

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

**NPTEL
NPTEL ONLINE CERTIFICATION COURSE**

Marketing Research

**Lec – 30
Cluster Analysis – 1**

**Dr. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology Roorkee**

Welcome everyone to the class of marketing research and analysis till now we have discuss about a several tools and techniques used in the you know in the marketing space how companies utilize them and what do they gain out of it how they make advantage out of those by using this techniques we have discuss some of them one of this which we had recently discussed was a technique which was an interdependent technique in which we try to bring in large number of variables to few one's and create factors out of it which we said as factor analysis right.

So in that we had you know reduce large number of variables may be 100 to a few meaningful ones which I always repeat the word meaningful right once and then give a name to those factors and then try to understand how this factors are going to determine or effect may be some other you know prediction some other kind of a you know relationship. Today we are going to discuss similarly another technique which is equally important right very much utilized in the market basically by marketers right.

Let us say there is a case a company wants to sell let us say some you know some kind of phones right now it is lot of phones are coming mobile phones so it wants to sell mobile phones and it has go the target of about let us say 100000 okay 100000 pieces of you know phones to be sold.

(Refer Slide Time: 01:59)



Now the company want to know how should I sell them right so while selling he wants to target the customers so while targeting the customers he has got certain variables like age of the people let us say income of the people kind of a you know habit of a people some kind of habit let us say so now on basis of these variables the company wants to divide a particular may be state or place be because suppose this is a whole place let us say and now they want to say are we going to target the whole state or no if we because if we are going to target t the whole state for 1lakh pieces of mobile it might be little difficult right.

So is there a way that I can you know divide this state in to several groups or clusters then what it does is he tries to divide the state not a geographically may be on certain other parameters taking these three may be on some clusters may be right some clusters. So now it says out of this clusters they find all these clusters have got different characteristics okay. so it finds that may be once I break this in to four clusters, cluster two is the one which is very promising it seems to the company that it is a very promising you know cluster, why?

Because the people in this cluster are once who are interested to buy such kind of mobile phones which has all the characteristics that this companies phone has got right. So in such a condition this becomes highly useful for the company to understand how we could easily segment and then target the market so that we can achieve better results right. So to do this we are suing the technique which is called cluster analysis, so cluster analysis as understand is to divide or you

know create cluster out of a large pool of respondents right we say respondents cases whatever you can say.

So these respondents are grouped in to several similar clusters right so that each cluster has got people inside or you know who are highly similar or highly homogenous in nature right. So what we understand is the all the respondents within the clusters let us say these are the respondents all the people or a respondent what it say there highly homogenous sorry homogenous in nature that means their behaviors is more or less highly similar they are same right but we understand the two clusters let us say cluster one and cluster four cluster two and cluster three they all different from each other right so this is the basic understanding of clusters analysis that within the cluster there is very menial difference or distance and two clusters between the clusters.

(Refer Slide Time: 05:17)

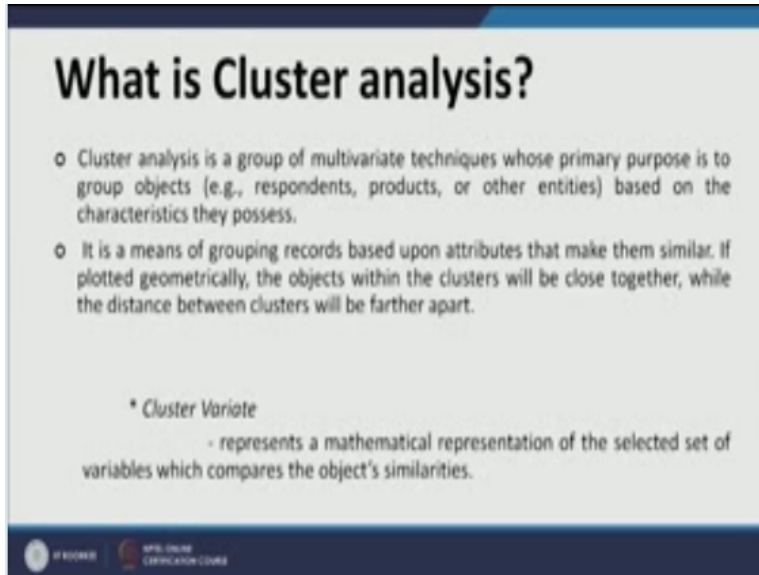


Let us say the distance is high okay so let us see how it functions what happens in a cluster analysis so as I started so what I cluster how to define it says it is a group of multivariate technique who is primary purpose is to group objects right example resonance it could be products it could be entities whatever right so you are trying to group these despondence or products are something into several clusters.

Several groups which other stimulates similar nature right the other thing if you remember in fact analysis we said there were trying to group the variables right so that was variables here

despondence or products or whatever other entities right, so what you are saying it is a means of grouping records based up on certain attributes.

(Refer Slide Time: 06:10)



What is Cluster analysis?

- Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects (e.g., respondents, products, or other entities) based on the characteristics they possess.
- It is a means of grouping records based upon attributes that make them similar. If plotted geometrically, the objects within the clusters will be close together, while the distance between clusters will be farther apart.

* Cluster Variate
- represents a mathematical representation of the selected set of variables which compares the object's similarities.

WU BOKERS WU ONLINE CERTIFICATION COURSE

Now what are the attributes on which one it is grouping now the attributes in this case are for example age.

(Refer Slide Time: 06:17)



In term level let us say habit spend let us say spending habit right or let us say may be technologically savvy how technologically tech savvy are people so maybe that is an indicator of measuring right, so you can have certain criteria certain attributes up on which you can create those similarity right.

So and each similar group is a cluster right so if plotted geometry is objects within the clusters will be close together so that is what I was saying if you plot it then the objects with the clusters are very close to each other that is why it is said they are homogeneous in nature and two clusters are the variables the despondence within two different clusters are highly hydrogenous in nature right.

Then we have said more highly different right this is what cluster basically tells you so cluster variant represents a mathematical representation of the selected set of variables which compares the objected object similarities now what it is saying so mathematically you are trying to plot may be a plot or try to find out some similarity so that it becomes easier for a marker to understand okay which are the clusters he or she cater to right.

It becomes very simple they say take so many example companies of Maruthi is coming up with a new car now should Maruthi target everybody or should Maruthi have a specific thing specific policy in mid how to target whom to target so when it does it has to take certain attributes on this basically attributes they finally decides okay.

(Refer Slide Time: 08:03)

Cluster analysis vs Factor analysis

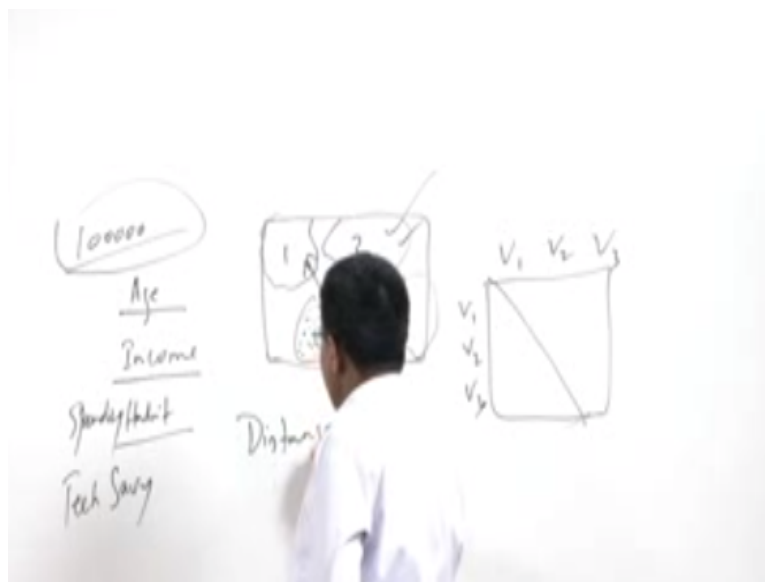
Cluster Analysis	Factor Analysis
- grouping is based on the distance (proximity)	-grouping is based on patterns of variation (correlation)

Factor analysis, we form group of variables based on the several people's responses to those variables. In contrast to Cluster analysis, we group people based on their responses to several variables.

OF BUSINESS MPFL ONLINE CERTIFICATION COURSE

So clusters was a factor analysis clustering is based on the distance matrix and factor we said was in a correlation matrix, so if you remember in factor analysis you said that the most important thing was.

(Refer Slide Time: 08:16)



We were trying to find out a correlation right so we were trying to find a correlation right so when we are trying to find a correlation and this correlation was saying how close where the variables to each other but in cluster analysis we do not take the correlation rather we take the distance what we take we take the distance so the grouping is based on the distance how far or the variables from each other or close or the variables from each other the you know from each other.

So as it says factor analysis we form groups of variables based on the several peoples responses to those variables in contrast cluster analysis we group people based on the responses to several variables now how people have responded to several variables what score they have given so that tells basically their behavior that their thing pattern and all so by taking those variables there thinking pattern in all these things you can classify these dependence into several groups right.


Which as I said is basically a nothing but a mental distance or a distance that is measured right so this distance is basically mental distance sometimes we feel something are very close to us but they are actually not where as some other things some other places might be felt that very far of but they are actually not so far this is actually nothing but a mental perception a mental distance right.

We feel that it is very far of but actually that might be closer right so that is what happens so where are the applications some application are other mentioned.

(Refer Slide Time: 10:06)

Application:

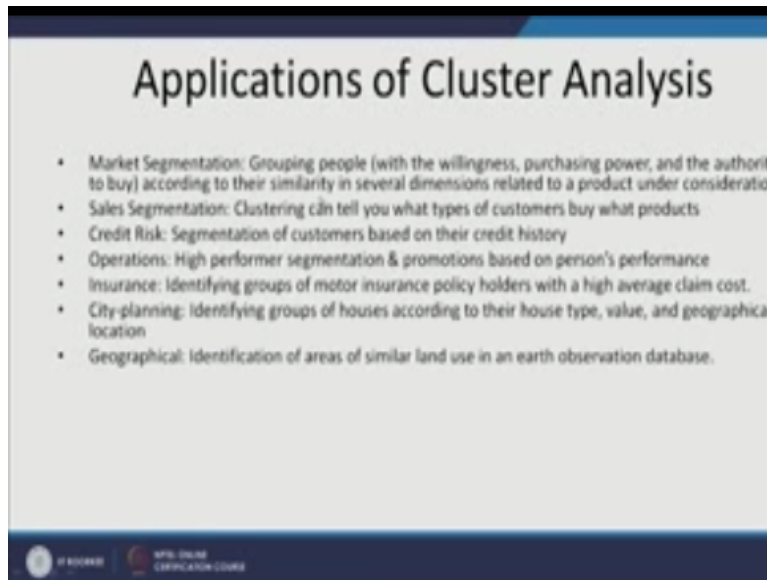
- **Field of psychiatry** - where the characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate form of therapy.
- **Biology** - used to find groups of genes that have similar functions.
- **Information Retrieval** - The world Wide Web consists of billions of Web pages, and the results of a query to a search engine can return thousands of pages. Clustering can be used to group these search results into small number of clusters, each of which captures a particular aspect of the query. For instance, a query of "movie" might return Web pages grouped into categories such as reviews, trailers, stars and theaters. Each category (cluster) can be broken into subcategories (sub-clusters, producing a hierarchical structure that further assists a user's exploration of the query results.
- **Climate** - Understanding the Earth's climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate.



Field of scarcity biology right information retrieval for example the world wide web contains billions of web pages so when you type maybe for example market research right, so all those pages which are linked to research or marketing research they will be clubbed they will brought together right so this is nothing but they are clustering in one way okay fits in the climate for example understanding the earth's climate requires finding pattern in the atmosphere and the ocean to that and cluster analysis have apply to fact fine patterns in the atmospheric pressure.

Of polar regions and areas of the ocean that are the significant impact so how do you say that some places are very similar climatic conditions because this is again where w use cluster analysis to divide the you know as per the all the distance by that two places might be far of but the climatic at the type of climate the weather all this things are very similar in two different places, so they can be still clubbed into one right so understanding these things is very from more applications I can show you market segmentation grouping people.

(Refer Slide Time: 11:18)



Right so grouping people and with the willingness purchasing power the authority to buy according to the similarity in several dimensions states segmentation can tell you what type of customers buy what products, so customer a would buy what kind of a product that is what a helps you know there is how the clusters is helps the marketer right so some examples are always there many more examples city planning, insurance right geographical you know examples so all these are basically are used to group respondents.

According to certain behavior attributes and find those clusters right once you find those clusters it becomes easy for you to polishing may for making policy making for selling something for you know maybe understanding the kind of trend or any trait and it behavioral trait maybe some genetics study you are wanted to do, so everywhere in fact to tell you the use of clusters analysis was not started did not start with marketing or something right it has it is roots here actually in taxonomy in biology.

Where different kinds of spaces where to be classified into different show in groups right so this is how it is all has started.

(Refer Slide Time: 12:41)

Common Roles Cluster Analysis can play:

- **Data Reduction**
-A researcher may be faced with a large number of observations that can be meaningless unless **classified into manageable groups**. CA can perform this data reduction procedure objectively by reducing the info. from an entire population of sample to info. about specific groups.
- **Hypothesis Generation**
- Cluster analysis is also useful when a researcher wishes to develop hypotheses concerning the nature of the data or to examine previously stated hypotheses.

© IIT KANPUR | IIT KANPUR ONLINE CERTIFICATION COURSE


So common role the cluster analysis can play right first is data reduction so as factor analysis also was helpful in data reduction similarly cluster analysis also helps you in data reduction researcher maybe face it larger number of observations that can be meaningless right unless classify to a managerial groups so how many groups suppose you have 10,000 respondents 1 lakh respondents but 1 lakh respondents individually try if you understand this nothing you can do so if you either create 10 clusters out of it.

Then it makes more much of better is meaning out of it right second is hypothesis generation cluster analysis also useful when a researcher wishes to develop hypothesis concern th nature of the data or to examine previously stated hypothesis so cluster analysis is also useful in developing a hypothesis right to it because it has it gives you in sight the knowldege about certain things so it helps you to develop a hypothesis right and even finally test it.

(Refer Slide Time: 13:52)

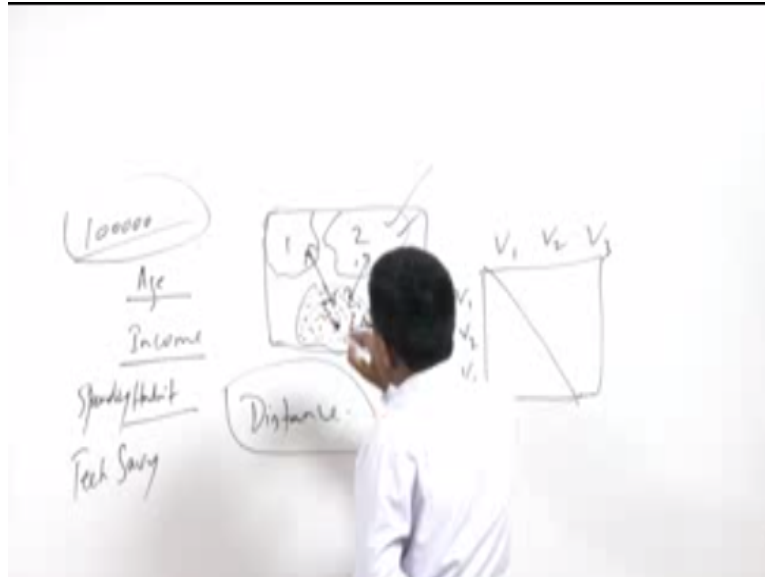
Objectives of cluster analysis

- Cluster analysis used for:
 - **Taxonomy Description.** Identifying groups within the data
 - **Data Simplification.** The ability to analyze groups of similar observations instead all individual observation.
 - **Relationship Identification.** The simplified structure from CA portrays relationships not revealed otherwise.
- Theoretical, conceptual and practical considerations must be observed when selecting clustering variables for CA:
 - Only variables that relate specifically to objectives of the CA are included.
 - Variables selected characterize the individuals (objects) being clustered.



So what are the objectives so taxonomy description so identifying groups data simplification the ability to analyze groups on similar you know observations instead all individual which I said is very, very taxing and difficult impossible sometimes then finally his relationship identification where it says the simplified structure from cluster portraits relationships not revealed otherwise sometimes we can find out okay let say now I have said there are four clusters right so are this four clusters okay.

(Refer Slide Time: 14:21)



Fine there is no doubt that this four clusters are different but this is a possibility that cluster 1 and cluster 3 are actually very closed to each other sometimes it happens in your state where you stay that maybe one distinct the state has got 10 districts out of which two districts are extremely similar because of their maybe a language there food habits are something ore cultural habits so they are very similar so they can sometimes if you want you can even club them and make it like one cluster right.

You can make it so this relationship identification is there important job of cluster analysis how does work.

(Refer Slide Time: 15:01)

How does Cluster Analysis work?

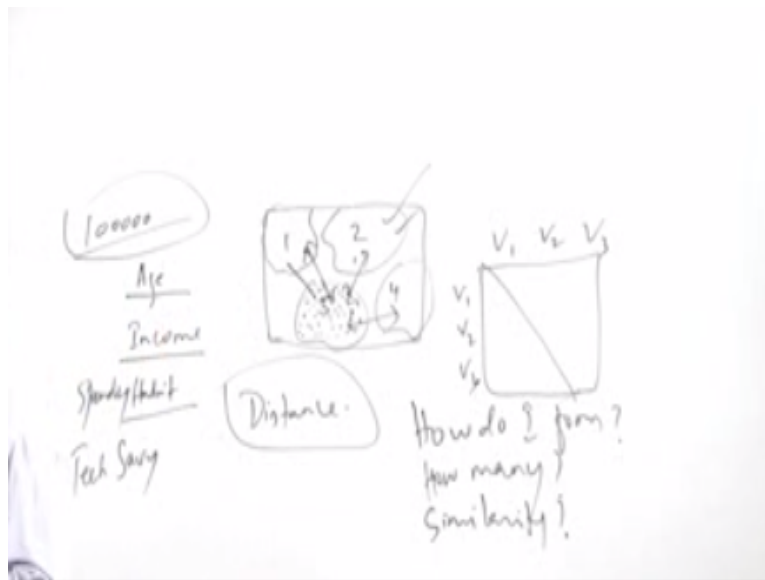
The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups. To accomplish this task, we must address three basic questions:

- How do we measure similarity?
- How do we form clusters?
- How many groups do we form?

So three basic things are there, now what are these three basic things so it says the primary objective of cluster analysis is to define the structure of the data right, so by placing the most similar observation into groups remember. If you are working with data and trying to find out similarity so that you can form groups then there must, there could be some problems with it also. Suppose the biggest problem that can affect cluster analysis is suppose you have a data which is got there is few outliers in the data.

Now if you have few outliers in the data then that could completely change the way the groups are formed, that could be a very important thing one should keep in mind okay. Three things that are very important what, how do we measure the similarity so when you are doing the cluster analysis the question comes how do I know key which groups are similar there are four clusters.

Now how do I know which are similar to each other, one second thing now do I form the clusters, how do we form the clusters it is not that key the data is given to us and we just do it, there must be a way right, so how do I form the clusters. Third, how many so the question is how do I form clusters, how do I form.
(Refer Slide Time: 16:25)



How do I form, right, how many do I form, right how many, how do I form, how many do I form and how do I measure a similarity, how do I measure the similarity so if I can measure the similarity then only I can do this right, so let us see but one thing is you have to keep in mind that in the cluster analysis we are not looking at the correlation and why we are not looking at the correlation I will explain you, see correlation could be like something like this you know suppose two things are moving like this.

The correlation might be very high among them but suppose this two verses let us say take this you know just for understanding, understand this way. suppose this two lines and these two lines if you look the correlation between the two both the lines pairs let us A, B,C,D, A,B and C and D suppose you these two are one group, these two are one group is more or less same the correlations more or less same.

But if you look at the distance is actually not same so that is what is the basic underlying difference between cluster and the distance and the correlation, okay.

(Refer Slide Time: 17:49)

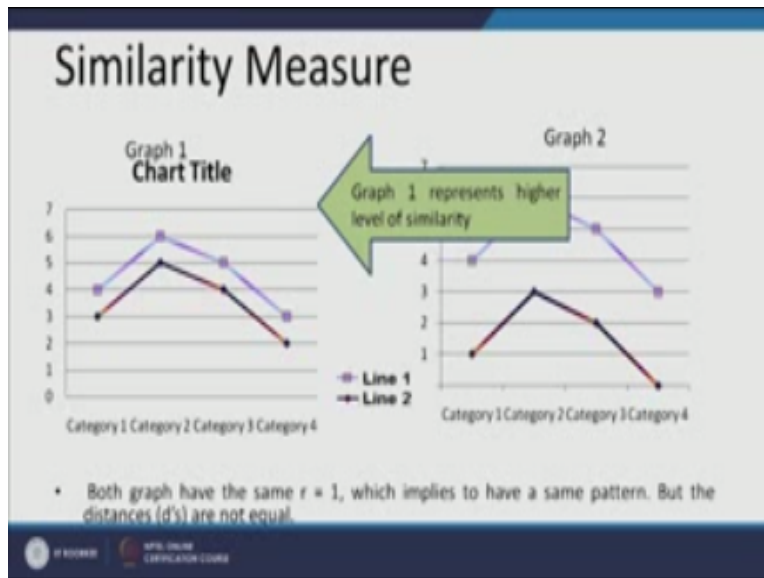
Measuring Similarity

- **Similarity** represents the degree of correspondence among objects across all of the characteristics used in the analysis. It is a set of rules that serve as criteria for grouping or separating items.
 - **Correlational measures.**
 - Less frequently used, where large values of r 's do indicate similarity
 - **Distance Measures.**
 - Most often used as a measure of similarity, with higher values representing greater dissimilarity (distance between cases), not similarity.

So what is similarity represents the degree of correspondents among objects across all the characteristics used in the analysis, so I have several attributes as I have used let us say I have ask several variables, there are several variables which I have used in the study some of them being like income, age and all these things. Now taking them together when I am trying to build a kind of a similarity matrix right, so this all this variables together will help me not one but together they will help me in defining a cluster, right, okay.

So as I said correlation basis are less frequently used only in case there is a special case where correlation is used where you know when there is something called this is called a molecular distance which is used when the certain variables do have a correlation we use it in special case I will come to that right, otherwise most often the similarity is measured through the distance, okay.

(Refer Slide Time: 18:58)



Now this is what I was trying to explain you see if you look at it, if you look at the two charts, the two graphs the graph 1 and graph 2 both the $R=1$, now what is R the correlation value so the coefficient of correlation is 1 that means they are highly correlated right, so which implies to have a same pattern right, but the distances are not equal, the distances between these and the distances between these are not equal so these two will get a similar different you know interpretation in cluster analysis but had not been enough case of a factor it would have been, it to due to very similar, right.

So that is the, thus the basic differences okay, so graph 1 represents high level of similarity right, and graph 1 because the distance is less, why it is saying now because if the distance is less that means they are close to each other as good as that, right as that right if it is close the similar there will be coming trending towards each other right if the distance is high or the correlation the trend is same but there is a sufficient gap right so all this things are very important to understand this right.

(Refer Slide Time: 20:35)



So the distance now how do you measure the distance several ways the basic way of measuring the distance is the Euclidean distance now Euclidean distance is the distance which we say is the straight line right so how do we calculate so D is equal to let say $\sqrt{x_2 - x_1^2 + y_2 - y_1^2}$ right so this is what basically how we measure right so the most commonly recognized 12 straight line distance right.

(Refer Slide Time: 20:54)

Distance Measures

• several distance measures are available, each with specific characteristics.

- **Euclidean distance.** The most commonly recognized to as straight-line distance.

$$d_{\text{Euclidean}}(B, C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

- **Squared Euclidean distance.** The sum of the squared differences without taking the square root.
- **City-block (Manhattan) distance. Not based on Euclidean distance.** Uses the sum of the variables' absolute differences

$$d_{\text{City-Block}}(B, C) = |x_B - x_C| + |y_B - y_C|$$

This is how you measure so the other forms also for example square Euclidean distance which says that you take the sum of the squared differences without taking the square root that means you only omit the square root right okay. So the other distance is like the city block distance Manhattan distance which is not like the Euclidean because this one takes the absolute value.

And it is sometimes it is it does not work well because of this nature of absolute value it does not work well right so one is to be very careful which you are mostly if you do not understand much simply you can go for the Euclidean distance because that is another safest way until the correlation among the variables right.

(Refer Slide Time: 21:43)

- **Chebychev distance.** Is the maximum of the absolute difference in the clustering variables' values. Frequently used when working with metric (or ordinal) data.

$$d_{\text{Chebychev}}(B, C) = \max(|x_1 - x_2|, |y_1 - y_2|)$$
- **Mahalanobis distance (D^2).** Is a generalized distance measure that accounts for the correlations among variables in a way that weights each variables equally. Best suited when variables are coorelated

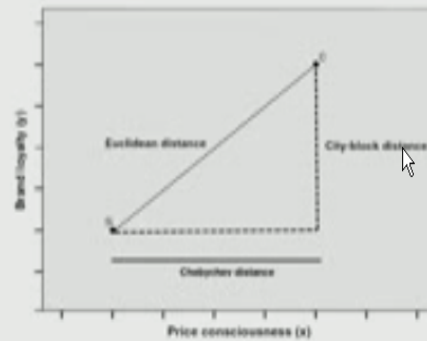
Then you have two more like Chebychev distance which is taking the maximum of the absolute difference in the clustering variables value for example maximum of $x_2 - x_1$ or $y_2 - x_1$ so the absolute value is only taken right they also which we are saying is the mahalanobis distance which measures accounts for the correlation among the variables in a way that each variables equally.

Now Mahalanobis distance is also very important tool which is used to also find out, out layers let me tell you this is may be not here but suppose you are doing the simple regression and you want to find out, out layers mahalanobis distance is the technique which is used to measure to find out those out layers okay so this is one way of doing it.

You remember only the condition applied is when the variables have a large correlation or high correlation among them that time it is preferable to use a mahalanobis distance over the other distances and these things you will find almost every were in the software nowadays right so you do not need to calculate it by hand okay.

(Refer Slide Time: 22:54)

Illustration:



Now this is how the Euclidean distance looks like all the three one go it is showing you now this is hypothesis measure right.

(Refer Slide Time: 23:02)

Simple Example

- Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores. A small sample of seven respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty- V_1 (store loyalty) and V_2 (brand loyalty)- were measured for each respondent on a 0-10 scale.

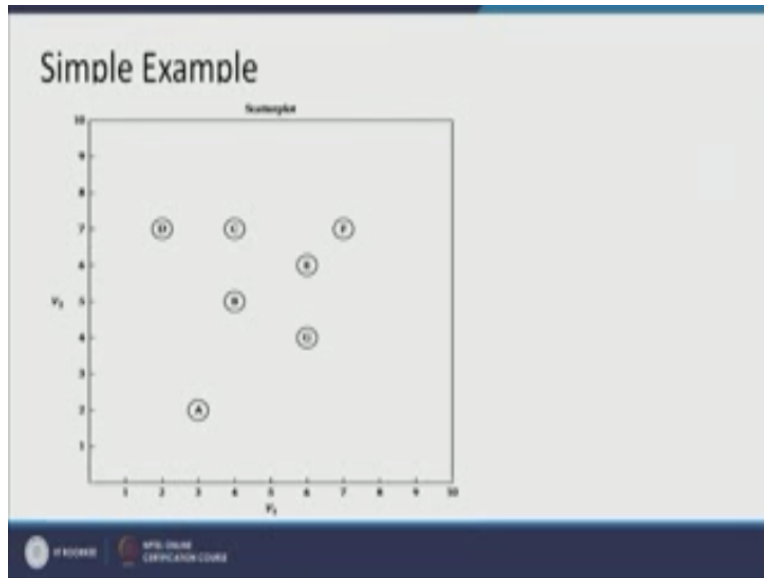
Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

So let us say this is an example knowledge starts with the example now this example if you can go through a market researcher a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to stores a small sample of 7 respondents is selected as a pilot two measures of loyalty store loyalty and brand loyalty V_1 and V_2 right were measured for each respondents on a 0-10 scale right.

The scores are given to you so these are the 7 respondents A B C D E F G and the score that they are given for store loyalty and brand loyalty is been given to you on the scale of 0-10 right now from this data let us see how we can get into the cluster analysis okay so when I place the data on a graph

(Refer Slide Time: 24:04)



So 3 2 4 5 4 7 so if you can see this A is 3, 2 right then B is 4, 5 right C is 4, 7 D is 2, 7 so we can see here D is 2, 7 right so we have just plot the graph and we place them on the graph right

(Refer Slide Time: 24:32)

How do we measure similarity?

Proximity Matrix of Euclidean Distance Between Observations

Observation	Observations						
	A	B	C	D	E	F	G
A	...						
B	3.162	...					
C	5.099	2.000	...				
D	5.099	2.828	2.000	...			
E	5.000	2.236	2.236	4.123	...		
F	6.403	3.606	3.000	5.000	1.414	...	
G	3.606	2.236	3.606	5.000	2.000	3.162	...

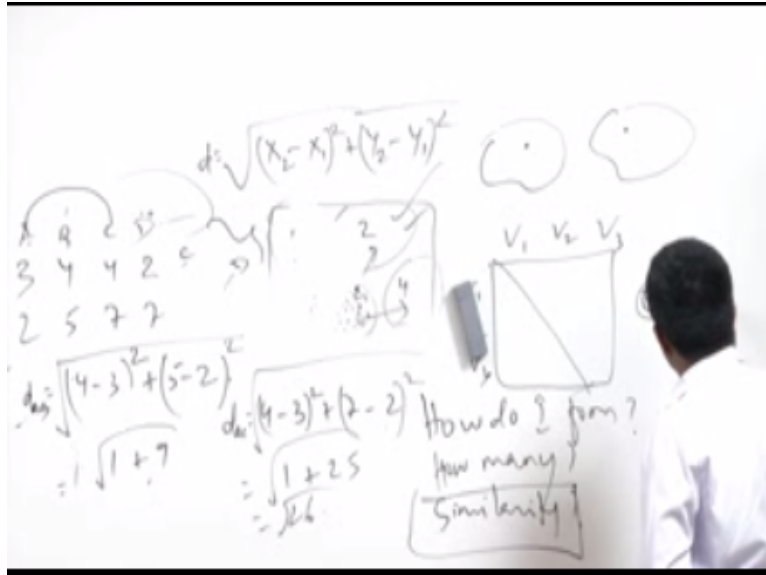
$$d_{\text{Euclidean}}(A, B) = \sqrt{(V_{1(A)} - V_{1(B)})^2 + (V_{2(A)} - V_{2(B)})^2}$$

$$d_{\text{Euclidean}}(A, B) = \sqrt{(3 - 4)^2 + (2 - 5)^2} = 3.162$$

Now how do we measure similarity the first question that we have how do we measure the similarity so to do this right how do we measure similarity I said we can use the Euclidean distance as a way of doing it now how it has done now for example let us take the distances now let us take two distances okay so among between two variables now for example this was 3 2 4 5 4 7 2 7 okay.

So let us take only this much so this ADCD I am taking it. Now suppose if we want to find out the distance, now how do we do it? Now do to do that? What is the way simply, for between A and B.

(Refer Slide Time: 25:34)



so the distance between A and B = $x_2 - x_1$ so 3- 4 right you can say 4- 3 also does not make a difference obviously you will square right, $(4 - 3)^2 + (5 - 2)^2$ sorry either you take it this way so $(4 - 3)^2 + (5 - 2)^2$ so what it is coming, so $1 + 9 = 10$. That means it is coming something around 3.162, similarly you can find the distance for all the other variables right, now the distance between A and B is 3.162.

(Refer Slide Time: 26:29)



How do we measure similarity?

Proximity Matrix of Euclidean Distance Between Observations

Observation	Observations						
	A	B	C	D	E	F	G
A	---						
B	3.162	---					
C	5.099	2.000	---				
D	5.099	2.828	2.000	---			
E	5.000	2.236	2.236	4.123	---		
F	6.403	3.606	3.000	5.000	1.414	---	
G	3.606	2.236	3.606	5.000	2.000	3.162	---

$$d_{Euclidean}(A, B) = \sqrt{(V_{1(A)} - V_{1(B)})^2 + (V_{2(A)} - V_{2(B)})^2}$$

$$d_{Euclidean}(A, B) = \sqrt{(3 - 4)^2 + (2 - 5)^2} = 3.162$$

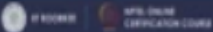



Obviously A and A would be same 1, now let us calculate one more A and C, these two we are measuring, if you measure these two it will be $\sqrt{(4-3)^2 + (7-2)^2}$ so that is $= 1+25 = \sqrt{26}$ so that is $= 5.099$. Something right, so it must be A and C IS 5.099, so for that everything you have measure right.

(Refer Slide Time: 27:18)

How do we form clusters?

- **SIMPLE RULE:**
 - Identify the two most similar(closest) observations not already in the same cluster and combine them.
 - We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own "cluster" and then combining two clusters at a time until all observations are in a single cluster. This process is termed a **hierarchical procedure** because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an **agglomerative method** because clusters are formed by combining existing clusters



So after measuring the next question was how do we form the clusters? So you have measured the distances, now find out the distance between two clusters, the minimum distance between the two clusters, the minimum why I am saying this is the closest, they are very close to each other. Had the distance been more they would have been far away from each other. So identify the two most observations, who are already in the same cluster and combine them.

So what you can do is two similar clusters, 2 observation for example as I explained, this is 1 cluster, so identify two values right and through certain ways, there are certain ways, I will explain that also, similar linkage, average linkage, centroids method, there are different ways how should I use right. So identify the two most similar observations, which are already in the same cluster and combine them so that is what the objective.

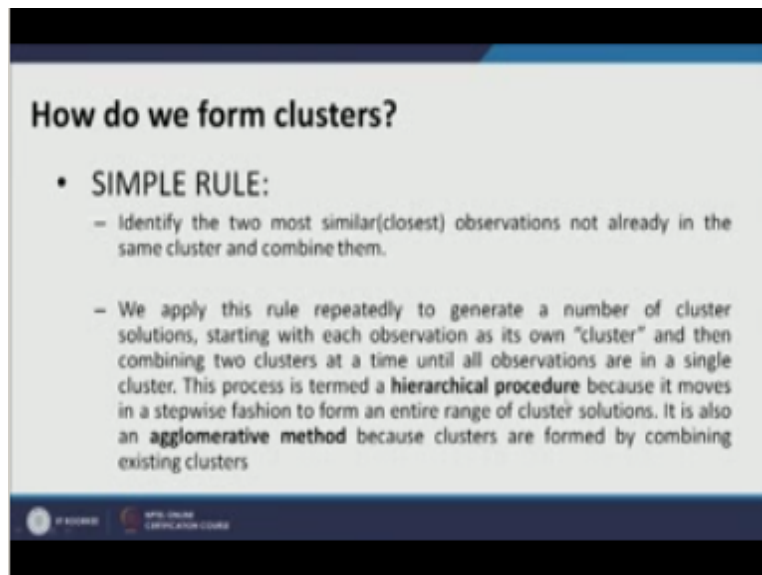
So we want to form the clusters, so once we have identified this, so apply this rule to generate the number of clusters, starting with each observation as it is own clusters right and then combining two clusters at a time until all observations are in the single clusters, that means what you are trying to add up the closest once, now this is one 1 and 2 are close, so next is say 5, so 5 is closest to them, so you add up now 5 then let say 8.

So you add up 8 then let us say 6 you add up 6 so it goes on adding the nearest variable in terms of the distance okay this method of this process is termed as hierarchical procedure why obviously it will say hierarchical procedure because your maintaining a hierarchy right you are

following a hierarchy so when you do this ultimately all the different respondent will be clip together to form a single cluster right single cluster.

But the question is if we have a single cluster then the whole meaning is lost right suppose I want to know to which state suppose the state of let say or this country our country India suppose you want to do it in India then if I say whole market is your market then it becomes very difficult for me to make an interpretation out of it so in those cases you have to understand well what should I do this is not sufficient for me how do I break it into several clusters first so how many clusters should India be broken into for the marker so that it can easily scatter to those clusters so we will see that right the process is termed hierarchical procedure

(Refer Slide Time: 30:20)



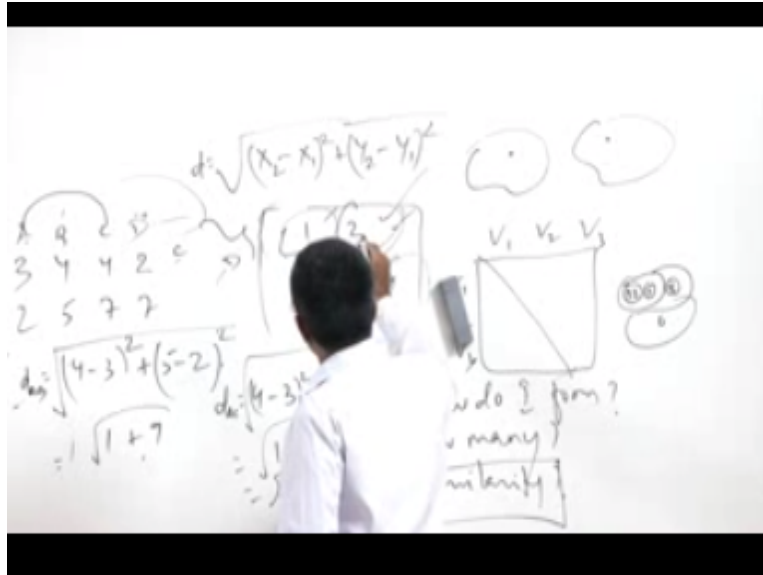
The slide is titled "How do we form clusters?". It features a blue header and footer. The main content is on a white background. A bullet point titled "SIMPLE RULE:" is followed by two sub-points. The first sub-point describes identifying the two most similar observations not in the same cluster and combining them. The second sub-point describes applying this rule repeatedly to generate cluster solutions, starting with each observation as its own "cluster" and combining two clusters at a time until all observations are in a single cluster. This process is termed a "hierarchical procedure" and is also an "agglomerative method".

How do we form clusters?

- **SIMPLE RULE:**
 - Identify the two most similar(closest) observations not already in the same cluster and combine them.
 - We apply this rule repeatedly to generate a number of cluster solutions, starting with each observation as its own "cluster" and then combining two clusters at a time until all observations are in a single cluster. This process is termed a **hierarchical procedure** because it moves in a stepwise fashion to form an entire range of cluster solutions. It is also an **agglomerative method** because clusters are formed by combining existing clusters

Because it moves in a step wise fashion to form an entire range of cluster solutions right it is also agglomerative method because clusters are formed by combining the existing clusters so what it is saying so you are trying to form the clusters because let us say cluster one as I have drawn here.

(Refer Slide Time: 30:43)



Let us say there was a cluster one there was cluster two so cluster three so if I add up 1+2+3+4 then it becomes the whole place the big country India now the question is how do I add up should I add up like anybody or everybody or should I have a mechanism now what is the mechanism the mechanism is find out the two clusters which are very similar or close to each other.

And once you can find out those two clusters which are close to each other and you start on adding and ultimately you will land up to the whole market that is India right so this is why it is called a agglomerative method okay now let us look at this so the agglomerative process is the process in which what we have done is if you look at it this is the minimum distance.

(Refer Slide Time: 31:26)

How do we form clusters?

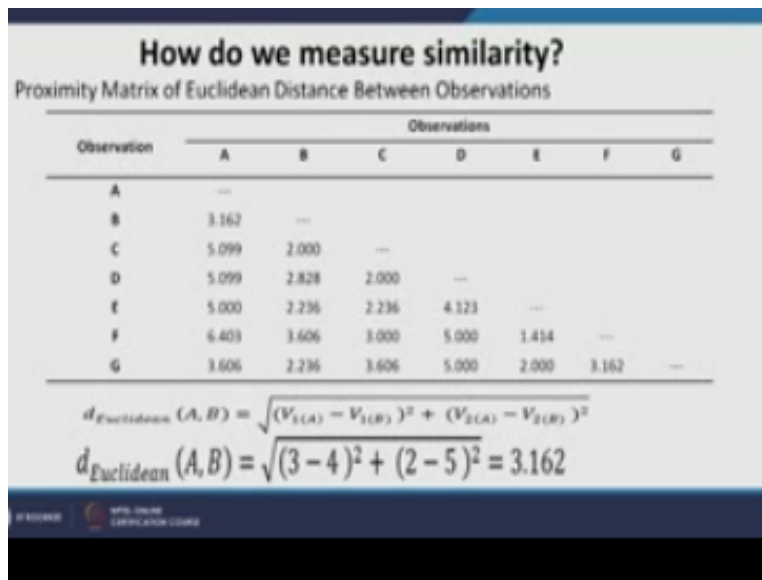
Step	AGGREGATIVE PROCESS		CLUSTER SOLUTION	
	Minimum Distance Uncolored Observations ^a	Observation Pair	Cluster Membership	Overall Similarity Measure (Average Within-Cluster Distance)
	Initial Solution		(A)(B)(C)(D)(E)(F)(G)	7
1	1.414	E-F	(A)(B)(C)(D)(E-F)(G)	6
2	2.000	E-G	(A)(B)(C)(D)(E-F-G)	5
3	2.000	C-D	(A)(B)(C-D)(E-F-G)	4
4	2.000	B-C	(A)(B-C)(D)(E-F-G)	3
5	2.236	B-D	(A)(B-C-D)(E-F-G)	2
6	3.162	A-B	(A-B-C-D)(E-F-G)	1

In steps 1,2,3 and 4, the OSM does not change substantially, which indicates that we are forming other clusters with essentially the same heterogeneity of the existing clusters.

When we get to step 5, we see a large increase. This indicates that joining clusters (B-C-D) and (E-F-G) resulted a single cluster that was markedly less homogenous.

That you have to calculated by now we have done this through the distance now after doing this what were the distances now distance we have arranged it so 1.414 to 2.236 now the pairs of respondent which were related to it are the first one is E and F the second one is EFG I hope you can recall that.

(Refer Slide Time: 31:56)



Let go back and look at it so the lowest value is at E and F okay so E and F is 1.414 is the lowest so the next one is there are three next lowest 2 right in this line and then in this line again you have a 2 and again you have 2 so now you can take the closest one for example in this case what we have done E and F for the first pair then second pair we took E and G right because U was already there so the closest to E we found out and then C and D and finally we took we started doing each right if you can see the cluster membership here abcdefg so there are seven clusters.

So we have not done any grouping right now after this what we did was we club E and F together here to form a single cluster so they become 6 clusters then right then we did abcdef and g we added EFG together so 5 clusters then we added ABCD together then EFG right so what we have done is basically we have shorten we have reduced the number of clusters to a few so that we can finally land up into one right so by doing this where I will explain you may be in the next session how do we calculate this part also right this oval similarity measure the average within the cluster distance this also I will explain.

So we are identifying the number of cluster and through this we can say finally by through this value this data how many clusters in this case we should have right so well this is what just the introduction of the cluster analysis so we will in the next session we will get into more details rights how one should form the clusters and how then one should interpret the clusters okay thank you for this session.

For Further Details Contact

**Coordinator, Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee 247 667**

**E-Mail Ecellitrke@gmail.com etcell@itr.ernet.in
Website: www.itr.ac.in/centers/ETC. www.nptel.ac.in**

Production Team

Sarath Koovery

Mohan Raj. S

Jithin. K

Pankaj saini

Graphics

Binoy. V.P

Camera

Arun. S

Online Editing

Arun.S

Video Editing

Arun.S

**NPTEL Coordinator
Prof B.K Gandhi**

**An Educational Technology Cell
IIT Roorkee Production**