StatNews #30

Cluster Analysis
September 1998

When exploring and describing large data sets, it is sometimes useful to summarize the information by assigning each subject or item to a group with similar characteristics. This process is called classification. Cluster analysis is a family of methods for discovering natural classifications of items. Cluster analysis was originally developed to aid biologists in classifying plants and animals, but there are many different applications in fields from marketing to medicine.

For example, a researcher was interested in describing dietary attitudes. Each subject was given a list of attitudes towards food and asked whether he or she agreed with the attitudes. No criteria existed for classifying subjects on the basis of their food attitudes, and the researcher did not know in advance how many or what groups might be found. Cluster analysis was used to determine whether a natural grouping existed. Based on their answers to the questions, subjects were classified into three groups that were interpreted as identifying particular types of eaters.

A cluster analysis involves many decisions. First, a general approach must be chosen. Hierarchical cluster analysis is the most common for small- to medium-sized data sets, and disjoint (K-means) cluster analysis is useful for larger data sets.

Hierarchical cluster analysis starts by considering each item as a single cluster. It finds the two items that are most similar and joins them into a cluster. From that point on, the cluster is treated as a single item. The process is repeated, each time joining the two most similar items or clusters, until all the items are joined into one large cluster. The results may be displayed graphically so that the researcher can determine how many clusters to use.

Disjoint or K-means clustering requires the researcher to decide ahead of time how many clusters are in the data. The data set is arbitrarily divided into clusters, and then items are reassigned one by one to different clusters on the basis of their similarity to the other items in the cluster. The process continues until no items need to be reassigned. Disjoint clustering is more efficient for large data sets.

A second important decision is how to determine which items are similar to each other. Many different ways of measuring similarity have been proposed, and it is important to choose one that is appropriate for your data.

Third, a criterion for adding items and to clusters and combining clusters must be chosen. For example, an item may be added to a cluster if it is similar to one item in that cluster, or if it is similar to the average item in the cluster. Different criteria can result in very different classifications of items.

Finally, the number of clusters must be chosen. In the food attitudes example, the cluster identifying "healthy eaters" may be divided into vegetarians and non-vegetarians, or may be left as a single large cluster.

Cluster analysis can be performed by most common statistical software packages, including SAS, SPSS, and Minitab. If you would like to discuss using cluster analysis, contact the CSCU office.

Author: Cara Olsen

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).