1 Cluster Analysis

Objective: create clusters of items, individuals or objects that have similarity with the others in the cluster but with differences between clusters.

The items, individuals or objects being placed into clusters will be referred to as cases. The degree of similarity or dissimilarity may be determined from the recorded values for one or multiple characteristics for the cases. There are no dependent variables for cluster analysis. Clustering procedures require that similarity be quantified. One quantitative measure for interval scale data is the distance between cases. Euclidean Distance measures the length of a straight line between two cases. The numeric value of the distance between cases depends on the measurement scale. If the measurements are recorded using different measurement scales then one should use a transformation to assure similar variability of measurements for all characteristics being used to create the clusters (see Transform Values section below tell how this can be done in SPSS Hierarchical Cluster Analysis, otherwise this needs to be done prior to the analysis, for example, when using SPSS K-Means Cluster Analysis). Other measures may be used to create a dissimilarity or distance matrix that can be used as the basis for creating clusters (see Measures for Interval Data section under Hierarchical Cluster Analysis).

A key issue in obtaining a set of clusters is the determination of the number of clusters. Hierarchical procedures provide information that allows the analyst to decide on the number of clusters based on the output. This is often done by examining tabular or graphical output to identify the gaps that define logical clusters. The SPSS K-Means Cluster Analysis procedure requires that the number of clusters be specified to run the analysis. The K-Means procedure is applicable for data sets with a large number of cases while the hierarchical procedure may be preferred when there are a limited number of cases.

Clustering Mathods

SPSS provides several alternative methods for determining a summary measure of distance when a cluster has multiple cases in the cluster. The SPSS options for the clustering method are Between-groups linkage, Within-groups linkage, Nearest neighbor, Furthest neighbor, Centroid clustering, Median clustering, and Ward's method.

- For the **nearest neighbor** or **single linkage** method, the dissimilarity between cluster A and cluster B is represented by the minimum of all possible distances between the cases in cluster A and the cases in cluster B.

- For the **furthest neighbor** or **complete linkage** method, the dissimilarity between cluster A and cluster B is represented by the maximum of all possible distances between the cases in cluster A and the cases in cluster B.

- For the **between-groups linkage** or average linkage method, the dissimilarity between cluster A and cluster B is represented by the average of all the possible distances between the cases in cluster A and the cases in cluster B.

- For the **within-groups linkage** method, the dissimilarity between cluster A and cluster B is represented by the average of all the possible distances between the cases within a single new cluster determined by combining cluster A and cluster B.

- For the **centroid clustering** method, the dissimilarity between cluster A and cluster B is represented by the distance between the centroid for the cases in cluster A and the centroid for the cases in cluster B. Note that this distance is not mathematically equivalent to the average of the distances used in the average linkage method. Also note the SPSS warning below about using squared Euclidean distance rather than Euclidean distance for this procedure.

> The squared Euclidean measure should be used when the CENTROID, MEDIAN, or WARD cluster method is requested.

- For **Wards method**, the dissimilarity between cluster A and cluster B is represented by the loss of information from joining the two clusters with this loss of information being measured by the increase in error sum of squares. For a cluster the sum of squares is the sum of squared deviations of each case from the centroid for the cluster. The error sum of squares is the total of these for all clusters. When selecting clusters to join, the two clusters among all possible combinations that have the minimum increase in error sum of squares are selected. See the note above about using squared Euclidean distance rather than Euclidean distance for this method.

- For the **median clustering** method, the dissimilarity between cluster A and cluster B is represented by the distance between the SPSS determined median for the cases in cluster A and the median for the cases in cluster B. See the message in the note above about using squared Euclidean distance rather than Euclidean distance for this method.

Cluster analysis can be an effective tool to identify extreme data values in a multivariate data set. Extreme points will be a cluster by themselves while the vast majority of the other points are in one or more well populated clusters.

When performing hierarchical cluster analysis one can cluster cases or variables in SPSS by selecting either cases or variables in the initial menu. The default is cases since it is probably done more than clustering variables but clustering variables may be desirable in certain situations.

Clustering Procedures:

Hierarchical Cluster Analysis

This procedure attempts to identify relatively homogeneous groups of cases (or variables) based on selected characteristics, using an algorithm that starts with each case (or variable) in a separate cluster and combines clusters until only one is left. You can analyze raw variables or you can choose from a variety of standardizing transformations. Distance or similarity measures are generated by the Proximities procedure. Statistics are displayed at each stage to help you select the best solution. Statistics include agglomeration schedule, distance (or similarity) matrix, and cluster membership for a single solution or a range of solutions. Plots include dendrograms and icicle plots.

- Agglomeration schedule. Displays the cases or clusters combined at each stage, the distances between the cases or clusters being combined, and the last cluster level at which a case (or variable) joined the cluster.
- Proximity matrix. Gives the distances or similarities between items.
- Cluster Membership. Displays the cluster to which each case is assigned at one or more stages in the combination of clusters. Available options are single solution and range of solutions.
- Dendrograms can be used to assess the cohesiveness of the clusters formed and can provide information about the appropriate number of clusters to keep.
- Icicle plots display information about how cases are combined into clusters at each iteration of the analysis. (User can specify a range of clusters to be displayed) Orientation allows you to select a vertical or horizontal plot.

Data.

The variables can be quantitative, binary, or count data. Scaling of variables is an important issue–differences in scaling may affect your cluster solution(s). If your variables have large differences in scaling (for example, one variable is measured in dollars and the other is measured in years), you should consider standardizing them (this can be done automatically by the Hierarchical Cluster Analysis procedure).

Case Order.

If tied distances or similarities exist in the input data or occur among updated clusters during joining, the resulting cluster solution may depend on the order of cases in the file. You may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution.

Assumptions.

The distance or similarity measures used should be appropriate for the data analyzed. Also, you should include all relevant variables in your analysis. Omission of influential variables can result in a misleading solution. Because hierarchical cluster analysis is an exploratory method, results should be treated as tentative until they are confirmed with an independent sample.

Measures for Interval Data

The following dissimilarity measures are available for interval data:

- Euclidean distance. The square root of the sum of the squared differences between values for the items. This is the default for interval data. Squared Euclidean distance is the default on the classroom version of SPSS. This is reasonable given the warning below that SPSS puts in the output. The squared Euclidean measure should be used when the CENTROID, MEDIAN, or WARD cluster method is requested.

- Squared Euclidean distance. The sum of the squared differences between the values for the items.

- Pearson correlation. The product-moment correlation between two vectors of values.

- Cosine. The cosine of the angle between two vectors of values.

- Chebychev. The maximum absolute difference between the values for the items.

- Block. The sum of the absolute differences between the values of the item. Also known as Manhattan distance.

- Minkowski. The pth root of the sum of the absolute differences to the pth power between the values for the items.

- Customized. The rth root of the sum of the absolute differences to the pth power between the values for the items.

Transform Values

The following alternatives are available for transforming values:

- Z scores. Values are standardized to z scores, with a mean of 0 and a standard deviation of 1.
- Range -1 to 1. Each value for the item being standardized is divided by the range of the values.
- Range 0 to 1. The procedure subtracts the minimum value from each item being standardized and then divides by the range.
- Maximum magnitude of 1. The procedure divides each value for the item being standardized by the maximum of the values.
- Mean of 1. The procedure divides each value for the item being standardized by the mean of the values.
- Standard deviation of 1. The procedure divides each value for the variable or case being standardized by the standard deviation of the values.
- Additionally, you can choose how standardization is done. Alternatives are By variable or By case.

To Obtain a Hierarchical Cluster Analysis From the menus choose:

```
Analyze  >  Classify  >  Hierarchical Cluster...
```

If you are clustering cases, select at least one numeric variable. If you are clustering variables, select at least three numeric variables.

K-Means Cluster Analysis

This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that can handle large numbers of cases. However, the algorithm requires you to specify the number of clusters. You can specify initial cluster centers if you know this information. You can select one of two methods for classifying cases, either updating cluster centers iteratively or classifying only. You can save cluster membership, distance information, and final cluster centers. Optionally, you can specify a variable whose values are used to label casewise output. You can also request analysis of variance F statistics. While these statistics are opportunistic (the procedure tries to form groups that do differ), the relative size of the statistics provides information about each variable's contribution to the separation of the groups.

Statistics. Complete solution: initial cluster centers, ANOVA table. Each case: cluster information, distance from cluster center.

Assumptions. Distances are computed using simple Euclidean distance. If you want to use another distance or similarity measure, use the Hierarchical Cluster Analysis procedure. Scaling of variables is an important consideration–if your variables are measured on different scales (for example, one variable is expressed in dollars and another is expressed in years), your results may be misleading. In such cases, you should consider standardizing your variables before you perform the k-means cluster analysis (this can be done in the Descriptives procedure). The procedure assumes that you have selected the appropriate number of clusters and that you have included all relevant variables. If you have chosen an inappropriate number of clusters or omitted important variables, your results may be misleading.

Case and Initial Cluster Center Order.

The default algorithm for choosing initial cluster centers is not invariant to case ordering. The Use running means option on the Iterate dialog box makes the resulting solution potentially dependent upon case order regardless of how initial cluster centers are chosen. If you are using either of these methods, you may want to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution. Specifying initial cluster centers and not using the Use running means option will avoid issues related to case order. However, ordering of the initial cluster centers may affect the solution, if there are tied distances from cases to cluster centers. Comparing results from analyses with different permutations of the initial center values may be used to assess the stability of a given solution.

To Obtain a K-Means Cluster Analysis From the menus choose:

```
Analyze  >  Classify  >  K-Means Cluster...
```

- Select the variables to be used in the cluster analysis.
- Specify the number of clusters. The number of clusters must be at least two and must not be greater than the number of cases in the data file.
- Select either Iterate and classify or Classify only.