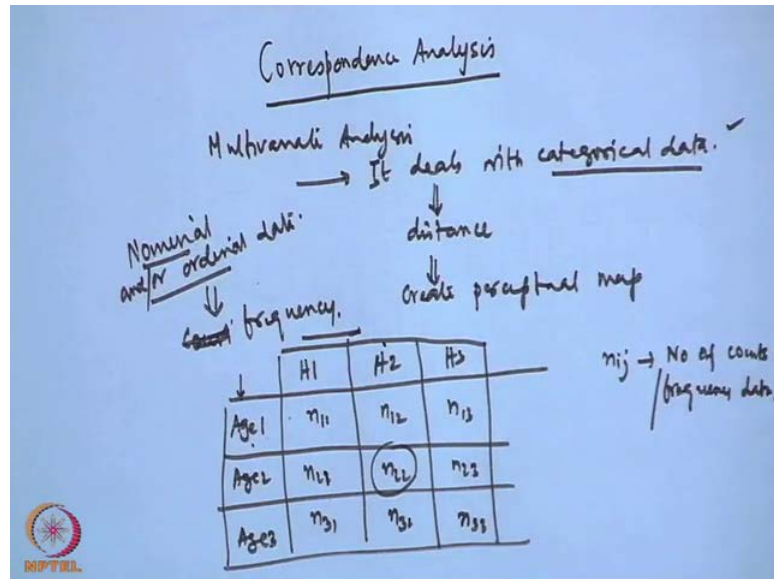


Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 41
Correspondence Analysis

(Refer Slide Time: 00:25)




Good morning. Today, we will discuss correspondence analysis. Correspondence analysis is a multivariate analysis, multivariate analysis. It has two components; one - it deals with categorical data, it deals with categorical data and then it converts this data to certain distance, distance. Using those distance values, it creates perceptual map, create perceptual map. So, the one of the advantage of correspondence analysis is that that it can it can work with categorical data. By categorical data we mean the nominal or ordinal data, nominal and or ordinal data.

Actually, in correspondence analysis, the data are observed in terms of counts or frequency, in terms of frequency. For example, let there are three age group of people, age 1, age 2 and age 3. Now, we can see their food habits, food habits, habit 1, habit 2 and habit 3. Now, if you see a particular population and we may find that in general, this age group with this food habit there may be n_{11} , n_{12} , n_{13} , n_{21} , n_{22} , n_{23} , n_{31} , n_{32} , n_{33} . Now, this n_{ij} , this is number of observations or I can say counts or say that frequency that is frequency data.


The question comes, that is there any relationship between these frequencies observed with the variable called age, which is categorized into three groups. Is there any relationship between the observations here, for example this with the food habits? Is there relationship between age and food habit in realizing this frequency counts. So, all those things can be captured through correspondence analysis.

(Refer Slide Time: 03:46)

Introduction



- Correspondence analysis (CA) is a geometric approach to multivariate descriptive data analysis
- Developed by the French linguist and data analyst **Jean-Paul Benzecri** and his colleagues in 1960s




Now, if we go back to the history of correspondence analysis, we will find that correspondence analysis was developed by Jean Paul Benzecri and his colleagues in 1960s. This is a geometric approach on multivariate descriptive data analysis. So, the way Jean Paul Benzecri and his team developed the correspondence analysis that same approach we will follow here.

(Refer Slide Time: 04:19)

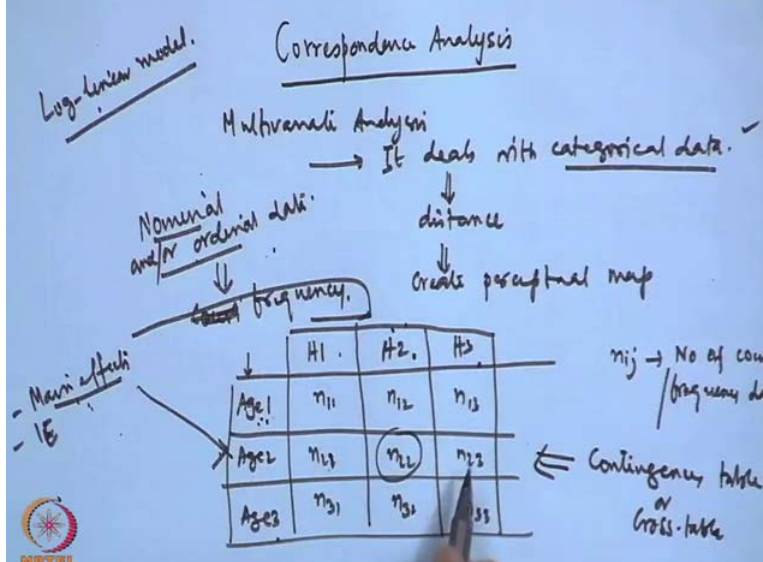
Introduction

- CA is a dimension reduction technique similar to factor analysis but extends factor analysis in two counts:
 - handling of categorical variables, particularly measured in nominal scale
 - developing perceptual maps of extracted components



So, let us see that what a correspondence analysis can do. Correspondence analysis is a dimension reduction technique similar to factor analysis, but extends factor analysis in two counts; handling of categorical variable, particularly measured in nominal scale and developing perceptual maps of extracted components. So, what I mean to say here?

(Refer Slide Time: 04:45)



The diagram illustrates the Correspondence Analysis process. It starts with 'Log-linear model' and 'Multivariate Analysis'. 'Multivariate Analysis' leads to 'It deals with categorical data', which then leads to 'distance' and 'create perceptual map'. 'Nominal and/or ordinal data' leads to 'frequency', which leads to a contingency table. The contingency table is a 3x3 grid with rows labeled 'Age1', 'Age2', 'Age3' and columns labeled 'H1', 'H2', 'H3'. The cells contain frequencies n_{11} , n_{12} , n_{13} , n_{21} , n_{22} , n_{23} , n_{31} , n_{32} , and n_{33} . The cell n_{22} is circled. A note on the right says 'n_{ij} → No of counts / frequency data'. A note on the bottom right says 'Contingency table or cross-table'. On the left, there are notes: '- Main effect' and '- IE'.

	H1	H2	H3
Age1	n_{11}	n_{12}	n_{13}
Age2	n_{21}	n_{22}	n_{23}
Age3	n_{31}	n_{32}	n_{33}

So, what I mean to say here is that you require developing a contingency table. Contingency table or cross table, both are same. When you say at contingency table or cross table, this is basically bivariate table and here you see although age is a continuous

variable, but age is made as it is a nominal variable by categorizing into three age groups, but food habit is definitely a nominal variable. So, if you want to use correspondence analysis, you have to have the variables of interest in nominal scale.

Now, if you, if you use ordinal data also, it will be treated as nominal data. If you use continuous data, then you convert them to nominal data. The sole purpose is to understand whether there is correspondence between between the observed frequencies and the variables considered either independently or jointly. It is similar to in case of ANOVA. We have seen that there are two things in ANOVA. One is main effects and interaction effects. Now, main effects are related to the factors like your age and food habits, interaction effect between the factors in similar way, but in MANOVA.

In ANOVA, what is required? It is required, the data, the response variables will be continuous in nature here it is different and correspondence analysis has also relations similarity with log linear model. In log linear model, we also try to find out main effects and interaction effects using certain parameter, but there definitely are a lot of differences between all the correspondence analyses between these models.

First of all, correspondence analysis uses frequency data. Log linear model also uses frequency data, but log linear model will not go for dimension reduction. Correspondence analysis will go for dimension reduction. Also, it is one way a principal component analysis or similar to factor factor analysis. Like in another way, it is having also a capability to identify the relationship between the categories across variables, categories between variables. It also has the potential to create map. By seeing the map, we are in a position to talk about the association along the categories of the variables and between the categories of two variables; the associations can be carefully interpreted.

(Refer Slide Time: 08:03)

Introduction (contd.)

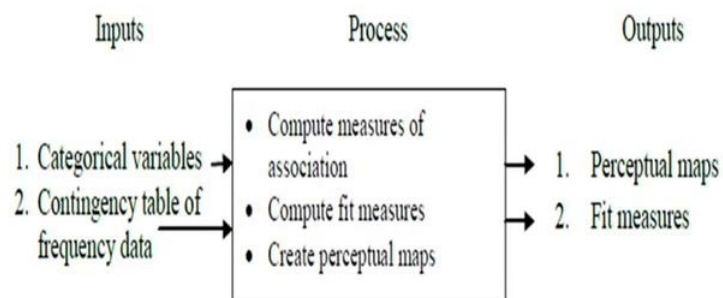
- While factor analysis captures linear relationships, CA captures non-linearity between the variables represented in contingency tables
- The other names of CA are:
 - Dual scaling
 - Method of reciprocal averages
 - Optimal scaling
 - Canonical analysis of contingency tables
 - Categorical discriminant analysis
 - Homogeneity analysis
 - Quantification of qualitative data



Now, see, what are the different names we find in the literature related to correspondence analysis? Dual scaling, method of reciprocal averages, optimal scaling, canonical analysis of contingency tables, categorical discriminant analysis, homogeneity analysis, quantification of qualitative data, what is in, what is mean and seen also that in these different names and that similar to correspondence analysis are, in fact the correspondence analysis is used in the literature and also published. Now, correspondence analysis another important issue here is that it captures the non linearity between the variables represented in the contingency table.

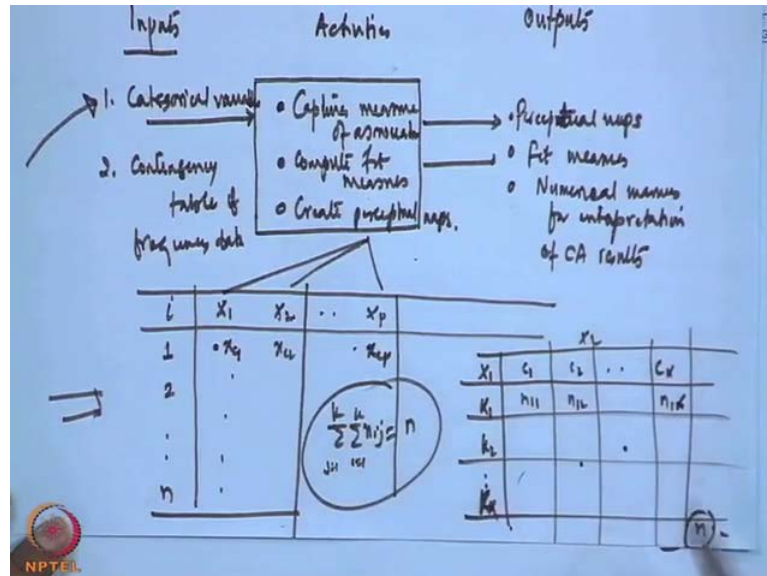
(Refer Slide Time: 09:07)

CA Process



So, let us see that how correspondence analysis works. Now, this one, we will be representing in terms of correspondence analysis process.

(Refer Slide Time: 09:23)



That process model some is inputs, then activities, then outputs. Inputs means what correspondence analysis requires. So, if you see that it requires some certain number of variables, which are categorical variables, then you have to create a contingency table, contingency table of frequency data frequency data. Then what it will do? It will do three things; captures measure of the association, captures measure of association, then compute feet measures and then create perceptual maps.

So, what are the outputs? Then there are some one is the perceptual maps, perceptual maps. Then second output is different feet measures and actually we will get many numerical numerical measures, numerical measures for interpretation for interpretation of ca results. So, that means if you want to use correspondence analysis for some purposes, what you have to do? You have to first find out what are the variables of interest, list those variables, collect data, then what will be the data structure? Data structure will be similar to like this i 1 to n number of observations. Now, your categorical variable 1, categorical variable 2, so like categorical variable p , then what is required? What will be required?

Suppose observation 1, it may fall with one categorical variable with some category some category. Suppose it is basically x_c 1 category, x_c 2 category x_c p category,

similarly different categories. So, when there are more than two variables with multiple categories, then you require multiple correspondence analyses to include all, all the categorical variables. But, if we consider only two variables like this x_1, x_2 with this category, then you will be able to develop suppose you write x_1 here and this side x_2 . You know it has category, category 1 category 2 to suppose category k and it has category, category 1, category 2, suppose category kk .

Then you will finally, get a frequency table like this n_{11}, n_{12}, n_{1n} , I think it is k . So, it will be k , a similar table will be, so that means your data will be coming like this. Then I will show you one example and and from this data, you take two variables at a time with them. These are the categories of the variables, related variables. Using this and n , we will be able to find out this correspondence table or contingency table. We will be able to find out this contingency table.

Now, a sum total of all these n values that will be capital N because we have considered n number of observations. This is small n , sorry not capital N , small n , we are measuring this small n , so n_{11}, n_{12} , what I do? I mean I mean to say that sum summation i equal to suppose if I say this is k, j equal to 11 to capital K n_{ij} , this will be n sum total of all will be n . So, this is our data. Now, from this data this way, you are getting the contingency table or correspondence table. If you have more than two variables, two categorical variables, you have to go for multiple correspondence or contingency table and or we can say the nested contingency tables.

Then, what we are trying to say that what are the activities that will be performed? You have to first find out the measure of association. How do you find out this measure of association, this is an issue and we will discuss next. Then once you find out the association measure, and then you will you will, using those values, you do some action and then some results you get. Based on these results, you find out the feet measures and finally, maps you create.

(Refer Slide Time: 15:33)

A Case Study

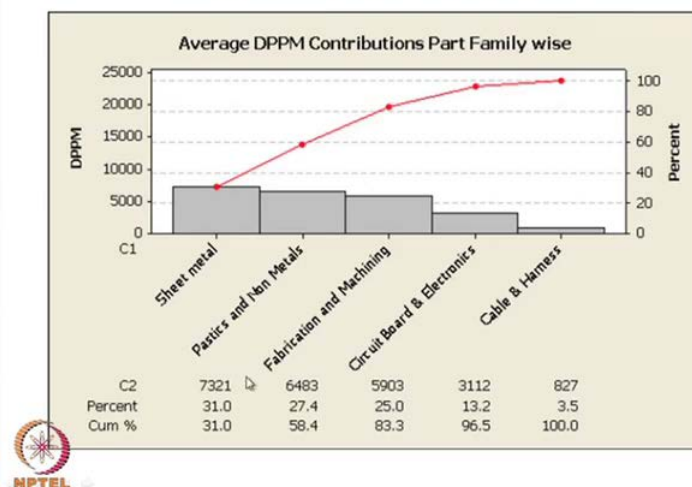
- An organization which manufactures diagnostic medical equipment
- The organization does assembly of final device and has outsourced all the components required for the final assembly



So, let us start with a case study because it will help us to understand ca, an organization which manufactures diagnostic medical equipment. This is our case. The organization does assembly of final device and has outsourced all the components required for the final assembly. This one is one of our case study one paper; that is reference is Aravindan S and Maiti 2012. A framework for integrated analysis of quality defects in supply chain, ASQ quality management journal, volume 19, issue 1, page number 34 to 52. Now, let us see that what is this case study in totality?

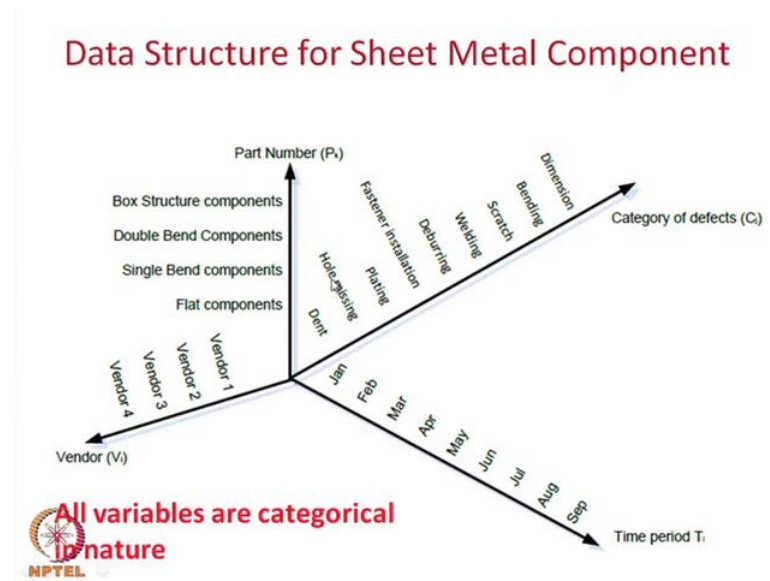
(Refer Slide Time: 16:27)

Case Study (contd.)



That we found out that that compare, if you see the what are the different components that sheet metal component, plastic and non metal component, fabrication and machining, circuit board and electronics, cables and hamess, these are the different components where ultimately defects per million, defective parts per million that is captured from six sigma point of view. What we found out that the sheet metal component has been the highest, highest defects. So, it is a case. So, we can say this is the sheet metal case we have to consider for improvement.

(Refer Slide Time: 17:20)



Then, when we capture the defect data, unrelated data with respect to the sheet metal component, what you find that there are four things, four categorical variables, which can be used to capture the that data set totally in totality. By this, I am not saying that there are no other variables which are contributing to the sheet metal defects. It is there, but in for the present study what has happened with discussion with the personnel, who are basically working there in this particular company. And who are dealing day to day activities related to this assembly operation and with discussion with the, we find that these are the things that can be considered.

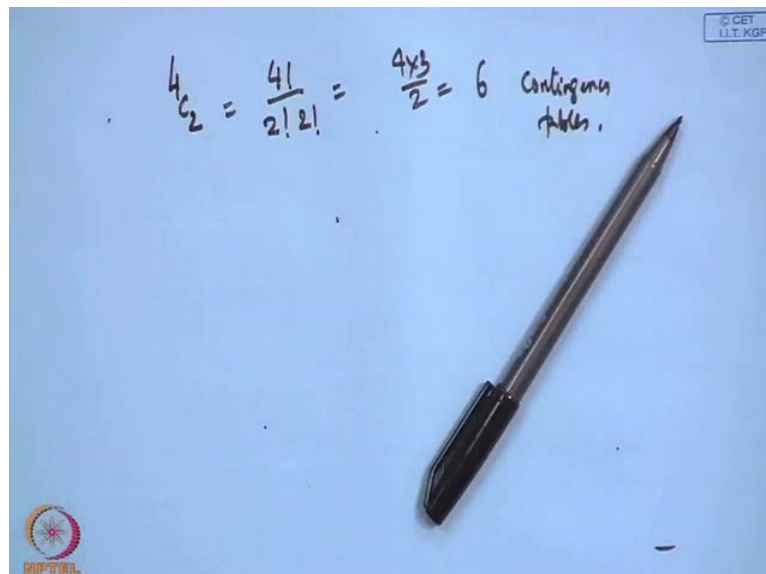
One is for sheet metal component; there is part number, box structure component, double bend component, single bend component, flat component. That means four categories. There are different vendors. Actually, that company having more than fifty vendors, around seventy something vendors are there, but four numbers of vendors, they are

supplying this sheet metal component, so vendor 1, vendor 2, vendor 3, vendor 4. This is another categorical variable.

Then, we our definitely, we want to minimize the defects. So, in order to minimize the defects, we what we have done? We have first seen that what types of defects are generated during assembly operations or after assembly final product when it is sent to the customers, what type of defects are reported plus in house that quality checking what type of defects are identified? Based on this, we found that the category of defects. Anyway, this is also categorical variable having one two three four five six seven eight nine different categories.

Then, it is obvious that there may be seasonal effect, which was which was found after discussion with the personnel; personnel involved in this assembly operation January February to September because we have collected data from January to September of a particular year. So, we could not collect beyond that because ultimate in this assembly operation where you have consulted our study that started in this month and that was that up to September that data was available. So, by saying this, what I mean to say that all variables are categorical in nature.

(Refer Slide Time: 20:43)



${}^4C_2 = \frac{4!}{2! 2!} = \frac{4 \times 3}{2} = 6$ Contingency tables.

Now, let us see that how do we develop contingency table. So, as there are four categorical variables and we want to take two at a time, so that means we will be able to


find out 4 C 2 contingency tables. So, 6 contingency tables that mean six two way two way relationships will be formed out of these four variables.

(Refer Slide Time: 21:09)

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

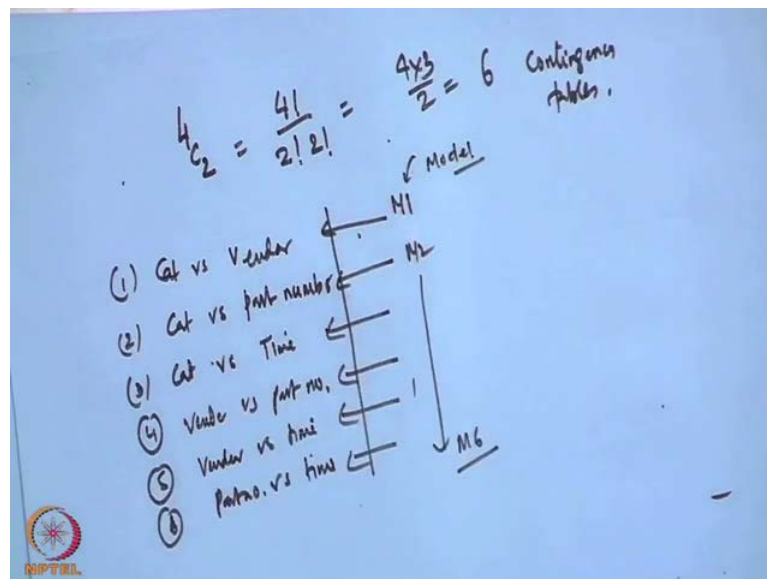

BD - Bend/dent; D - Deburring; DM - Dimension related; FI - Fastener installation; HM - Hole missing; P - Plating; S - scratch; W - Welding

Let us see one after another here categorical data in terms of frequency. So, this is the categories of defects, these are the vendors and what does it mean that vendor 1 supply components having this BD. BD means BD is hole missing bending dimension. I think it is basically bend and dent, B is bend and dent D stands for deburring, DM dimension related, FI fastener installation, so like this, so 150 bend dent defects can be attributed to vendor 1. Similarly, 142 responsible vendor 2, for 146 vendor 3 and vendor 4, 57.

So, if you see this particular column, what you find that ultimately vendor 4 is having less number of defects related to bend and dent in comparison to other 3 vendors that is vendor 1, 2 and 3. They are almost equally, equally contributing towards defects related to bend and dent. Similar, in similar manner, you will be able to find out this. This that means deburring defects, dimension related defects like this. So, if I go for the total, what you found out that that vendor 1 1076, vendor 2 1141, vendor 3 1043, vendor 4 1181, but here in first case, we are finding out vendor 4 is least contributing. Whereas, if I see the total, we are able to see that this vendor 4 is contributing the maximum although with respect to vendor 2.

The difference is not that significant. So, that means there are some relationship may be. What I mean to say as here it is less that means somewhere it is contributing more. For example, if you see the dimension related things, vendor 4 is contributing 269 defects compared to others where vendor 1 is 207. So, in this manner, what I mean to, what I mean to say that, so there may be relationship between the vendors and categories of defects. So, this table is known as contingency table. Now, you can generate similar contingency table for what are those things?

(Refer Slide Time: 24:17)



First one, we have done category of defects versus vendor, this is one. Then category of defect versus you can go for part number, then your category of defect versus your time, then your vendor versus part number, then you can say vendor versus time, then number is, sixth one is your vendor, part number versus time, six tables. Now, six simple correspondence analyses can be done. So, we can say this is M1, M2, so like this. So, there will be M6 model 6, M stands for model.

(Refer Slide Time: 25:30)

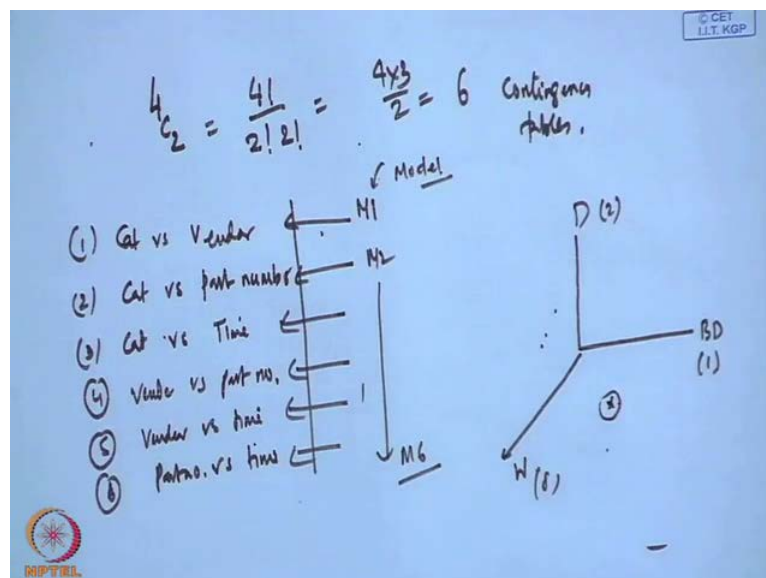
Key Questions

- 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?



Now, with respect to the first model, what we have is the category versus vendor. We have put few questions. What are those questions? What are the similarities and differences among the 4-vendors with respect to the 8-categories of defects. So, we are saying that there are 8 categories 1, 2, 3, 4, 5, 6, 7, 8, 8 categories of defects. Now, that means if I want to see the vendor position, then you are basically getting an eight dimensional issue coordinate. That means what I mean to say?

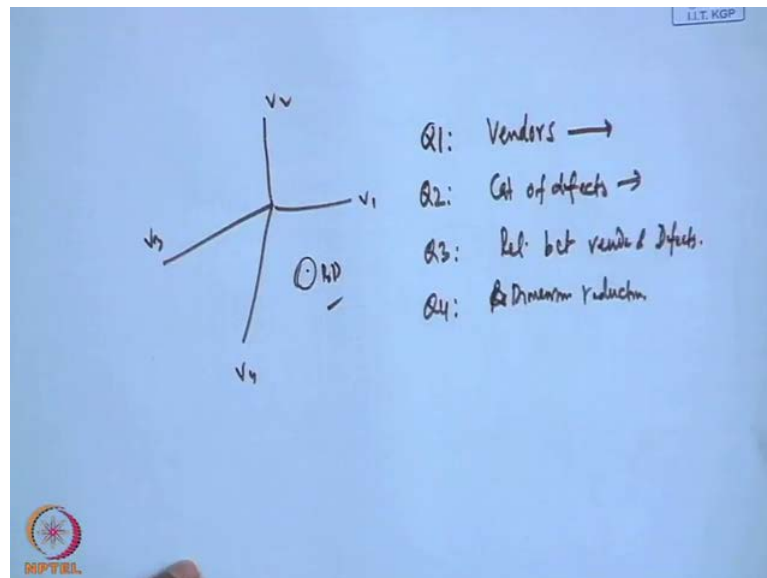
(Refer Slide Time: 26:16)



I mean to say that category of defects that is BD, that D, similarly W, so if it dimension 1, dimension 2, this is dimension 8, not visible here, not visible here, but it is very difficult. What will happen? Now, the vendor 1 has a position suppose somewhere here, so this one is having eight coordinate values that means vendor 1 is measured using this eight categories in terms of one that values are 157, 137, 207, 91 76, 210, 185, 20. Vendor 2 coordinate values are like this, vendor 3 coordinate values are like this, vendor 4 coordinate values are like this.

What we are saying that then what are the similarities and differences among four vendors. So, we using this, this row, this row, this row values, this row values, we want to see whether there is differences between vendor 1, vendor 2, vendor 3, and vendor 4. Are they similar? Are they dissimilar? Note only vendor 1 to vendor 4 four with BD or with D or with DM or FI, we are talking about vendor similarity or differences between vendor 1, vendor 2, vendor 3 and vendor 4 in terms of all the defects. Second one is what are the similarities and differences among the eight categories of defects with respect to four vendors? What does it mean? Let us simply say that we have four vendors.

(Refer Slide Time: 28:05)



We are now considering it four dimensional system that vendor 1, vendor 2, vendor 3, vendor 4, 1, 2, 3, and 4. Now, for a particular defect, you may be say seeing somewhere here in a four dimensional phase, BD is there. Now, what is this BDs point, then that coordinates, then 150, 142, 146 and 57.

Now, if I, if we want to know that our this question, what are the similarities, differences among eight categories of defect with respect to four vendors, so you want to see BD, D, DM, FI, P, S, W, they are similar or dissimilar or if similar, how much similar or how much dissimilar that this vendors with respect to, sorry with respect this, these as well as this eight categories with respect to this four vendors, how they were similar and dissimilar.

That is what is written here, the similarities and differences amongst the eight categories of defects with respect to the four vendors. Now, obviously the third question is what is the relationship between the vendor and the category of defects? These are the vendors and this category of defect. So, when you talk about the first one that what are the similarities and differences among the four vendors with respect to the eight categories of defects, it is similar to main effects of vendor. Second one is similar to main effects of categories of defect bit. Third question is talking about the interactions between vendor and defects. Those interactions we are considering here.

Now, the fourth one that can these relationships be represented graphically in a joint low dimensional space because you have seen that eight vendors, eight categories of defects. If we consider and if you measure vendor 1 to vendor 4, then a particular vendor has eight coordinates, eight dimensional issues. So, but again, when we are measuring the defects category, each of the category with respect to vendors is four dimensional issues. Now, question is the, more the dimension, the complexity is more difficult, very difficult to make decisions. So, what we require? We require reducing the dimension as well.

So, that is why that is the fourth question can these relationships be represented graphically in a joint to low dimensional space. It is possible if there is a relationship amongst the categories of the defects amongst the vendors as well as between categories of defects and vendors. If there is relationship dependency, then the reduction is possible, low dimensional reduction is possible. If it is not there, they are independent, then you cannot reduce the dimension; either you have to go by the vendor wise, that four dimensional issue or category wise eight dimensional issue. You cannot reduce reduction. Reduction is possible only when there is some dependency, some similarity between the anti similarities.

So, now we want to see that again repeat the questions. Question one is our that we are basically four vendors, your target is vendor similarity and dissimilarity, in question two category of defects similarity and dissimilarity, question three what you are trying to say? You are trying to find out relationship between vendor and defect category of defects and question four; you want to reduce the dimensional reduction. So, what are the answers to these questions? How do we know?

(Refer Slide Time: 32:37)

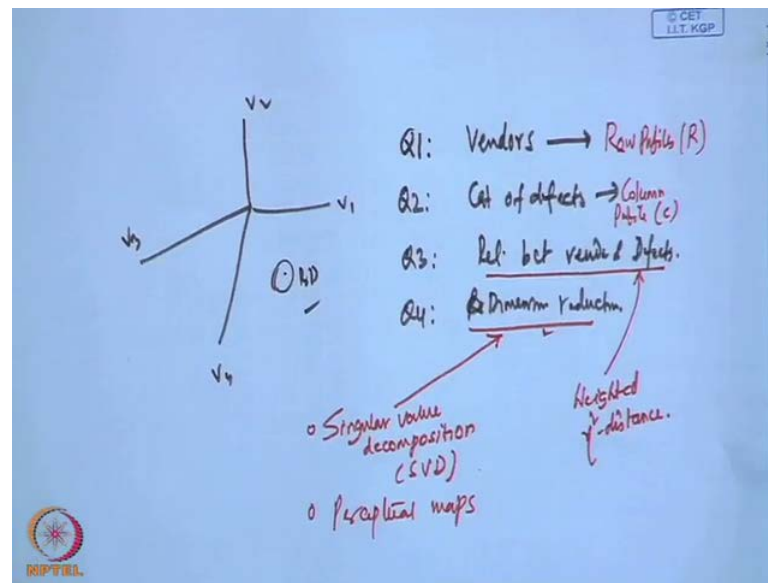
Answers using CA

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



Question number one is answered by row profiles, row profiles.

(Refer Slide Time: 32:44)



Please keep in mind that here along the row; vendors are there along the rows, first row vendor 1 like this. That is why when we are talking about vendor similarity and dissimilarity; I am talking about row profile, but if you change interchange, suppose the categories of defects will come along the rows and vendor along the columns. Then when we talk about row profile, it will be related to the category of defects. So, it is the way you design the contingency table and accordingly the row profile and column profile will give you the concerned reply or responses or answers.


Now, for categories of defect is column, row profile is R and it will be column profile that is C. Now, is there any relationship between these or not? That we will be capturing through weighted chi square distance. The final one that is dimension reduction possible or not, you will be capturing through singular value decomposition SVD. Actually, what happened? SVD will you has different new dimensions, hidden dimensions, latent dimensions and then we will consider two of the most prominent dimensions. Then we create perceptual map using this dimensions, perceptual maps leaving this two where we are seeing answering the fourth one. Now, let us see that what is row profile?

(Refer Slide Time: 35:17)

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**


So, there are different steps to perform correspondence analysis. The contingency table what you see earlier, this one, we are defining this as X, capital X. That 4 vendors, 1, 2, 3, 4, 8 defects, 1 to 8, everywhere some value is there. This is X and this value is x11, x12, x18, like this, similarly x41, x42, and x48. This is basically frequency data contingency table, whatever way you can define like this.

(Refer Slide Time: 36:22)

Row Profiles [R]

- Step 1 – Obtain correspondence matrix Z where each element in Z, i.e., $z_{ij} = x_{ij}/N$

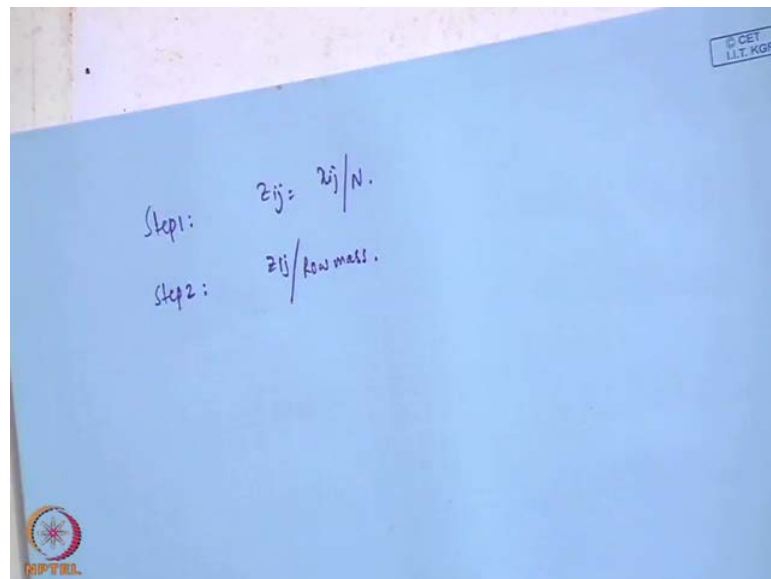
	BD	D	DM	FI	HM	P	S	W	Row mass
Vendor 1	0.034	0.031	0.047	0.020	0.017	0.047	0.042	0.005	0.242
Vendor 2	0.032	0.031	0.045	0.027	0.024	0.050	0.042	0.007	0.257
Vendor 3	0.033	0.029	0.043	0.026	0.020	0.046	0.033	0.005	0.235
Vendor 4	0.013	0.015	0.061	0.059	0.020	0.036	0.054	0.009	0.266
Column mass	0.111	0.107	0.196	0.132	0.080	0.179	0.170	0.025	1.000

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

Now, what is our step one? Now, what is our step one? Step one is you develop correspondence matrix. You see what is it in there obtain correspondence matrix Z where

each element in Z is that x_{ij} divided by capital N , where capital N is the total number of observations. So, what I mean to say here? I mean to say this 4441, this is the total number of defects found for this particular case for I think up to September that is nine months data. Then this is capital N , 150 here, this is x_{11} , 137 is x_{12} , so we are now creating z_{11} , which will be 150 by 4441. That means each of the element in this x table here will be divided by the total that is what we are talking about here. So, 0.34 is nothing but 150 divided by 4441 and 4441 when divided by 4441, this will be 1. So, this is your first step. Then so that mean if I write here that what is our step one.

(Refer Slide Time: 37:56)




Step one is let us found out z_{ij} which is x_{ij} by N .

(Refer Slide Time: 38:09)

Row Profiles [R] (contd.)

- Step 2 – Divide each element of matrix Z by the respective row mass

	BD	D	DM	FI	HM	P	S	W	Row mass
Vendor 1	0.139	0.127	0.192	0.085	0.071	0.195	0.172	0.019	0.242
Vendor 2	0.124	0.122	0.175	0.105	0.092	0.194	0.162	0.025	0.257
Vendor 3	0.140	0.125	0.185	0.109	0.083	0.197	0.142	0.019	0.235
Vendor 4	0.048	0.058	0.228	0.220	0.074	0.135	0.202	0.036	0.266
Column mass	0.111	0.107	0.196	0.132	0.080	0.179	0.170	0.025	1.000

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

Now, go to step two, divide each element of matrix Z by the respective row mass. If you see what is row mass here, now you go back to the original table 1076 divided by 4441, this is what is your row mass? So, that means row total divided by total, so first step one is this step two each of these elements will be divided by row mass. So, that is why, what happened 0.034 divided by 0.242, it is giving you 0.139. So, each of the elements of the Z matrix is divided by their corresponding row mass. So, vendor 1 case, each of the element here, these elements are divided by this vendor 2 case. Each of these elements of z matrix is divided by 0.257, so like this.

So, let us write down the step two, step two we are saying that z_{ij} when we divided by row mass, keep in mind the corresponding row mass. You see that in the first step one, the column values mass if you see that column mass is your column mass is this, 495 divided by 4441, 474 divided by this. That means the column total divided by the total grand total. So, column total divided by grand total is like this. If you sum up, row mass is 0.242, 0.257, 0.235, 0.266, this will lead to 1. If sum up the column masses, this will lead to 1.

So, while calculating row profile, first we are finding out this that the correspondence matrix. Then what you are doing? You are basically dividing each of the elements of the


correspondence matrix by their corresponding row masses, but see you are not considering column mass here because we are talking about row profiles.

(Refer Slide Time: 40:50)

Column Profiles [C]

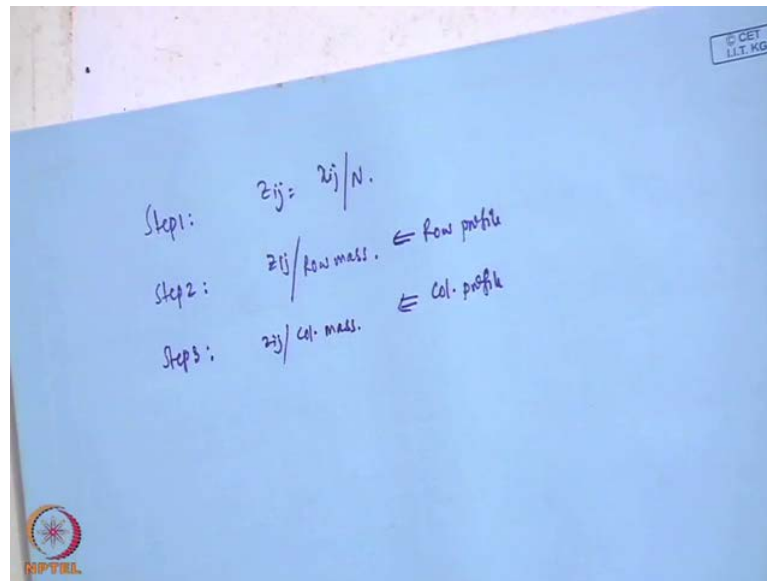
- Step 3 – Divide each element of matrix Z by the respective column mass

	BD	D	DM	FI	HM	P	S	W	Row mass
Vendor 1	0.303	0.289	0.238	0.156	0.214	0.264	0.244	0.180	0.242
Vendor 2	0.287	0.293	0.230	0.205	0.296	0.278	0.244	0.261	0.257
Vendor 3	0.295	0.274	0.222	0.195	0.245	0.258	0.196	0.180	0.235
Vendor 4	0.115	0.143	0.310	0.444	0.245	0.200	0.316	0.378	0.266
Column mass	0.111	0.107	0.196	0.132	0.080	0.179	0.170	0.025	1.000

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

So, now I think you understand now that what is your row profile. So, that vendor 1 position, if we consider in terms of row profiles, so that mean with respect to this eight categories of defects. So, vendor 1 position is 0.139, 0.127, 0.192, 0.085 like this, vendor 2 like this, vendor 3 like this, vendor 4 like this. So, under eight dimensions, you are basically putting that where is vendor 1. Getting me? Where is vendor 2? So, you are getting these values. So, in the similar manner, you can find out the column profiles.

(Refer Slide Time: 41:45)



Step three is we want to find out, so this is your row profile, you got in second stage. Third stage, you want to know the column profile. What you will do? Again, you will basically z_{ij} by column mass that this work into basically, so when I talk about divide each element of the correspondence matrix Z by the respective column mass, this is the case. So, each one this is my that original correspondence matrix Z , 0.034 is divided by 0.111. This is what is given here, 0.303.

Similarly, 0.287 like this. So, what I does it mean? That means that when you are talking about column profile, we are talking about here different categories of defects. So, in a four dimensional space, where the dimensions are created by the four vendors, where does each category of defect lies? So, I think that the first that row profile will give you similarity, dissimilarity and column of the row variables categories, column profile will give you the similarity, dissimilarity between the column variable categories. Now, we want to combine the two. So, we require having some other measure what is known as step four.

(Refer Slide Time: 43:29)

Weighted χ^2 -distances [D]

- Step 4 – $D = D_r^{-1/2} (Z - rc^T) D_c^{-1/2}$

	BD	D	DM	FI	HM	P	S	W	Row mass
Vendor 1	0.043	0.027	-0.002	-0.036	-0.010	0.017	0.003	-0.003	0.242
Vendor 2	0.020	0.018	-0.019	-0.020	0.014	0.018	-0.006	0.002	0.257
Vendor 3	0.042	0.021	-0.011	-0.016	0.005	0.018	-0.026	-0.003	0.235
Vendor 4	-0.094	-0.069	0.031	0.070	-0.005	-0.050	0.030	0.007	0.266
Column mass	0.111	0.107	0.196	0.132	0.080	0.179	0.170	0.025	1.000

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

D can be used to explain the relationships between vendors and category of defects.

Here that weighted chi square distance, so weighted chi square distance is calculated like this.

(Refer Slide Time: 43:44)

Step 1: $z_{ij} = x_{ij} / N$
 Step 2: $z_{ij} / \text{row mass.} \leftarrow \text{row profile}$
 Step 3: $z_{ij} / \text{col. mass.} \leftarrow \text{col. profile}$
 Step 4: Weighted χ^2 -distance = $D = D_r^{-1/2} (Z - rc^T) D_c^{-1/2}$

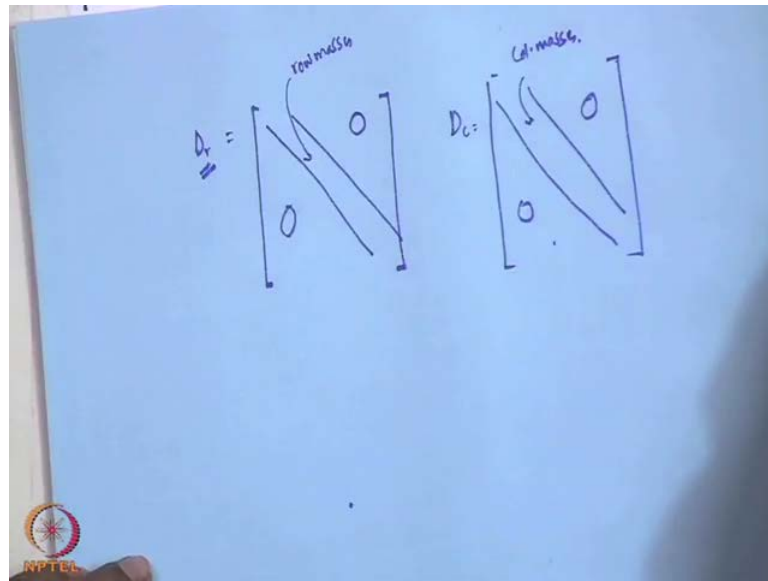
$Z_{4 \times 8} = \begin{bmatrix} 0.242 \\ 0.257 \\ 0.235 \\ 0.266 \end{bmatrix}$ $C = \begin{bmatrix} 0.111 \\ 0.107 \\ 0.196 \\ 0.132 \\ 0.080 \\ 0.179 \\ 0.170 \\ 0.025 \end{bmatrix}$ $rc^T = (4 \times 1)(1 \times 8)$

Your step four is that weighted chi square distance, which we are talking about as D and we are writing this D equal to Dr minus half Z minus rc transpose into Dc minus half. So, this is basically r stands for row, c stands for column. What way it is done? Now, Z, what is the Z matrix? This we have seen that 4 categories of 4 vendors cross 8 categories

of defects. Now, r which is row profile, which is 4 cross 1, so that means this one is nothing but 0.242, 0.257, 0.235, 0.266.

Now, what is your c ? c is 8 cross 1, 8 cross 1 that is 0.111, in the same manner that 0.025. So, if I multiplied this two rc transpose, this will ultimately lead to 4 cross 1, 1 cross 8. So, 4 cross 8, which is matching with this one, this totality, it is a matrix of dimension 4 cross 8. Then what is our D_r ? You also require to know what is D_r ?

(Refer Slide Time: 45:45)



D_r here, it will be a diagonal matrix, where the diagonal elements for these will be the row masses and for D_c , it will be the again a diagonal matrix, diagonal elements will be the column masses, what is off diagonal elements zero, both the cases. So, what will happen ultimately? Then by D minus half, you are taking the inverse and then square root of the inverse basically what you are doing here. So, if I say like this 4 cross 4, then 8 cross 8. Then this will be 4 cross 8. Ultimately, your resultant matrix will be 4 cross 8, total matrix is 4 cross 8.

This is what is chi square distance and this chi square distance, if we use this formulation where your rc , D_r , D_c everything is available because you have the contingency table, the data is given to you. You see I have this is r , this one my c . Then I put create a 4 cross 4 matrix putting diagonal, all this values, we will be getting D_r ; putting diagonal, all these values, we will be getting D_c . Then you take that inverse square root of each of the values there. Then you start multiply multiply this with this. Then the resultant

will be multiplied with this. So, ultimately you will be getting this. This is what we are talking about chi squared distance.

What is the issue here is that you see r and c ? You are seeing r and c , r is this row mass and this column mass. These two you are multiplying, these are again mass which is divided by the total. This two you are multiplying and you are subtracting from from Z , original Z , this Z value, so this value minus this cross this. So, if there is no independency between this two, vendor versus this from the marginal probability concept point of view, what we know this joint probability here, this will be the sum total of that is the multiple case of the marginal probability independent case.

So, in that case, what will happen? These cross these or this minus this cross, this will become 0, distance will be 0. There is some distance that means there is some relationship between these two sides. Any problem? I do not think there will be any problem, but is a very straight forward thing. So, what is written here. You see D can be used to explain the relationships between vendors and categories of defects this distance, but it is see it is a really difficult proposition. It is very easy to say, but when you talk about vendor, it has distance with respect to this all categories of defects.

Similarly, when you talk about the defects, it has also with respect to the vendors. So, it is not straight forward as the way we thing, but even then but this is beautiful of of see from the categorical data, nominal data where no nothing is available. Only some frequency is available. Now, how nicely you are able to convert to it in a distance measures and the chi square distance? So, if I know distance, distance is a position. I know one vendor position, another vendor position in terms of distance. So, then it is possible for me to see what is the difference in distance and some conclusions can be made.

(Refer Slide Time: 50:49)

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.

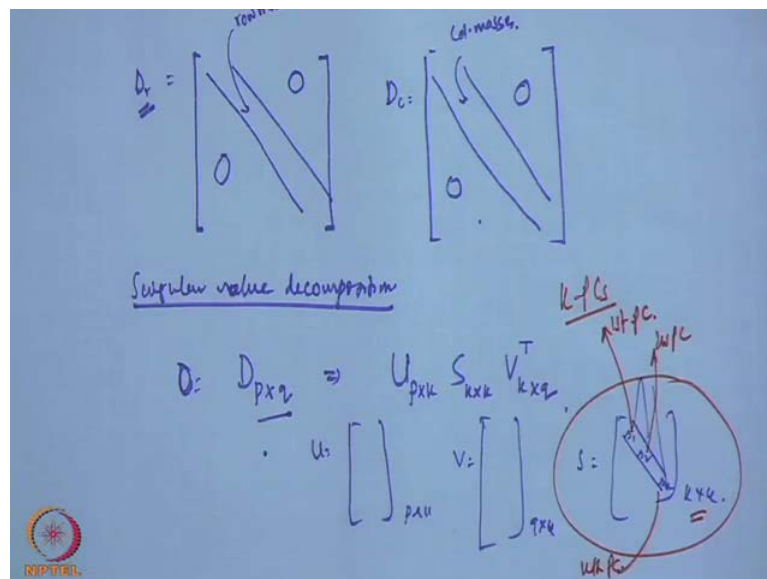


$$k = \min\{p-1, q-1\}$$

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

Then, what is required? Then we will go for singular value decomposition after step four.

(Refer Slide Time: 50:59)



Then, step five is your singular value decomposition. Now, singular value decomposition, I think what is required here that it is a matrix algebra issue. I think that any matrix can be decomposed depending on the nature of matrix. If it is a square matrix, you can go for Eigen value, Eigen vector decomposition. When it is not square, then you cannot go for that. So, singular value decomposition can be composed is non square

matrices. For example, in this case, our D is 4 into 8. Now, you want this is not a square matrix, 4 cross cross 8 is not a square matrix.

What you want to do? We want to decompose this matrix and we we want this decomposition is very, very interesting. I think in my PCA lecture physical component analysis, we have seen that we have decomposed a matrix into Eigen value and Eigen vectors. The way that $\sum \lambda_j = 1$ to p λ_j is a, is a transpose that is since we decompose means one matrix p cross p matrix converted to like this one scalar and with vectors it is also similar way it is done. Now, you see that SVD, what it does? It basically partitions the D matrix into three matrices U , V and S .

In case of Eigen value, Eigen vector you find out the Eigen vectors and Eigen values. Here, three different matrices we are subtracting that decomposition is taking place where U is a p cross k matrix, V is a q cross k matrix and S is a k cross k matrix with diagonal elements. These S diagonal elements are s_1 greater than equal to s_2 greater than equal to s_k . k is the reduced dimension. What we have I am trying to say, we have 4 cross 8 matrix.

Now, we are saying there is independence and because of this feature, we are saying that this matrix easily can be, can have reduced dimensions. So, that reduced dimension, what is reduced dimension? That k is minimum p minus 1 q minus 1. What is p ? p is suppose p is the rows, number of rows. We have 4 rows, we have 8 columns. So, p minus 1 is 4 minus 1, 3, q minus 1 is 8 minus 1, 7. So, the k can be minimum. k is minimum of 3 and 7, which is three.

Now, so that mean, what I mean to say this 4 cross 8 matrix, it can be decomposed into lower dimensions that is the maximum, that dimensions will be 3 like principal component we will be extract. We can be extracting 3 PCS from this D matrix chi squared distance matrix. So, two components are always preferable because then the visual interpretation, the graphing of the two components, all those things will be possible, but three component also possible to map graphically.

So, that means essentially what you are doing then? Your D matrix, which is we are talking about D is p cross q matrix, this p cross q matrix, this you are creating like this U that is p cross k , S k cross k and V k cross q . So, if you say no it is then V transpose you write because we want U will be p cross k , V will be q cross k and S definitely k cross k .

The element of this S , the diagonal element of this S , which are s_1, s_2 like s_k , these are important things. These are the things what we we want to extract. Getting me?

Now, see what we are, we are doing here, s_k is the singular value. This s_k value is known as the singular value for the k th PC. So, that means what we are saying here? We are saying that we are we are able to extract k PCs, k principal components. s_1 is related to first principal component, s_2 is related to second principal component like this s_k is related to k th principal component. So, this is the singular, all singular values, this singular value if you square it, you will be getting Eigen value. All of you know by this time that Eigen values are the variance explains component.

What is the amount of variance of the original data set captured by the k th principal component in this case is given by s_k square. It is similar to λ_k where λ_k is the Eigen value. So, that means Eigen value is the square of singular value. So, represent the weighted variance explained by the k th principal component.

We will stop here. Then I will in the next class, I will continue with this singular value decomposition approach and we will see with this particular case, how ultimately will lead to developing perceptual maps means creating different principal components, their coordinate values, then mapping them. Then finally, what will happen from the numerical measures of CA as well as the maps, we want to infer about the defects in a design process for this particular case what we have considered.

Thank you very much.