Cluster Analysis (CA)

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Cluster Analysis has been used in marketing for various purposes. Segmentation of consumers in cluster analysis is used on the basis of benefits sought from the purchase of the product. It can be used to identify homogeneous groups of buyers.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

The variables on which the cluster analysis is to be done should be selected by keeping past research in mind. It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher. An appropriate measure of distance or similarity should be selected; the most commonly used measure is the Euclidean distance or its square.

Distance Measure

Mahalanobis Distance or Euclidean Distance (with quantitative measurements)

$$d(X_r , X_s) = d_{rs} = \sqrt{(Xr - Xs)' \sum^{-1} (Xr - Xs)}$$

Clustering procedures

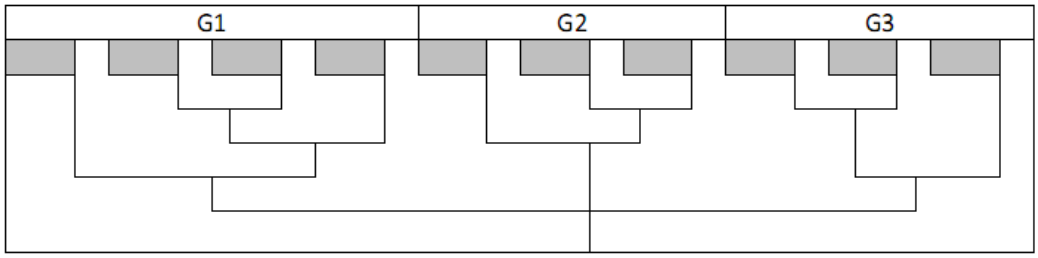
Clustering procedures in cluster analysis may be hierarchical, non-hierarchical, or a two-step procedure. A hierarchical procedure in cluster analysis is characterized by the development of a tree like structure. A hierarchical procedure

can be agglomerative or divisive. Agglomerative methods in cluster analysis consist of linkage methods, variance methods, and centroid methods. Linkage methods in cluster analysis are comprised of single linkage, complete linkage, and average linkage.

The non-hierarchical methods in cluster analysis are frequently referred to as K means clustering. The two-step procedure can automatically determine the optimal number of clusters by comparing the values of model choice criteria across different clustering solutions. The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful. The clusters should be interpreted in terms of cluster centroids.

There are certain concepts and statistics associated with cluster analysis:

- Agglomeration schedule in cluster analysis gives information on the objects or cases being combined at each stage of the hierarchical clustering process.
- Cluster Centroid is the mean value of a variable for all the cases or objects in a particular cluster.
- A dendrogram is a graphical device for displaying cluster results.



- Distances between cluster centers in cluster analysis indicate how separated the individual pairs of clusters are. The clusters that are widely separated are distinct and therefore desirable.

- Similarity/distance coefficient matrix in cluster analysis is a lower triangle matrix containing pairwise distances between objects or cases.
- The number of cases (sample size) and the number of variables used is expected to be correlated, as large numbers of variables (high data dimensionality) require large data sets. Due to a lack of rules, the only recommendation that can be given concerning sample sizes and variable numbers is to critically question if the dimensionality is not too high for the number of cases to be grouped. Formann (1984) suggests the minimal sample size to include no less than 2^pcases (p = number of variables), preferably $5 * 2^p$.

Steps of Single Linkage method (Nearest neighbor method)

- Start with n of the clusters where each cluster includes one observation.
- Combining the two nearest points according to one of the adopted distance measures.
- The adoption of the spacing between this new cluster and any other point as the smallest distance between these two points in the cluster and this other point.
- Continuing to merge the closer clusters to each other, thus the number of clusters will decrease by one with each step. The spacing between any two clusters represents the distance between the two closest clusters.
- Therefore, this method begins with n of the clusters, where each cluster includes one observation, and we continue to merge the points and clusters until the process ends with one cluster that includes all the observations or points.
- The appropriate number of clusters is based on their number at the beginning and their number at the end. There are several methods of selecting the number of clusters, including a logical view in this regard. One of these methods that helps in doing this is building a hierarchical tree shape.