DSSV 2023 Data Science, Statistics & Visualisation
EC DA 2023 European Conference on Data Analysis

Methods to generate contingency tables satisfying user-specified properties

M. van de Velden[1], R. S. H. Willemsen[1], and W. van den Heuvel[1]

[1]Erasmus University Rotterdam, Econometric Institute, Burg. Oudlaan 50, Rotterdam, 3062PA, the Netherlands

Synthetic data can be used to study how models behave under varying conditions. Moreover, synthetic data are also crucial in the development of machine learning methods, where models are increasingly complex and require vast amounts of data to be estimated. In this paper, we consider data generation for synthetic categorical data. In particular, we consider methods for generating contingency tables that satisfy user-specified properties related to univariate (the marginals) as well as bivariate properties (the association between variables). Generating contingency tables with specified marginals has been studied by [1], [2], [3] and [4]. We consider the problem of generating contingency tables with user specified marginals, explained inertia (fit of a low-dimensional approximation) and Cramer's V (a measure of the strength of association between two nominal variables). We propose a three-step framework, offering both a mathematical programming and a heuristic approach for generating contingency tables, in which the user can select which conditions should be satisfied. We note that there is an inverse relationship between the values of explained inertia and Cramer's V, which indicates that it is difficult to generate tables for which the explained inertia and Cramer's V are close to the limit.

[1] J. M. Boyett, "Algorithm as 144: Random r × c tables with given row and column totals," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 3, pp. 329–332, 1979.

[2] W. M. Patefield, "Algorithm as 159: An efficient method of generating random r × c tables with given row and column totals," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 30, no. 1, pp. 91–97, 1981.

[3] D. W. Balmer, "Algorithm as 236: Recursive enumeration of r x c tables for exactly likelihood evaluation," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 37, no. 2, pp. 290–301, 1988.

[4] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu, "Sequential monte carlo methods for statistical analysis of tables," *Journal of the American Statistical Association*, vol. 100, pp. 109–120, 2005.