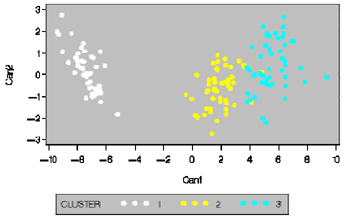
Cluster Analysis



- How should the objects for analysis be selected?
- 2. Which variables should be used to describe the objects?
- 3. Should any standardization or differential weighting of variables be undertaken?

- 4. How should a relevant measure of dissimilarity be constructed from a pattern matrix?
- 5. Which clustering and graphical procedures should be used in the analysis of data?
- 6. How should the results of the study be validated and summarized?

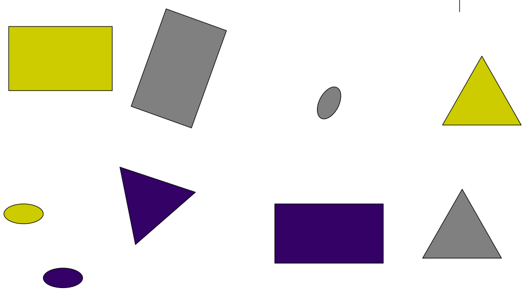
- deals with classification
- classification - the grouping of objects into classes, based on the same similarity of properties or by relationships between the objects.
- object - an individual with attributes

- Cluster analysis is not as much a typical statistical test as it is a "collection" of different algorithms that "put objects into clusters."
- The point here is that, unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any *a priori* hypotheses, but are still in the exploratory phase of our research

principles of classification

- 1) Classifications should be designed for a specific purpose; they rarely serve two purposes equally well. Purpose and use must be linked.
- 2) The classification of any group of objects should be based upon properties which are properties of the objects; it follows that differentiating characteristics should be properties of the objects classed.
- 3) The differentiating characteristics must be important for the purpose of classification or else the classification is trivial.

Classify these objects



- 4) Classifications are not final and must be changed as more knowledge is gained about the objects.
- 5) Classification should proceed at every stage and as far as possible on one principle. If this principle cannot be used for the entire classification, the properties used at the higher class must be more important than those used in lower classes.

- Classification methods may be grouped into two categories :
- 1) The testing of a Priori Classification.
  - i.e. X^2, ANOVA, discriminant analysis.
- 2) The Development of a Classification.
  - i.e. Factor analysis, clustering procedures.
- Cluster analysis is a general term for multivariate techniques that find groups or clusters of similar objects. Typically they use measures of similarity to determine if two objects should be fused into a group.

- The basic procedure is similar for most hierarchical agglomerative procedures.
- 1) Determine a matrix of similarity between pairs of objects. Each measurement scale has a number of similarity measures

- The most common measure for the metric scales is the square Euclidean distance
  - with X_{ij}, X_{ik} are the data matrix values of attribute *i* for object *j* and *k*.

$$d_{jk} = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2$$

- 2) Find the smallest d_{jk} and combine objects *j* and *k* into the same group.
- 3) A new similarity matrix is constructed for the remaining groups. It is in this step that the most common procedures differ. Each has its own method of determining the new distances.
- 4) Search the new matrix and find the smallest distance or greatest similarity value. Combine the groups defining that distance.
- 5) Continue until only one group exists.

There are six common approaches to clustering :

- 1) Nearest Neighbour : (also known as single linkage clustering) The distance between groups is the minimum distance between any pair of members of the two groups. This technique is prone to chaining - the appearance of elongated groups.

- 2) Furthest neighbour : (also known as complete linkage clustering) the distance between groups is the distance between the most remote pair. The method generally leads to tight, hyper-spherical discrete clusters.

- 3) Centroid : The distance between two groups is the distance between their centroids. The procedure does not yield monotonic results. The procedure suffers from reversals in the joining of objects (clusters).
  - A reversal occurs when an object joins a cluster after the cluster has formed, but joins at a higher similarity level than at which the cluster formed

- 4) Median : (Also known as the weighted pair - group centroid method) the distance between groups is the distance between the centroids, where the centroid is defined as the midway point between the centroids of the two clusters that fused to create a given cluster. This method also does not guarantee monotonic results.

- 5) Group average : (Also known as the unweighted pair - group method using arithmetic averages) the distance between groups is the average between all pairs of the members of two clusters. It is probably the most used technique.

- 6) Ward’s method : The technique uses the within group sum of squares. That is, the sum of squares of the distances from each cluster member to its parent cluster mean.
- Related to ANOVA

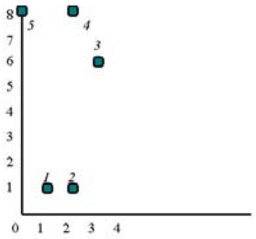
Similarity Measures

- 1) Correlation Coefficient
- 2) Distance Measures
- 3) Association Coefficients, i.e. Jaccard's (similarity ratio), Gower's.

Example: Using centroid cluster analysis.

		Variable	
		1	2
Individual	1	1.0	1.0
	2	1.0	2.0
	3	6.0	3.0
	4	8.0	2.0
	5	8.0	0.0

- We'll assume the data has been standardized



Distance Matrix is

	1	2	3	4	5
1	0	1	29	50	50
2	1	0	26	49	53
3	29	26	0	5	13
4	50	49	5	0	4
5	50	53	13	4	0

matrix is squared Euclidean distance

- We now fuse the two individuals who are the closest then we get :

	1	2
(12)	1	1.5 →centroid of group
3	6	3
4	8	2
5	8	0

	(12)	3	4	5
(12)	0	27.25	49.25	51.25
3	27.25	0	5	13
4	49.25	5	0	4
5	51.25	93	4	0

- we fuse the closest

	1	2
(12)	1	1.5
3	6	3.0
(45)	8	1.0

	(12)	3	(45)
(12)	0.0	27.25	49.25
3	27.25	0.00	8.00
(45)	49.25	8.0	0.00

- Three would be fused into group (45).

K means clustering

- Suppose that you already have hypotheses concerning the number of clusters in your cases or variables.
- You may want to "tell" the computer to form exactly 3 clusters that are to be as distinct as possible.
- This is the type of research question that can be addressed by the k- means clustering. In general, the k-means method will produce exactly *k* different clusters of greatest possible distinction.

- Computationally, you may think of this method as analysis of variance "in reverse."
- The program will start with *k* random clusters, and then move objects between those clusters with the goal to (1) minimize variability within clusters and (2) maximize variability between clusters.
- In k-means clustering, the program tries to move objects (e.g., cases) in and out of groups (clusters) to get the most significant ANOVA results.