Confidence ellipses when analyzing simultaneously several contingency tables resulting from free-text descriptions

Cadoret Marine[1,*], Buche Marianne[1], Lê Sébastien[1]

1. Agrocampus Ouest, laboratoire de mathématiques appliquées, Rennes, France
* Contact author: marine.cadoret@agrocampus-ouest.fr

Keywords: confidence ellipse, correspondence analysis, multiple factor analysis, free-text description

Textual data are often analysed using correspondence analysis (CA) on a contingency table where, for instance, rows correspond to the texts of the corpus to be analysed and columns to the words used in the corpus: at the intersection of one row and one column there is the number of occurrences of the word (associated with the column) in the text (associated with the row) (Lebart, Salem & Berry, 1997).

In our particular application, a given set of items is described by two groups of subjects using free-text description. For each group of subjects we may build a contingency table where rows correspond to the items to be studied and columns to the words used to describe the items; we may then obtain a representation of the items per group using CA. The comparison of those two representations may be obtained by the simultaneous analysis of both contingency tables using the so called intra-sets multiple factor analysis (Bécue and Pagès, 1999; Escofier, Pagès, 1988-1998). This method provides a global representation of the rows as well as a partial representation of the rows from the point of view of each contingency table, within a single framework.

For each of those different representations, global and partial, we may wonder what might have been the positions of the items if the description had been generated by some other subjects. To answer that question, we propose a methodology that allows building confidence ellipses around the items that would represent the variability of the positions the items might have taken for other subjects (Lê, Husson & Pagès, 2004).

To build such ellipses, the idea is to resample the subjects with replacement and to build from those particular subjects' description a contingency table to be projected as supplementary elements on the axis issued from the analysis of the original groups of subjects. We then obtain a new representation of the set of items from a virtual group of subjects. Ellipses are finally obtained after having resampled a great number of times.

This methodology will be illustrated using as items the Rorschach's inkblots, two groups of subjects, a first one that has analyzed the cards following the official order of the Rorschach's test, a second one that has analyzed the cards following a random order.

References

Bécue, M., Pagès, J. (1999). Intra-Sets Multiple Factor Analysis. Application to textual data. *Proc. of the 9th International Symposium on Applied Stochastic Models and Data Analysis*, J. Jansen *et al.* (eds), Universidade de Lisboa Editor, 51-60.

Escofier, B., Pagès, J. (1988-1998). *Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation*, Dunod, Paris.

Lebart, L., Salem, A., Berry, L. (1997). *Exploring textual data*, Kluwer.

Lê, S., Husson, F. & Pagès, J. (2004). Confidence ellipses in HMFA applied to sensory profiles of chocolates. The 7th Sensometrics meeting, Davis (USA).