Data analysis

By the terms "data analysis" is meant in statistics a set of multi-dimensional descriptive methods. These methods generally require information organised in the following way: "n" statistic individuals (spatial entities, households, firms...) described by "p" variables. These methods allow summarising information contained in data tables with large dimensions (table n lines x p columns). Two "families" of methods may be distinguished:

- Factor analyses : they consist in transforming the initial data table into a new table containing the same information, but in a *hierarchically organised* form. It is composed of *factor axes.* The first factor axis corresponds to the linear combination of initial variables which most differentiates the individuals from each other. It presents *maximal variance.* Factor axes are independent from each other and ordered in function of their variance. Generally a small number of factor axes (three or four) is sufficient to express the essential part of the information contained in the initial table. Interpretation of these factor axes allows to enlighten the form of interrelations between the studied variables and the similarities and dissimilarities between individuals with regard to these variables. The two most commonly used methods are the *principal components* analysis (appropriate for heterogeneous data combining variables expressed into different scales of measure, or also for variables expressed in percents) and *correspondence* analysis (appropriate for contingency tables or tables of qualitative variables).
- Classifications : they allow working typologies out and grouping individuals into classes according to their similarities with respect to all variables. A criterion frequently used from a technical point of view is to search for the classification that minimises intra-classes variance (variability among individuals of a same class) and maximises inter-classes variance (variability among classes). The most classical methods are the *ascending hierarchical* classification and the classification by *K-means clustering*.

See also : quantitative variables

Bibliographie