Multivariate Analysis of Histogram Data

Paula Brito
Faculdade de Economia & LIAAD - INESC TEC Universidade do Porto, Portugal, mpbrito@fep.up.pt

Palavras–chave: Clustering ; Discriminant Analysis ; Histogram data ; Regression ; Symbolic Data Analysis.

Abstract:
In classical Statistics and Multivariate Data Analysis data is usually represented in a matrix where each row represents a statistical unit, or individual, for which one single value is recorded for each numerical or categorical variable (in columns). This representation model is however too restricted when the data to be analysed comprises variability. That is the case when the entities under analysis are not single elements, but groups formed on the basis of some given common properties. Then, for each descriptive variable, the observed variability within each group should be taken into account, to avoid an important loss of pertinent information. To this aim, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or, more generally, distributions over a given domain. Symbolic Data Analysis provides a framework for the representation and analysis of such data, taking into account their inherent variability. In this talk, we consider the case of numerical data described by empirical distributions, known as histogram data. We introduce alternative representations of histogram observations, and consider descriptive statistics and distance measures. Methods for multivariate analysis of such data are then presented, which allow taking into account the variability expressed in the data representation.

Bibliografia

[1] Brito, P. Symbolic data analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4 (4), 281–295, 2014.

[2] Brito, P. and Chavent, M. Divisive Monothetic Clustering for Interval and Histogram-Valued Data. In: *Proc. ICPRAM 2012* - 1st International Conference on Pattern Recognition Applications and Methods, Vilamoura, Portugal, 2012.

[3] Dias, S. and Brito, P. Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8 (2) , 75–113, 2013.