DSS Statistics Seminar

June 17, 2022, 13:30

In person Room 34 (CU002)
Webinar https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09
Passcode: 432940

A general framework for implementing distances for categorical variables

Michel van de Velden
Econometric Institute, Erasmus University Rotterdam
Joint work with
A. Iodice D'Enza *Dipartimento di Scienze Politiche, Università di Napoli Federico II*
A. Markos *Department of Primary Education, Democritus University of Thrace*
C. Cavicchia *Econometric Institute, Erasmus University Rotterdam*

In many statistical methods, distance plays an important role. For instance, data visualization, classification and clustering methods require quantification of distances among objects. How to define such distance depends on the nature of the data and/or problem at hand. For distance between numerical variables, in particular in multivariate contexts, there exist many definitions that depend on the actual observed differences between values. It is worth underlining that often it is necessary to rescale the variables before computing the distances. Many distance functions exist for numerical variables. For categorical data, defining a distance is even more complex as the nature of such data prohibits straightforward arithmetic operations. Specific measures therefore need to be introduced that can be used to describe or study structure and/or relationships in the categorical data. In this paper, we introduce a general framework that allows an efficient and transparent implementation for distance between categorical variables. We show that several existing distances (for example distance measures that incorporate association among variables) can be incorporated into the framework. Moreover, our framework quite naturally leads to the introduction of new distance formulations as well.

SAPIENZA
UNIVERSITÀ DI ROMA

DSS - Dipartimento di Scienze Statistiche - www.dss.uniroma1.it