k is minimum of this, but you will get even minimum of this that if p minus 1 is the minimum, then even pth some value also you will get what I have seen a while using this. This is because of the all the rounding errors effects calculation computational problem.

Then, another issue here is that suppose many a times when you go for multiple correspondence analyses, so you will find out that the two principal components cannot explain the variability, desired variability to be explained. This will not happen. I have run I have seen that two components are able to explain around 24, 25 or 30 percentage of variability. So, that means you require more components. Then ultimately your

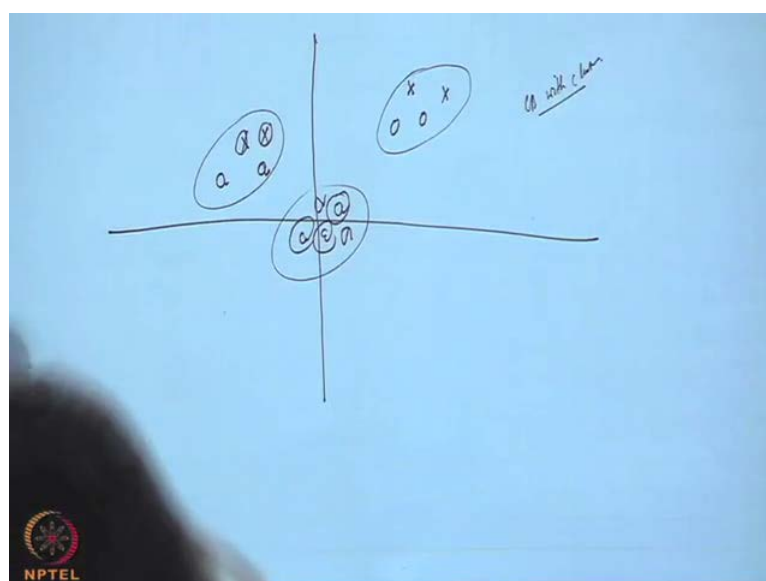
purpose is not served. So, in that case, what to do? Can we just stop there or there is some other way of doing things? I think I have seen some papers where they said.

(Refer Slide Time: 53:50)



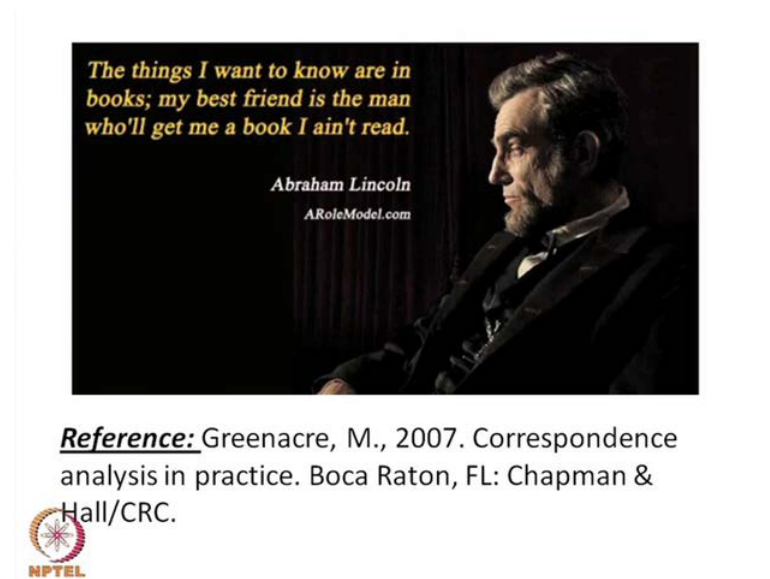
You extract as many PCs as possible, PC1 to suppose PCk. Then you can go for; you can go for clustering using these PC values. These are all categories one two, there will be suppose m categories, then for all m categories here, PC, PC1 coordinates are there, PC2 coordinates are there, PCk coordinates. Sorry, extremely sorry.

(Refer Slide Time: 54:27)



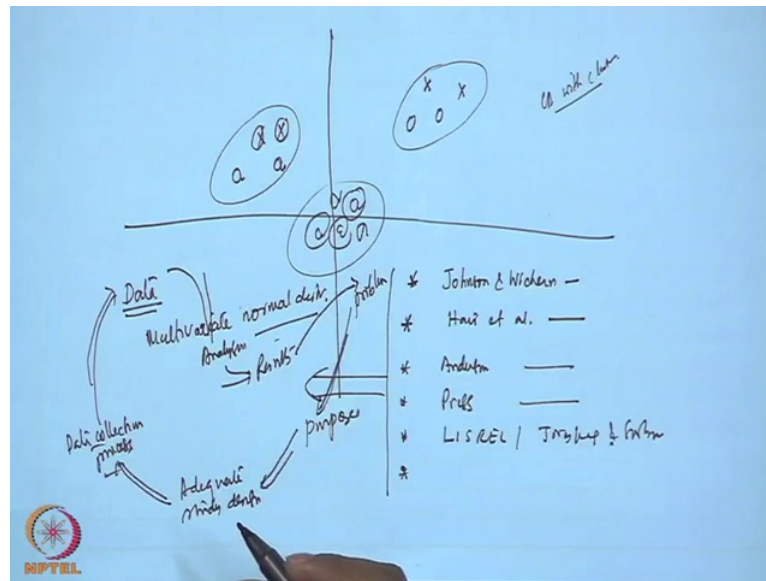
So, in that case, what happened if I go for clustering? You may find out that values may be clustered. So, clustering may help us to find out, this is one cluster, this is another cluster, and this is another cluster. So, you will be able to group them. So, what I am trying to say? Then that why cannot, we go for CA with clustering, it is it is available also.

(Refer Slide Time: 55:30)



Similarly, other variables can be now known. I will I will finish today's lecture with this reference as well as the quote. You see the reference for correspondence analysis Greenacre, M, 2007, correspondence analysis in practice. This is a very good book on correspondence analysis and this is a very famous quote from Abraham Lincoln. The things I want to know are in books; my best friend is the man who will get me a book I ain't read. When that is what is also true, very much true because it is not because of it is said by Abraham Lincoln, whether he is a famous great personality, whatever they will say definitely that has value. But, we being ordinary people, when we when we realize this and do something, we will be immensely benefited by this great person like this. I have seen because that I have to find out first this book and I have gone through this book. So similarly, for all the topics covered in this multivariate statistical modeling course, let me tell you very important things here.

(Refer Slide Time: 56:51)



Your work will be simple if you collect the good books. I suggest Johnson and Wichern book, Johnson and Wchern multivariate data analysis. There is wonderful, another wonderful book by Hair et al that is for the research scholars. For the for the concept building, this book is very, very good. There is book by Anderson also for multivariate. Then Press suppose for structural equation modeling and all those things that Lisrel software and Joreskog and Sorbom, you will get that correspondence analysis greenacre. You will get couple of good books on factor analysis also; your work is first find out the good books.

Another important issue is data. So, theory is one thing fine, but if you do not apply these theories, whatever you are learning with real data, your knowledge will not be 100 percent perfect. The reason is when I applied to the models in the real data, what we find out that surprisingly my expectation meets only 50 percent only less number of, less amount of time.

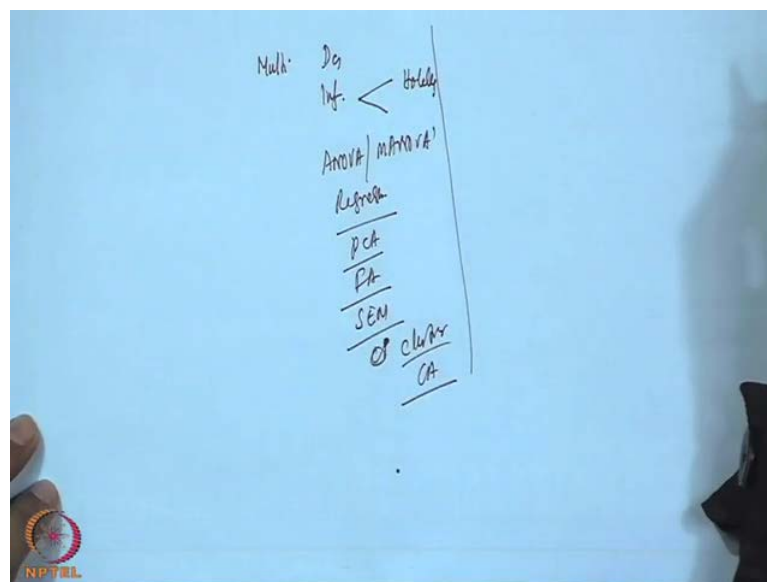
Most of the times, we are finding out that what we thought of and what we are getting from the analysis has not matching. This is because most of the models are developed based on several assumptions. One of the big assumptions is multivariate, multivariate normal distribution, but how do we know that it multivariate normal or not many a times this normality assumptions are violated. But, because we know how to use the software without bothering about all those things, we use and we get some results. Sometimes,

they are favoring, sometimes they are not. When they are favoring, we are using, but maybe we are using wrongly that favor is a wrong favor.

So, I say that one is data and thus another important issue is this data. How you are collecting the data, data collection, data collection process in the first class, I told about this different study data collection process should be governed through a, through adequate study design. You must know what is your purpose? So, the study design will be governed by purpose and your purpose definitely will be governed by the problem, what problem at hand.

So, your data, then data analysis, then you are getting the results, but this result must ultimately solve the problem. Getting me? So, these issues, these are the very, very key issues; basically one issue is all the models. We have basically come across several models starting from basic statistics under this descriptive statistics, then your inferential statistics, multivariate statistics.

(Refer Slide Time: 01:00:34)



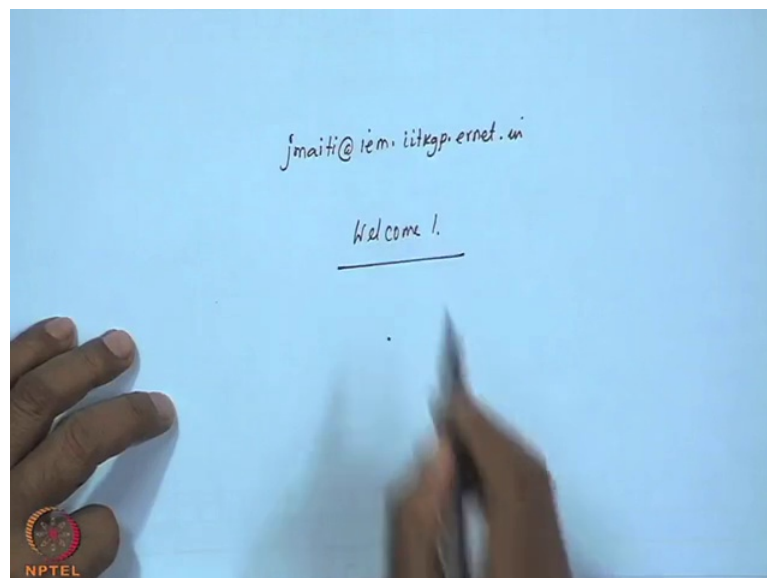
Multivariate, descriptive statistics, descriptive statistics, inferential statistics; under inferential, you learnt hotelling t square for different group categories. Then you have gone through ANOVA. Then you have gone through MANOVA.

You have gone through regression. You have gone through PCA. You have gone through factor analysis. You have introduced structural equation modeling. You have seen cluster analysis. You have done CA, so many things. But, essentially whatever we have discussed in this course, they are very much introductory in nature. Every topic has inbuilt in depth mathematics and if you all are interested to know the in depth mathematics behind it, then you have to go through further good books.

For example, johnson and wichern is a good book, but the how the parameter estimated, what is the optimization techniques applied, what is the method applied or when we are using this, some other methods, maximum likelihood method, all those things, so those derivations and those integrities, some numerical methods, those things also are very very important in understanding in totality. The data analysis or data modeling steps, this is also required.

So, that nevertheless this lecture is basically an introduction to multivariate statistical modeling for the persons who are basically in the application side; preferably engineers and scientists who have data want to analyze data in the multivariate domain. Thank you very much. For any query, whatsoever, this is my email address.

(Refer Slide Time: 01:02:46)



jmaiti at iem dot iitkgp dot ernet dot in. Welcome.

Thank you very much.