Marketing Research and Analysis - II (Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture 60
Cluster Analysis in Practice - II

Welcome friends to the lecture of Marketing Research and Analysis. So in the last class, we are talking about a new technique called cluster analysis. So we understood that cluster analysis is a technique very similar to factor analysis, which is also called an interdependence technique. It is a subjective measure where we are trying to classify objects or create groups or clusters, which have lot of similarity within themselves, right.

And our intention is to create such clusters, so that within the clusters there is a lot of homogeneity, but between 2 clusters, there is a lot of heterogeneity. That means 2 clusters are as far from each other is possible, but the data points within the clusters are very much similar or close to each other. That is the objective of the study and where it is used? It is used for in areas like climate change even biology, botany, in genetics, and largely also in marketing.

So our interest is mostly into marketing because we want to segment the market. When we want to segment the market, how do you segment the market? For that you have to understand the consumer’s buying behavior and patterns, right. So to do that cluster analysis is very helpful, another area which is nowadays becoming where cluster analysis also being largely used is in the political system, right.

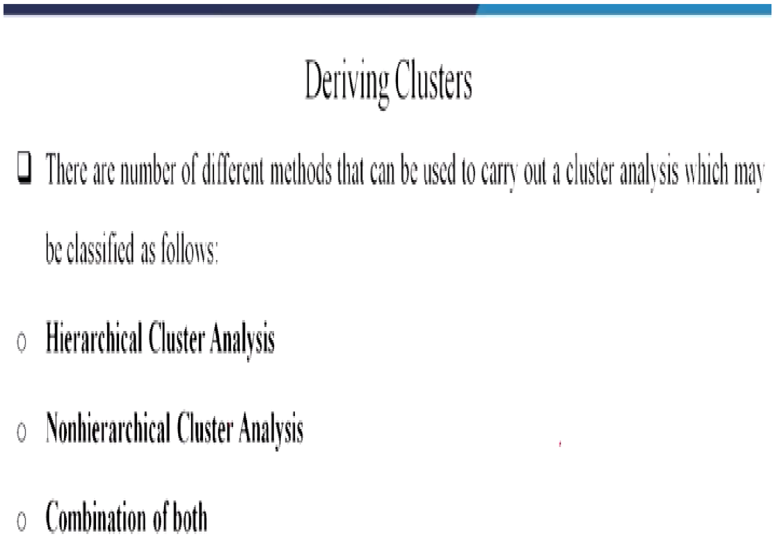
During elections where people want to when leaders are giving the speech, they try to create a speech according to behavior of a particular cluster. So suppose some cluster is a very traditional, for example some leader wants to give a speech in a place where the people are mostly very modern and fashion oriented, so he would talk more about simple such kind of issues.

Suppose somebody where is a very sensitive place and people are more open to relationships and all, so they talk more about such kind of rules and regulations. In some places, people might be very orthodox people, so they would like to talk more about holding to the basics and fundamentals and all. It helps leaders also to create a different kind of speeches for themselves and accordingly catered to the audience. So cluster analysis is being largely utilized in every sphere nowadays, right.

And we said that it uses instead of the core relation measure, it uses the distance measure, right and we learnt in the last lecture that basically there are 2 types of cluster analysis one is the hierarchical cluster analysis and the other is the non-hierarchical cluster analysis. The objective of the hierarchical cluster analysis is to identify the clusters. Basically it helps you to identify the clusters and the non-hierarchical cluster analysis helps you to identify the characteristic of the cluster.

For example, hierarchical only said to you how many cluster can be formed. That how many is explained through the hierarchical clustering and what is the characteristic of the cluster is explained in the non-hierarchical or k-means clustering, we say. So let see how it is done.

(Refer Slide Time: 03:53)



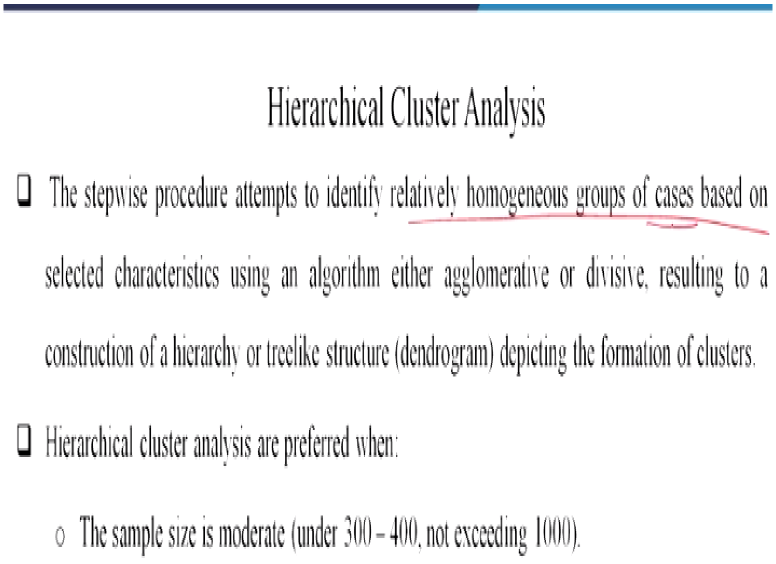
Deriving Clusters

❑ There are number of different methods that can be used to carry out a cluster analysis which may be classified as follows:

- ○ Hierarchical Cluster Analysis
- ○ Nonhierarchical Cluster Analysis
- ○ Combination of both

So hierarchical cluster analysis as I was just saying non-hierarchical cluster analysis and combination of both. So sometimes many of the times, it is better, it is advised that we used a combination of the hierarchical cluster and non-hierarchical cluster, so that here the number of groups or number of clusters are decided and here the characteristic of the clusters are decided or observed, of the cluster is observed, right and then the researcher or the marketer can decide, which cluster to cater to, right.

(Refer Slide Time: 04:30)



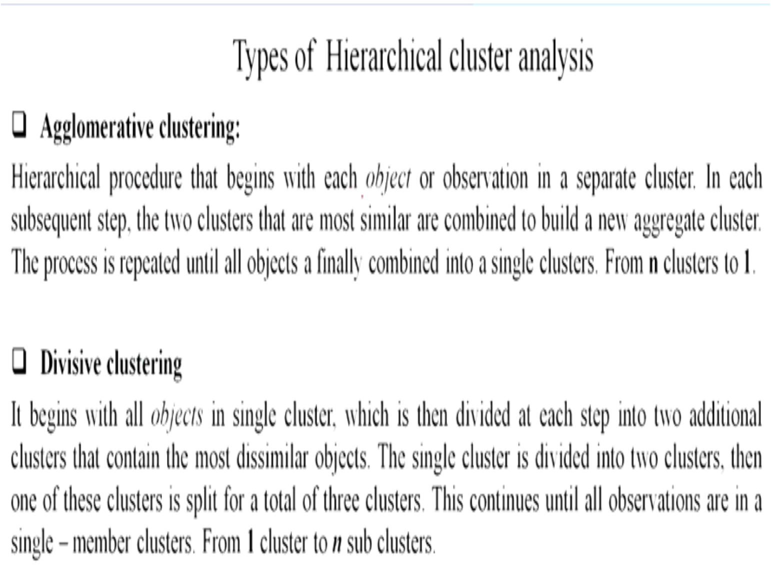
Hierarchical Cluster Analysis

- ❑ The stepwise procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics using an algorithm either agglomerative or divisive, resulting to a construction of a hierarchy or treelike structure (dendrogram) depicting the formation of clusters.
- ❑ Hierarchical cluster analysis are preferred when:
  - o The sample size is moderate (under 300 – 400, not exceeding 1000).

So let see the hierarchical cluster analysis here the stepwise procedure attempts to identify relatively homogenous groups of cases you remember this is basically the cluster analysis is done on cases, as factor analysis is done on variables, it is done on cases that means the respondents. Cases based on selected characteristics using an algorithm either a agglomerative or divisive. Agglomerative means combination, combing.

Divisive means breaking it up resulting to a construction of hierarchy or tree like structure depicting the formation of clusters. It is preferred when sample size is moderate under 3000-400 and not exceeding 1000. So hierarchical clustering is preferable when you have a moderate sample size.

(Refer Slide Time: 05:16)



Types of Hierarchical cluster analysis

❑ **Agglomerative clustering:**

Hierarchical procedure that begins with each *object* or observation in a separate cluster. In each subsequent step, the two clusters that are most similar are combined to build a new aggregate cluster. The process is repeated until all objects a finally combined into a single clusters. From **n** clusters to **1**.

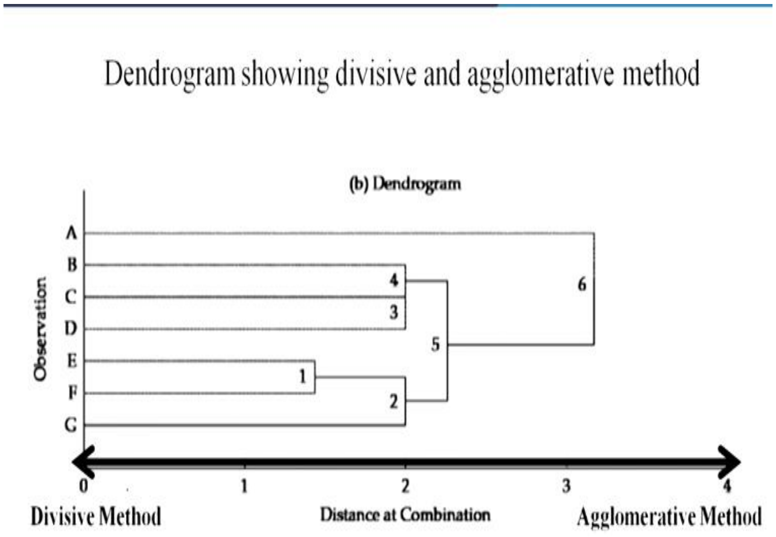
❑ **Divisive clustering**

It begins with all *objects* in single cluster, which is then divided at each step into two additional clusters that contain the most dissimilar objects. The single cluster is divided into two clusters, then one of these clusters is split for a total of three clusters. This continues until all observations are in a single – member clusters. From **1** cluster to ***n*** sub clusters.

So as I was saying there are 2 methods agglomerative and divisive. It begins with each object or observation in a separate clusters in each subsequent step, the 2 clusters that are more similar or combined to build a new cluster and this process goes on and on and on till you have a single cluster, from n cluster to 1 cluster. Divisive it begins with all objects in a single cluster, which is then divided into each step into 2 additional clusters that contain the most dissimilar objects.

So single cluster is divided into 2 clusters then 1 of this cluster is split for total of 3 clusters. This continues until all observations are in a single member clusters from 1 cluster to n subclusters. This is the opposite way.

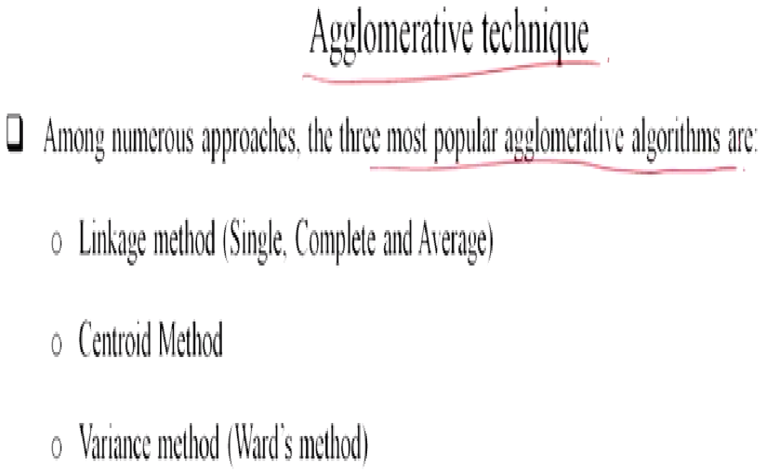
(Refer Slide Time: 06:02)



Dendrogram showing divisive and agglomerative method

So this is how it looks. So this is the divisive method. If you go this side from 6-5, 4, 3, 2, 1 and if you go this way that means this side it is the agglomerative way.

(Refer Slide Time: 06:18)

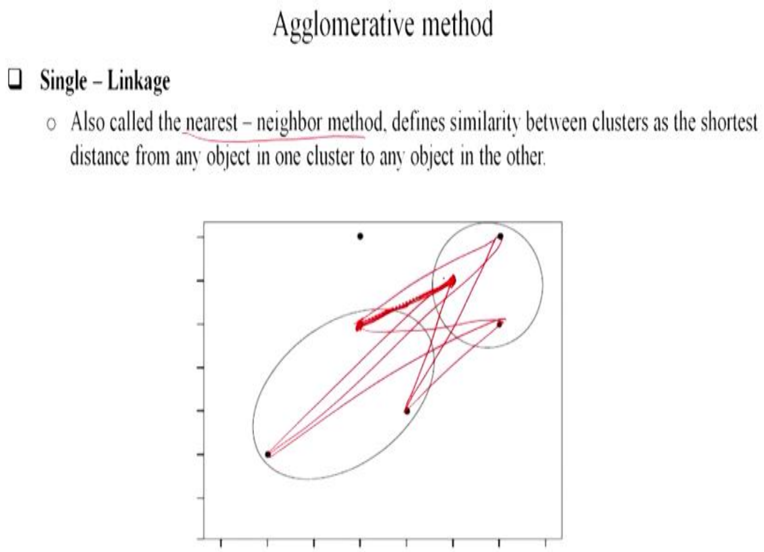


Agglomerative technique

- Among numerous approaches, the three most popular agglomerative algorithms are:
  - Linkage method (Single, Complete and Average)
  - Centroid Method
  - Variance method (Ward's method)

There are different types of methods when you talk about the agglomerative technique. So the 3 most popular agglomerative algorithms are linkage method, which has I will tell you what is that, centroid method and variance method, right. So when you use any software for example, SPS or something, it will ask you which approach do you want to use. So a single linkage method, a complete linkage method and average linkage what is that let see and then we will look at each one of them.

(Refer Slide Time: 06:45)



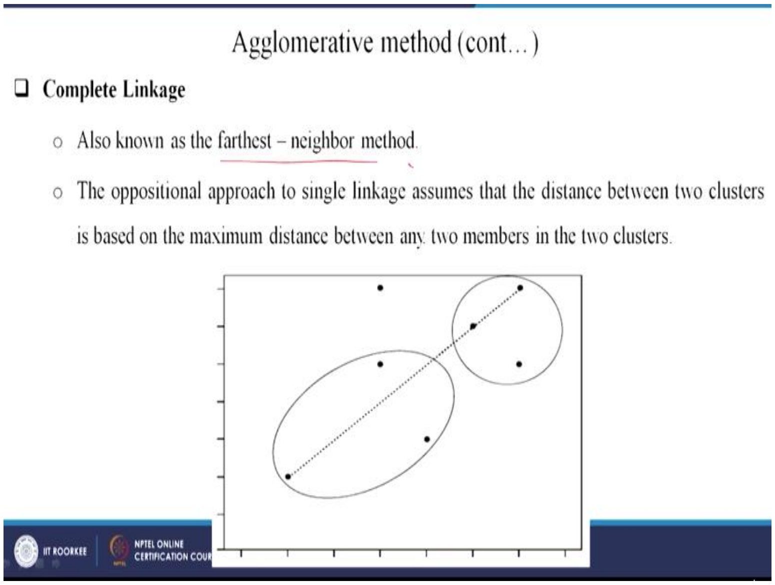
Agglomerative method

- **Single – Linkage**
  - Also called the nearest – neighbor method, defines similarity between clusters as the shortest distance from any object in one cluster to any object in the other.

It is also called the nearest neighbor method. Define similarity between clusters as the shortest distance from any object in one cluster to any object in the other. For example, this is one cluster and this is another cluster. Now there are several data points. You see this is the distance. This is the distance. This is the distance. Similarly, from here, if you this is one, this is one, this is one. Similarly, this is one, this is one, this is one, but which is the closest or the nearest point.

This one, so this one is the closest. So the most bold looking one. So now what it does is, it tries to add the clusters on basis of the nearest point taking the shortest points.

(Refer Slide Time: 07:35)



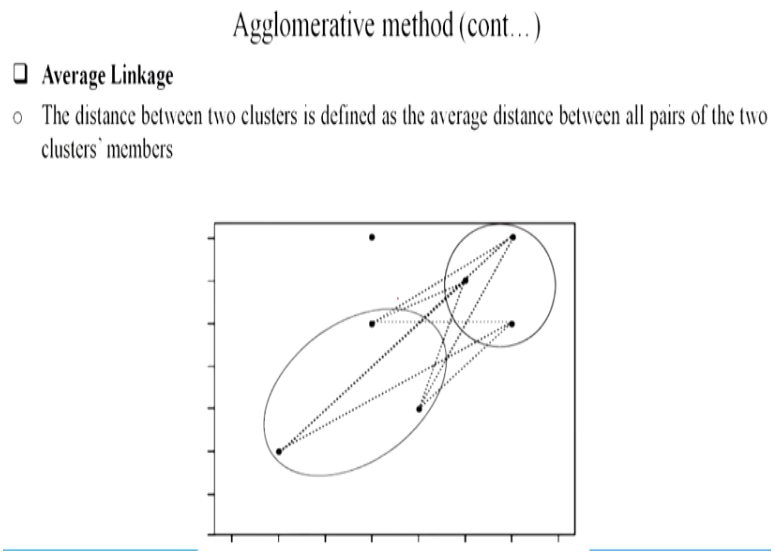
Agglomerative method (cont…)

❑ **Complete Linkage**

- Also known as the farthest – neighbor method.
- The oppositional approach to single linkage assumes that the distance between two clusters is based on the maximum distance between any two members in the two clusters.

Complete linkage is just opposite. It takes the data on base of the farthest neighbor. So the oppositional approach is to single linkage assumes that the distance between two clusters is based on the maximum distance between the 2 members in the 2 clusters. Third is the average. Now what is the average?

(Refer Slide Time: 07:57)



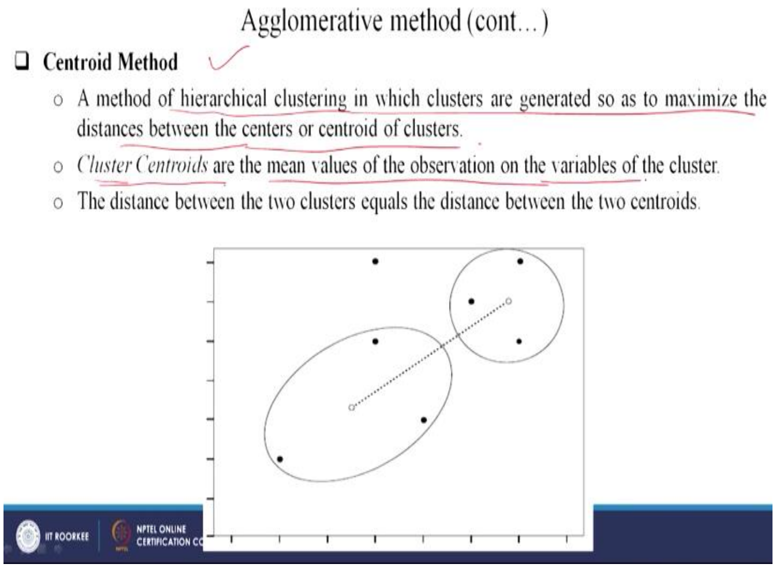
Agglomerative method (cont…)

❑ **Average Linkage**

- The distance between two clusters is defined as the average distance between all pairs of the two clusters` members

It identifies all the different points and adds them and takes the average distance between all pairs of the 2 cluster members. So that is why if you have large data points, you know hierarchical clustering becomes very difficult, because so many combinations will come, so many calculations have to be done. Then the next method is after seeing linkage, we have the centroid method.

(Refer Slide Time: 08:26)



Agglomerative method (cont…)

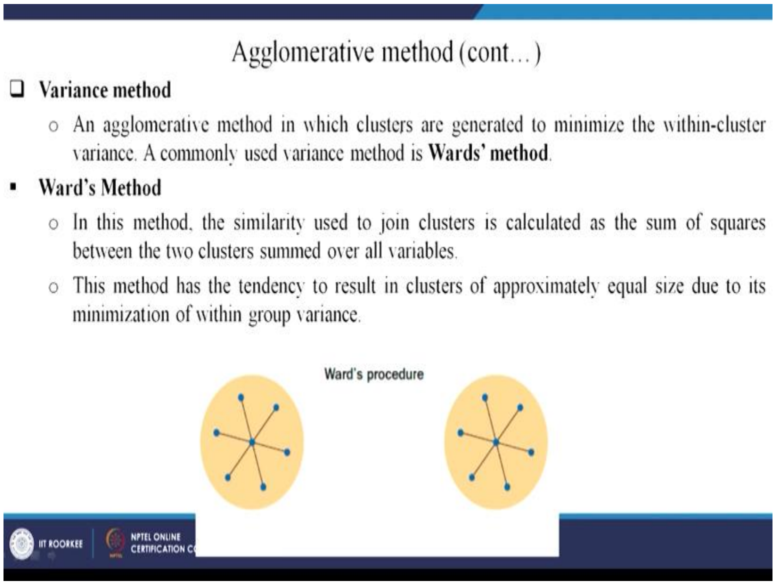
❑ **Centroid Method**

- A method of hierarchical clustering in which clusters are generated so as to maximize the distances between the centers or centroid of clusters.
- *Cluster Centroids* are the mean values of the observation on the variables of the cluster.
- The distance between the two clusters equals the distance between the two centroids.

A method of clustering, hierarchal clustering in which clusters are generated so as to maximize the distance between the centers or centroid of clusters. What are the centroids? Cluster centroids are nothing, but the mean values of the observations on the variables of the cluster. So we take the mean value. The distance between two clusters equals the distance between the two centroids. So it takes the mean value and finds the centroid, right.

For this cluster, this is the centriod or the mean point and then it takes the distance between these 2 points.

(Refer Slide Time: 09:00)



Agglomerative method (cont…)

❑ **Variance method**

- An agglomerative method in which clusters are generated to minimize the within-cluster variance. A commonly used variance method is **Wards' method**.

▪ **Ward's Method**

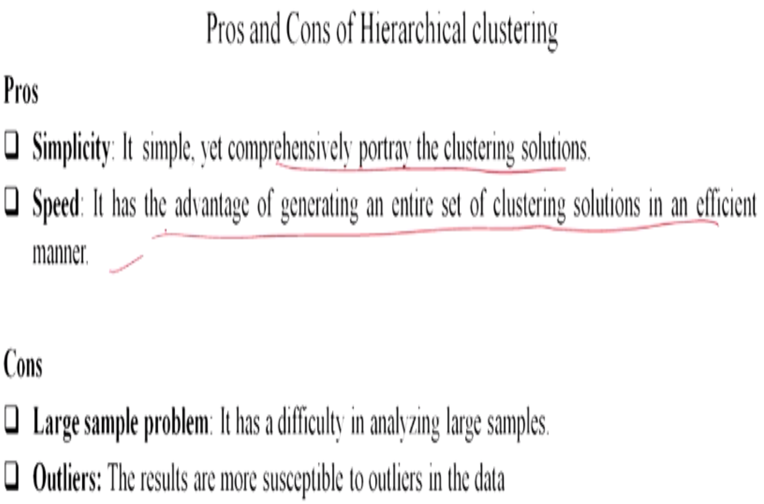
- In this method, the similarity used to join clusters is calculated as the sum of squares between the two clusters summed over all variables.
- This method has the tendency to result in clusters of approximately equal size due to its minimization of within group variance.

The third method is the variance method. What is this method doing? In this method, the clusters are generated to minimize the within cluster variants. So you know about within cluster variance means, so within the cluster you take all the distances and you measure the variance and then find out the minimum variance. So here the method used is called the Ward's method. In this method, the similarity used to join clusters is calculated as the sum of squares between the two clusters summed over all the variables.

This method has the tendency to result in clusters of approximately equal size due to its minimization of the within and you know the problem again the Ward's method makes it very complicated sometimes because of the calculation, within the software become very complicated because large number of calculations are to be done. So that is the only, otherwise for you and me it does not matter because we only change the output.

But for the computer, it takes lot of exercise. So let us do one. This is the pros and cons of the hierarchal clustering.

(Refer Slide Time: 10:08)



Pros and Cons of Hierarchical clustering

Pros

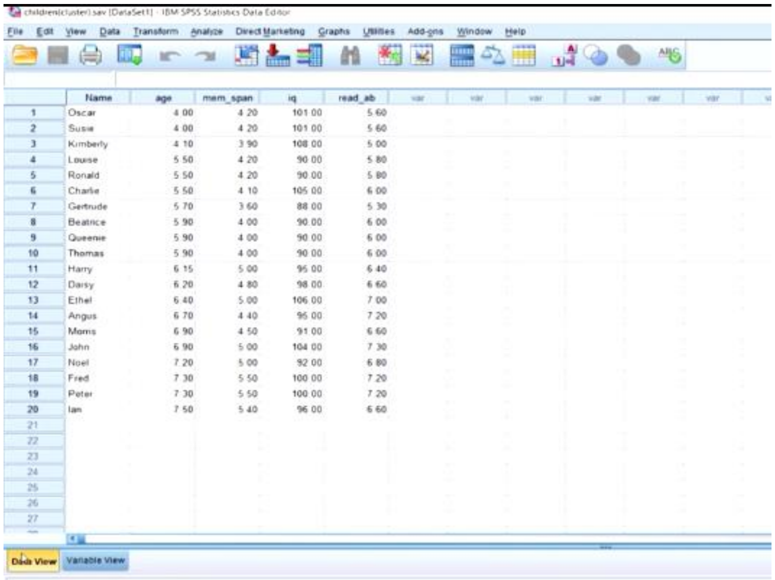
- ❑ **Simplicity**: It simple, yet comprehensively portray the clustering solutions.
- ❑ **Speed**: It has the advantage of generating an entire set of clustering solutions in an efficient manner.

Cons

- ❑ **Large sample problem**: It has a difficulty in analyzing large samples.
- ❑ **Outliers:** The results are more susceptible to outliers in the data

For example, it is simple and yet portrays the clustering solutions, speed advantage of generating an entire set of clustering solution in efficient manner, right. But what is the problem, difficulty in analyzing large samples. That is what I was just saying and the results are most susceptible to out layers. If there are out layers, then it will have a problem. So let us take one case. This is the case which I have brought.

(Refer Slide Time: 10:32)



	Name	age	mem_span	iq	read_ab	var	var	var	var	var	var
1	Oscar	4.00	4.20	101.00	5.60						
2	Susie	4.00	4.20	101.00	5.60						
3	Kimberly	4.10	3.90	108.00	5.00						
4	Louise	5.50	4.20	90.00	5.80						
5	Ronald	5.50	4.20	90.00	5.80						
6	Charlie	5.50	4.10	105.00	6.00						
7	Gertrude	5.70	3.60	88.00	5.30						
8	Beatrice	5.90	4.00	90.00	6.00						
9	Queenie	5.90	4.00	90.00	6.00						
10	Thomas	5.90	4.00	90.00	6.00						
11	Harry	6.15	5.00	95.00	6.40						
12	Daisy	6.20	4.80	98.00	6.60						
13	Ethel	6.40	5.00	106.00	7.00						
14	Angus	6.70	4.40	95.00	7.20						
15	Morris	6.90	4.50	91.00	6.60						
16	John	6.90	5.00	104.00	7.30						
17	Noel	7.20	5.00	92.00	6.80						
18	Fred	7.30	5.50	100.00	7.20						
19	Peter	7.30	5.50	100.00	7.20						
20	Ian	7.50	5.40	96.00	6.60						
21											
22											
23											
24											
25											
26											
27											

So this is the, there are different names of different children. Their age is given, memory span is given, I/Q is given, reading ability is given. This I think we have also used during regression, the same data set, we have used. So now suppose I want to do a hierarchical clustering. So I want to see how many clusters can be formed out of it. First question, so to do that, first of all what do you do is, I will explain. Go to classify and you see hierarchical clustering here.

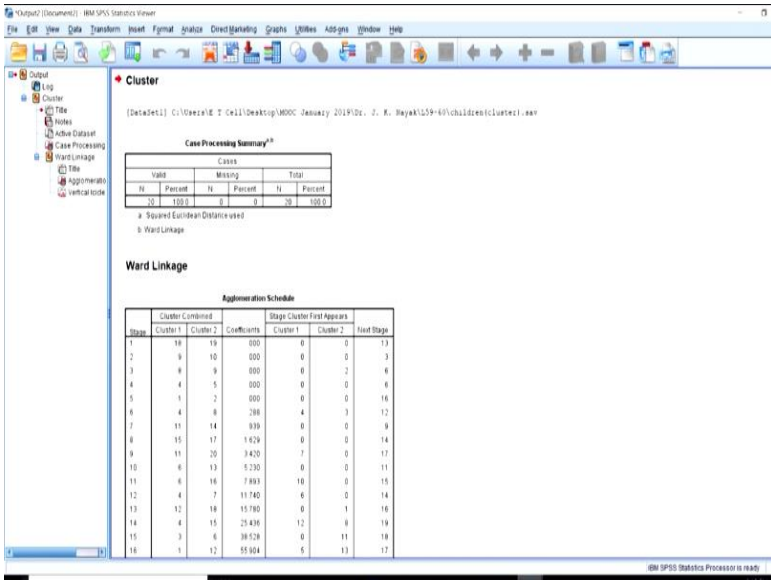
So when you go to hierarchical clustering, take the data points that you are interested in the variables, right. Now what do you want. Now go to statistics, here the agglomeration schedule is by default there, so no issues, continue. I am never interested in looking at the dentogram because I can find it through the agglomeration schedule, so I do not use the dentogram generally. So you can use that, but then it looks very odd sometimes, it is very complicated.

And I do not feel it is a required even. So what is the method you would use. So there are different methods. So generally one of the best method is the Ward's method and how do you want to measure the distance, the squared Euclidean distance this is by default. This is now important, now you see, if suppose do you want to standardize, it is giving an option, you need to standardize when your data are measured in different kinds of scales.

If they are not measured in different kind of scales, you did not have any problem, but if you have the suppose some in categorical, some in you kind of a continuous or such kind of scale issues are there, then you will have to just standardize. You do not have to do anything; you just have to go for these 2 options. Either z score or range between -1 to +1. So if you give any of the them, it will be standardized and then the scale issues are automatically gone.

So I am not doing it here, right. I will show you if you want to do in another case may be. So do you want to save the value, do you save it, well at the moment no, because it is just I want to find the number of clusters, so nothing here.

(Refer Slide Time: 12:55)



Cluster

```
[DataSet1] C:\Users\E T Cell\Desktop\MOOC January 2019\Dr. J. K. Nayak\L59-60\children(cluster).sav
```

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
20	100.0	0	.0	20	100.0

a. Squared Euclidean Distance used

b. Ward Linkage

Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	18	19	.000	0	0	13
2	9	10	.000	0	0	3
3	8	9	.000	0	2	6
4	4	5	.000	0	0	6
5	1	2	.000	0	0	16
6	4	8	.288	4	3	12
7	11	14	.939	0	0	9
8	15	17	1.629	0	0	14
9	11	20	3.420	7	0	17
10	6	13	5.230	0	0	11
11	6	16	7.893	10	0	15
12	4	7	11.740	6	0	14
13	12	18	15.780	0	1	16
14	4	15	25.436	12	8	19
15	3	6	38.528	0	11	18
16	1	12	55.904	5	13	17

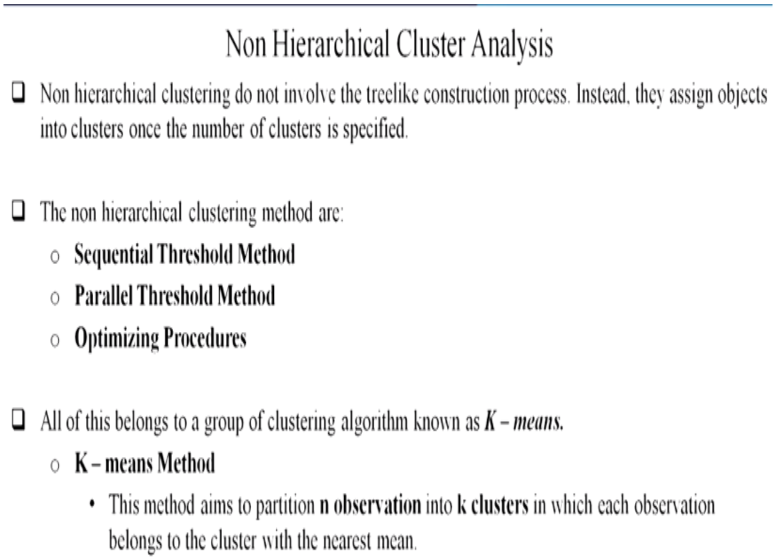
So let us see if you look at it. See this data is a very small data just for in class purpose I have brought it. Now look at the agglomeration schedule. How many clusters to be formed you can get it from here. Now for example look at the last value, go from backwards 794.185, the coefficient value right. Then the next one is 249, so there is a significant difference in the value, there is a great fall, if you come from this way great jump.

Here also there is a significant difference, here also to some extent here, so you up to that point you go till you feel there is a serious or a big difference between the 2 values, right. Suppose but remember 1 thing, if you create too many clusters, it is not very good. Why because the characteristics of the clusters will be very thin and there will be a very little difference between them sometimes very much possible. If you create only one cluster then there is no point.

If there is only one cluster, there is no point of doing a cluster analysis. So you should be reasonable enough to understand how many clusters to be formed. For example, in this case, I will take 1, 2, 3, and 4. Let us say 4 clusters and see whether there is any significant difference. So this part of understanding and identifying how many clusters are to be done, is done through this agglomeration schedule.

So you just have to look at these values and see till what point, there is a substantial change in the values and when you find the differences now not much, then you stop there that is what it does.

(Refer Slide Time: 14:37)



Non Hierarchical Cluster Analysis

- ❑ Non hierarchical clustering do not involve the treelike construction process. Instead, they assign objects into clusters once the number of clusters is specified.

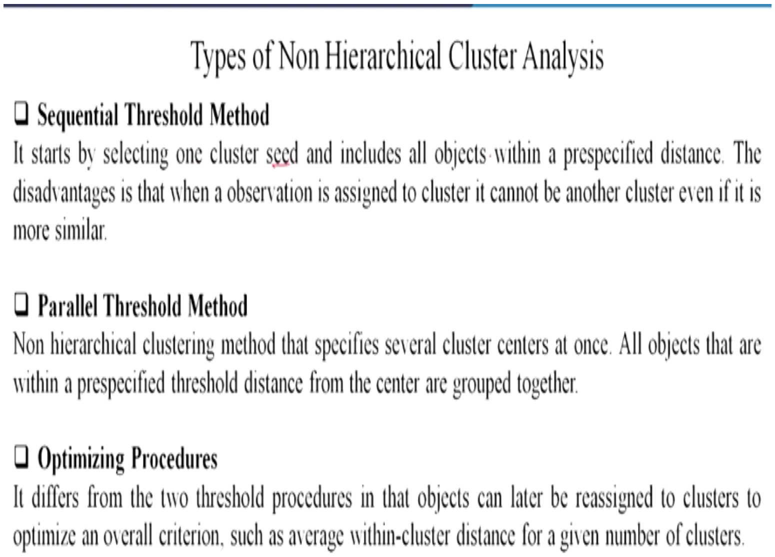
- ❑ The non hierarchical clustering method are:
  - o **Sequential Threshold Method**
  - o **Parallel Threshold Method**
  - o **Optimizing Procedures**

- ❑ All of this belongs to a group of clustering algorithm known as ***K – means.***
  - o **K – means Method**
    - This method aims to partition **n observation** into **k clusters** in which each observation belongs to the cluster with the nearest mean.

In the non-hierarchical cluster analysis, so after you have identified the number of clusters the next part. What it does is, the non-hierarchical cluster does not involve the tree like construction process instead, they assign objects into clusters once the number of clusters is specified. So you have to specify, here during the non hierarchical, you have to specify the number of seed points we say or the number of clusters. The non-hierarchical clustering methods are, basically there are 3 types basically.

But we are more bothered about this K-means method. So if you go to a spaces or anywhere, you will find that instead of non-hierarchical clustering, there is no name called non-hierarchical clustering only you find a K-means because it is almost all the time that the K-means clustering represents the non-hierarchical method. There are 3 methods as you saw here. So what are they?

(Refer Slide Time: 15:29)



Types of Non Hierarchical Cluster Analysis

❑ **Sequential Threshold Method**

It starts by selecting one cluster seed and includes all objects within a prespecified distance. The disadvantages is that when a observation is assigned to cluster it cannot be another cluster even if it is more similar.

❑ **Parallel Threshold Method**

Non hierarchical clustering method that specifies several cluster centers at once. All objects that are within a prespecified threshold distance from the center are grouped together.

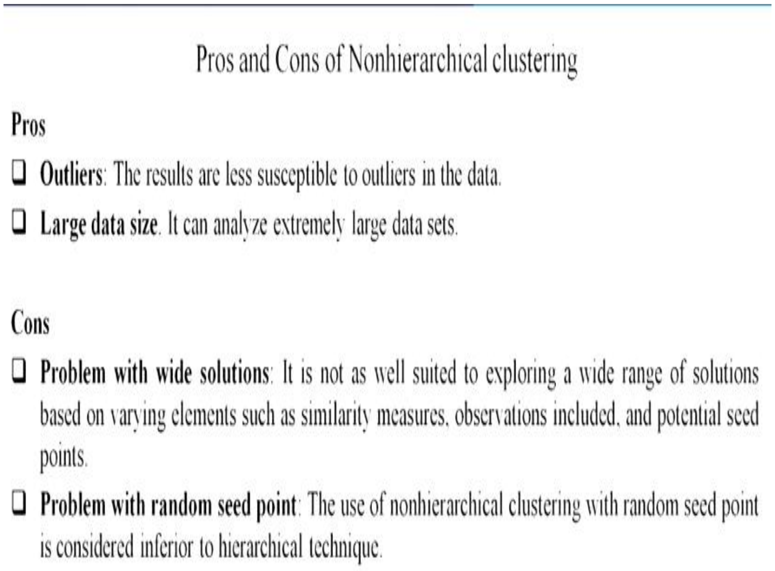
❑ **Optimizing Procedures**

It differs from the two threshold procedures in that objects can later be reassigned to clusters to optimize an overall criterion, such as average within-cluster distance for a given number of clusters.

It starts by selecting one cluster, the sequential method, one cluster seed and includes all objects within a prespecified distance because the whole measure is distance. The disadvantage is that when the observation is assigned to 1 cluster, it cannot be assigned to another cluster even if it is more similar. So that is one problem with the sequential. Parallel threshold method, this is the non-hierarchical method that specifies several cluster centers at one point more like a divisive method.

All objects that are within the pre-specified threshold distance from the center are grouped. So these are the 3 and the optimizing procedures, which is the K-means, it differs from the 2 in that objects can later be reassigned to clusters to optimize an overall criteria such as average within cluster distance. So average within cluster distance is what we are talking about the Ward’s method. So our objective is to reduce this minimum within cluster distance.

(Refer Slide Time: 16:37)



Pros and Cons of Nonhierarchical clustering

Pros

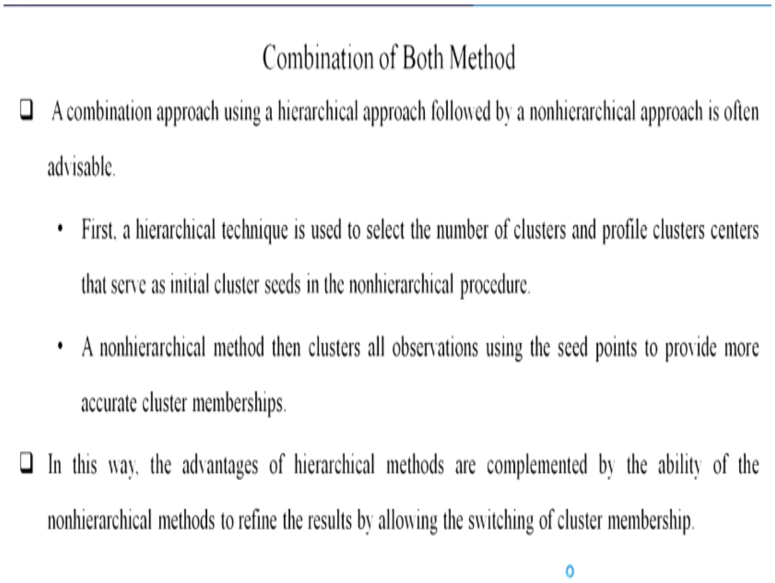
- □ **Outliers**: The results are less susceptible to outliers in the data.
- □ **Large data size**. It can analyze extremely large data sets.

Cons

- □ **Problem with wide solutions**: It is not as well suited to exploring a wide range of solutions based on varying elements such as similarity measures, observations included, and potential seed points.
- □ **Problem with random seed point**: The use of nonhierarchical clustering with random seed point is considered inferior to hierarchical technique.

So this is the some of the pros and cons you can check it later on.

(Refer Slide Time: 16:41)



Combination of Both Method

- □ A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable.
  - First, a hierarchical technique is used to select the number of clusters and profile clusters centers that serve as initial cluster seeds in the nonhierarchical procedure.
  - A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships.
- □ In this way, the advantages of hierarchical methods are complemented by the ability of the nonhierarchical methods to refine the results by allowing the switching of cluster membership.

We will use a combined approach. Now let us go back to the data set. Now once we have done it, we had suppose 3 or 4 clusters, now we will go back to the K-means cluster. Now again we take all the data with the variables to this side. Now I am interested to see whether this is okay. Do you want to save anything; do you want to save the particular respondent comes into which cluster membership. You can do that. You can even check the distance from the cluster center.

Continue and options, initial cluster center and cluster information continue and number of clusters here, I would like to make a change. So for example, I have gone with 4 clusters you may make it 3 or you may make it 5, there is no binding, it is the researchers own logic that has to be used. Now we have run it. Now you see, we said 4 clusters and 4 clusters have come, this is the initial cluster center and then it tells you the cluster membership.

So each case you see 1-20 respondents were there. Now first respondent is in cluster 1, second cluster 1, third cluster 3, four cluster 4, on basis of certain behavioural traits. Let us see what is that. So go to the final cluster centers. So there are 4 clusters, cluster 1, cluster 2, cluster 3, cluster 4. So what is the behavior of cluster 1. Let us see, cluster 1 is somebody whose mean age is 5.9, short term memory span is 4.88, I/Q is 101.2, reading ability 6.58.

Cluster 2 has got a higher age mean from cluster 1, but memory span is almost equal, I/Q is less than cluster 1, reading ability is slightly higher. Cluster 3 the mean of the age is quite less; short-term memory span is 4.33. Again it is quite less. I/Q is very high, and reading ability is poor because obviously we can connect that it is a young child because it is the child. So the age group is also less. The age minimum is less, so that is why the reading ability might be less also.

Fourth cluster is 6.06 again not very high, not very low, it is in between and look at the memory span, very, very poor, in comparison. I/Q is 90 the lowest and reading ability is again also the lesser side. Now once a company or a researcher identifies these clusters, the behavior of the cluster is also known to you. Now suppose let us say had it been the case of a marketer and the marketer knows what is the age group of the people who are using his products.

What is their memory span, what is their I/Q level, what is their reading ability? Suppose a bookseller for example, now he can say well I want to target, you cannot target all the groups right, because that is why segmentation is done, so that you can choose you target segment. Now out of this 4 maybe I will choose only one or two maybe three or whatever it is. So let us say, he wants to select 2 clusters out of it. So which 2 clusters will you select.

The question is you can see from the distance between the final cluster centers. Now cluster 1 and cluster 2, 5.525, 1 and 3, 5.2, 1 and 4, 11. So 1 and 4 are really far off from each other, that means there is a very huge distance between the two clusters. Cluster 1, 2, 1, 3, 1, 4, so 2, 3. Now

2 and let us say 3 10.4 again it is far off, 2 and 4 is 5. Cluster 3 and cluster 4 very far off. So we can see here the cluster 2 and cluster 4 seems to be the one which is very close.

My sample size is very less, but when you do, you can use it on a larger sample size. So by understanding this, the marketer can understand the cluster 2 and cluster 4 have some similarity and they can target maybe that particular cluster. So this is only make you understand to realize how it works, how do you use it, but then you use your logic and see what is to be done and what not is to be done. So this is how cluster analysis works.

So your combined approach is what I just noted, first through the hierarchical clustering you find the selected number of clusters and serve as the initial clusters sits for the non hierarchical and the non hierarchical clusters, then clusters all observations using the seed point to provide more accurate cluster memberships. Now I wanted to show something here. Let us go to the data set. Now you see everything is given to you now.

Which this particular respondent falls into which cluster and what is the cluster distance. Now if you see the higher the distance that means farther and which member, suppose you want to know this guy Ronald in which cluster does it fall, in 4. Suppose you are interested in cluster 2 and 4 as the marketer. So 2 and 4 let us see who are the people Luzy, Ronald, Getrade, Betrese, Quinee, Thomas, Morris and Noel. So these are the people who are a part of your target segment.

So you do not want to invest your resources and waste your resources for the cluster 1 and cluster 3 and cluster maybe 2. I am only interested for cluster 4 let us say for the time being, just to understand it. So then my focus all my resources will be diverted towards cluster 4. So that is how it helps a marketer a lot, okay and obviously you can see from here, if the same logic applies to as I was giving a political example.

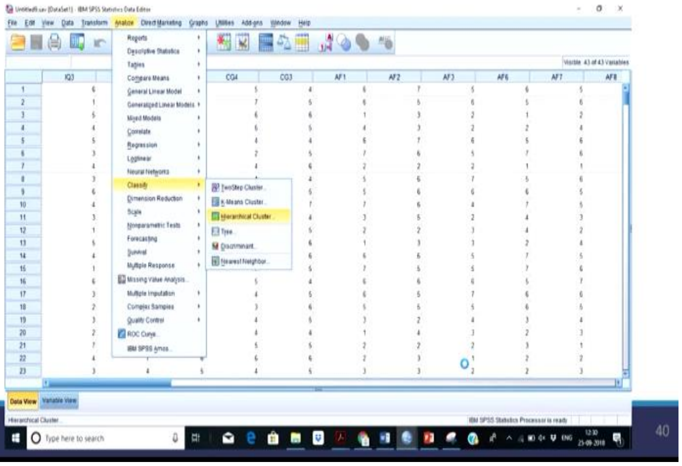
Now suppose, I know the cluster 1 is a, suppose, suppose I am just saying, is a very open minded person, cluster 3 is a very religious person, so when a politician goes to these 2 cities, city 1 is the very open minded city, people have like you know very modern and modern thinking approach. Cluster 3 is the very traditional religious kind of people. So the speech of the politician will differ in both the cities.

So that is why it is very important to understand what kind of cluster does the city fall to or the person to or the product fall to. So this is very important to understand. So this is what we have done. So the steps are all mentioned over here.

(Refer Slide Time: 24:00)

Hierarchical Cluster Analysis in SPSS steps

Step-1 Go to analyze > classify > hierarchical cluster.

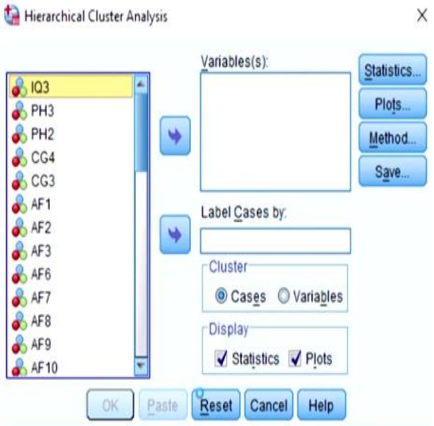


So now it will help you to understand how to proceed with the cluster analysis. So once you can do this, this steps I have already done it.

(Refer Slide Time: 24:11)

Hierarchical Cluster Analysis in SPSS steps (cont…)

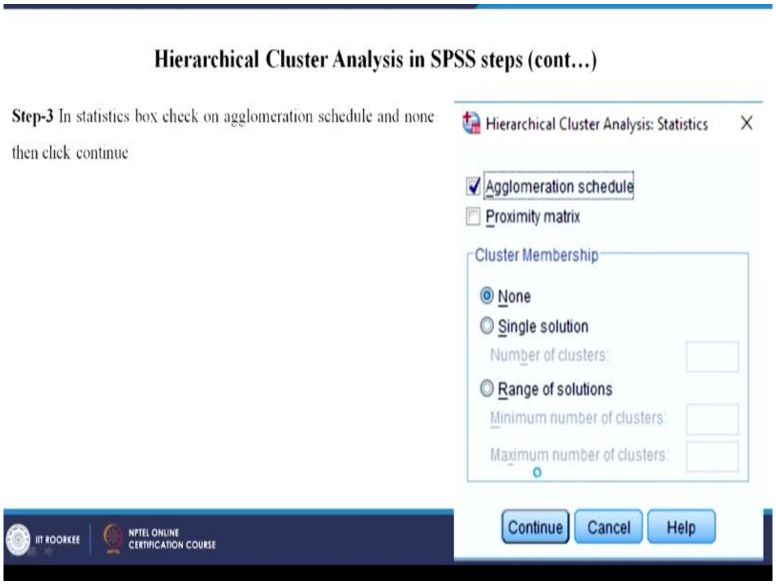
Step-2 A window of hierarchical cluster analysis will open. Then select the variables upon which we base our clusters and place them into variables section. Check the statistics and Plots boxes. Then click on statistics.



Cluster
● Cases ○ Variables

Display
☑ Statistics ☑ Plots

(Refer Slide Time: 24:13)



Hierarchical Cluster Analysis in SPSS steps (cont…)

Step-3 In statistics box check on agglomeration schedule and none then click continue

Hierarchical Cluster Analysis: Statistics

☑ Agglomeration schedule

□ Proximity matrix

Cluster Membership

● None

○ Single solution

Number of clusters:

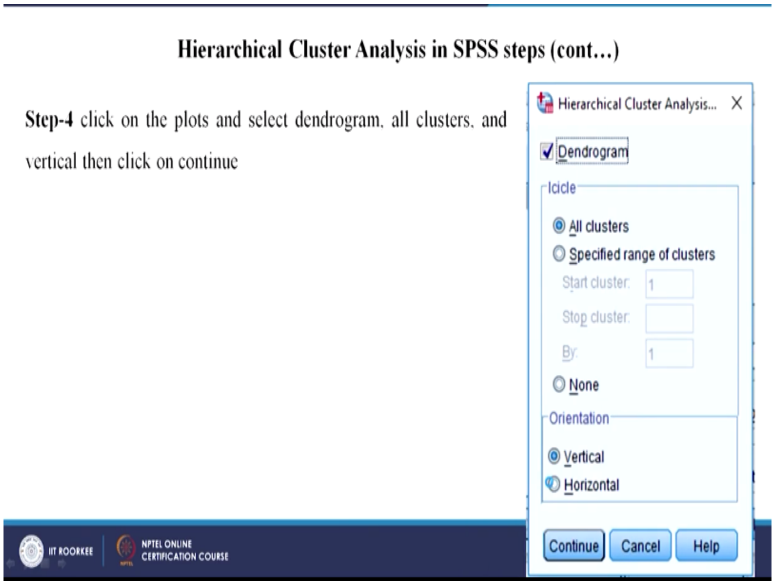
○ Range of solutions

Minimum number of clusters:

Maximum number of clusters:

And this is for your reference only how you do.

(Refer Slide Time: 24:14)



Hierarchical Cluster Analysis in SPSS steps (cont…)

Step-4 click on the plots and select dendrogram, all clusters, and vertical then click on continue

Hierarchical Cluster Analysis...

☑ Dendrogram

Icicle

● All clusters

○ Specified range of clusters

Start cluster: 1

Stop cluster:

By: 1

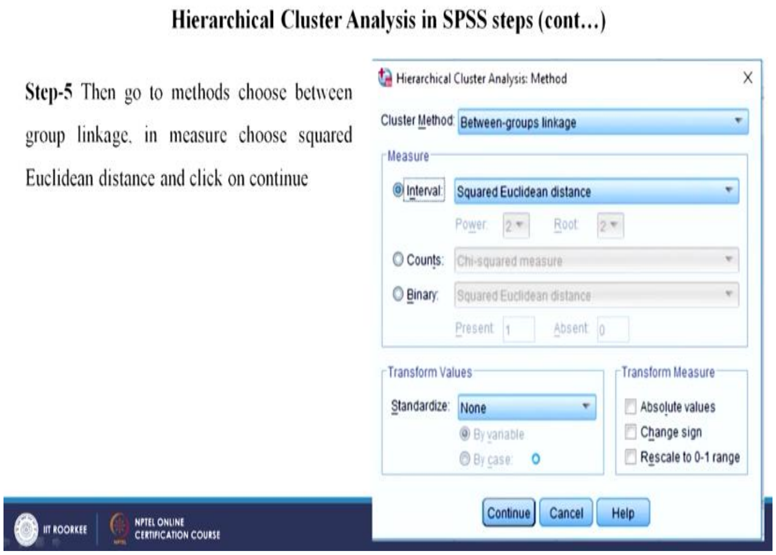
○ None

Orientation

● Vertical

○ Horizontal

(Refer Slide Time: 24:15)



Hierarchical Cluster Analysis in SPSS steps (cont…)

Step-5 Then go to methods choose between group linkage, in measure choose squared Euclidean distance and click on continue

Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure

● Interval: Squared Euclidean distance

Power: 2 Root: 2

○ Counts: Chi-squared measure

○ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: None

● By variable

○ By case:

Transform Measure

□ Absolute values

□ Change sign

□ Rescale to 0-1 range

And after you do, then you can finally write your cluster analysis report, the interpretation. For example, you may write that we did a cluster analysis and we first conducted hierarchical clustering analysis and we found 3 or 4 clusters and then after that we did K-means clustering and we found the behavior of the cluster. There are 4 clusters for example, the cluster 1 and write down the behavior of the traits of the cluster, entire.

So the age group was this, the memory span was this and this and this and similarly cluster 2, cluster 3 and cluster 4 and then you may give your opinion. Even you should do one thing, please remember that kindly give the names of the clusters. For example, cluster 1, now according to the behavior of the cluster, you can given a name to that particular cluster maybe young and very intelligent students. Second impulsive, but very sharp I/Q.

What you can do is, you can give a name to this clusters. So that will be very helpful for others to understand what is the basically the trait of this particular cluster. So this is how you do a cluster analysis and you interpret it. I think this should be extremely helpful for you in the future and if you use it well, you can really make good publications out of it. So wish you all and thank you so much.