Multivariate Statistics (Ralf Hansmann)

Exercise: Cluster Analysis I (Clustering cases)

A. Open the "margarine.sav" SPSS data.

1. Provide the proximity matrix for clustering the 5 cases (types of margarine) based on the distance measures:
   a) Squared Euclidian Distance
   b) Euclidian Distance
   c) City-Block
   d) Pearson correlation.

   *(For this you need to use the **statistics** and **method** subcommands of SPSS -> Analyse -> Classify -> hierarchical CA)*

2. Provide for the distance measure Squared Euclidian Distance the Dendrogram resulting from the clustering methods a) *average linkage between groups*, b) *average linkage within groups*, and c) *Ward*.

   *(For this you will need to use the **plot** subcommand of hierarchical CA)*

   Provide the Dendrogram for the clustering methods d) *average linkage between groups*, and e) *average linkage within groups* using the proximity measure *Pearson correlation.*

3. Compare the five Dendrograms. Which selection has been most fundamental (decisive) in terms of causing differences between the five Dendrograms?

B. Open the "graduate survey" SPSS data.

A set of 6 questions asked the graduates where they acquired their qualifications! Please use different methods of hierarchical CA for separating the cases into **two clusters** on basis of their values of the 6 corresponding variables:
F24: occupation before university, F24: occupation during university, F24: studies at university, F24: occupation after university, F24: further education after university
F24: outside engagement

- Use the **save** subcommand to *specify a two cluster solution* and to *require that the cluster* to which each case is assigned *be saved as a new variable*.
- **Deactivate the statistics and Plot options** of the hierarchical CA to save time.

Apply the distance measure Squared Euclidian Distance in the three clustering methods a) *Ward*, b) Complete Linkage (=furthest neighbour), and b) Single linkage.

1. Provide for each of the three methods, the *Frequencies* of the cases in the two resulting clusters. Which methods obtain rather balanced cluster sizes?
2. Provide for each of the three methods, the centroids of the two clusters!