Description of algorithm R_synthpop_ipf

Briefly

1. Make a table of the counts in a complete tabulation of all variables in the data. Number of cells in table denoted by N.
2. Identify any cells in the table that should be structural zeros, e.g. PINCP_DECILE not missing for AGEP <15.
3. Add a small value to all non structural zero cells in the table, Here default value of 1/N has been used.
4. Select the margins to be preserved in the synthetic data. Default is all two-way marginals
5. To make the synthetic data DP add Laplace noise with parameter epsilon to all the marginal counts. except structural zero cells. Ignore this step if not DP. Adjust the values to make them positive.
6. Use iterative proportional fitting to generate a complete table of the proportions in the cell from these margins.
7. If parameter ipf.rand is not set to FALSE, generate multinomial random counts from a multinomial distribution with parameters given by the proportions in the table and sample size equal to the original sample size.
8. Recreate a new data set from this table.

Source

1. Synthpop package for R https://cran.r-project.org/web/packages/synthpop/index.html
2. Code for the function `syn.ipf` can be found in the file `functions.R` at https://github.com/bnowok/synthpop/blob/master/R/functions.syn.r
3. A paper describing some results of this method, presented at PSD in Paris in 2022 and published in proceedings can be found here Utility and Disclosure Risk for Differentially Private Synthetic Categorical Data