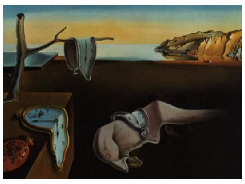

Journal of Statistical Software

November 2013, Volume 55, Code Snippet 1. *http://www.jstatsoft.org/*

A Greedy Algorithm for Representative Sampling: `repsample` in **Stata**

Evangelos Kontopantelis
University of Manchester

Abstract

Quantitative empirical analyses of a population of interest usually aim to estimate the causal effect of one or more independent variables on a dependent variable. However, only in rare instances is the whole population available for analysis. Researchers tend to estimate causal effects on a selected sample and generalize their conclusions to the whole population. The validity of this approach rests on the assumption that the sample is representative of the population on certain key characteristics. A study using a non-representative sample is lacking in external validity by failing to minimize population choice bias. When the sample is large and non-response bias is not an issue, a random selection process is adequate to ensure external validity. If that is not the case, however, researchers could follow a more deterministic approach to ensure representativeness on the selected characteristics, provided these are known, or can be estimated, in the parent population. Although such approaches exist for matched sampling designs, research on representative sampling and the similarity between the sample and the parent population seems to be lacking. In this article we propose a greedy algorithm for obtaining a representative sample and quantifying representativeness in Stata.

Keywords: representative sample, Stata, greedy algorithm.

1. Introduction

Randomization is a simple procedure that can often ensure that selection bias is not incurred, when attempting to estimate a causal effect. In the randomized controlled trial, arguably the best experimental design, randomly selected groups can only be randomly different on all observed and unobserved background covariates. When the group sizes are large, the simple randomization is expected to achieve balance in group sizes and covariates, since the random error is small. However, when group sizes are not large and balance is required on many covariates, alternative approaches are often considered. In stratified randomization, the covariates define strata of ‘similar’ subjects and a simple randomization is performed within each stratum. Practically, a small number of covariates can be factored in this design since the number of strata increases exponentially with covariates. A widely used deterministic alternative is the minimization method, which attempts to minimize the total imbalance between groups in selected covariates (Pocock and Simon 1975; Begg and Iglewicz 1980). In observational studies, propensity score matching is widely used to minimize covariate bias. Subjects are matched across groups on their propensity scores, the conditional probabilities of assignment given a group of observed covariates (Rosenbaum and Rubin 1983). More recently, a ‘fine balance’ method was proposed, which attempts to provide identical marginal distributions of the covariates of interest within each group, without a one-to-one matching of subjects (Rosenbaum, Ross, and Silber 2007). An alternative to propensity score matching is coarsened exact matching, under which the covariates of interest are reduced (or ‘coarsened’) into acceptable categories (bins) that define a range of strata. Each observation is then assigned to a stratum and one-to-one matching is performed within strata with 2 or more observations (Blackwell, Iacus, King, and Porro 2009; Iacus, King, and Porro 2009).

In representative sampling the aim is to obtain a sample that is as representative of a population as possible on selected available characteristics, possibly even before the ‘main’ data collection. A study using a sample that is a close match to its parent population on key covariates would be considered to possess external validity, allowing for valid generalizations from the sample to the population. Although the aim is different, representative sampling shares some of the challenges of randomization/minimization and matching. A simple random sampling algorithm is easy to implement but might fail to deliver a representative sample. Besides the potential inability to provide balance on key covariates, a simple random process would be affected by the possible presence of non-participation bias if the whole population was not available. Another issue can be logistical constraints, which might limit the selection of the sample from a population subset with different characteristics, in which case a random process would not provide a representative sample of the population. Therefore, a more deterministic approach to the sampling process may be more suitable. Such a framework does not seem to exist for representative sampling, although the problem shares some of the features of minimization, propensity score matching, coarsened exact matching and fine balance. However, it also faces unique challenges. In all methods outlined above either one-to-one matching is performed or a distance statistic is selected to quantify the distance between subjects (e.g., Mahalanobis distance) and is then used in the group allocation process – or both. For representative sampling, one-to-one matching is not relevant and even in its absence, as in ‘fine balance’, methods are not applicable to this context. In addition, the use of the term ‘representative’ is rather arbitrary and sometimes studies do not provide a comparison between the sample and the population characteristics on which the sample is deemed to be representative. Even when a comparison is presented the representativeness of the sample is not quantified.

We propose a new greedy algorithm in Stata (StataCorp. 2011b) that uses common discrete and continuous distribution comparison methods to provide a sample that is as representative as possible of (i) a measured and available population, or (ii) one or more theoretical distributions. The algorithm can be fully or partly deterministic and at each step it selects the case that minimizes the overall difference between the distribution(s) of the sample and the population. It also reports an overall measure of representativeness that can offer standardization and transparency and acts as a criterion on the external validity of a study.

2. Methods

Let us assume a population (or pool) of size N from which we wish to draw a sample of k cases, which is the most representative in terms of the distribution of n_1 continuous variables $X_1, X_2, ..., X_{n_1}$ and n_2 discrete variables $Y_1, Y_2, ..., Y_{n_2}$. We define the most representative sample as the one whose overall distance from the population or the theoretical distributions is the smallest. A simple description of the representative sampling algorithm is:

Step 1: A percentage of the sample is randomly selected.

Step 2: Each case in the eligible pool is temporarily added to the current sample in turn and the distance to the population or a theoretical distribution, for each of the n_1 plus n_2 variables, is estimated. Since the process can be computationally taxing, an early stop rule can be employed that stops the search when a sample that provides an overall distance below an arbitrarily set threshold is identified.

Step 3: An overall measure of distance is estimated across all variables for each of the possible inclusions and the case whose inclusion leads to the smallest overall distance is added to the sample.

Steps 2 and 3 are repeated until k cases are selected. In other words, the algorithm is of $O(N \cdot k)$ complexity and attempts to arrive at a solution for the overall distance minimization problem by selecting the local minimum for each case added to the sample. The methods used for Step 2 vary by sampling type (population or theoretical) and test type (asymptotic or exact) and are presented in Sections 2.1 and 2.2. Methods for Step 3 are consistent across sampling type and are presented in Section 2.3.

2.1. Population sampling

When the user requests the current dataset to be used as the whole population the second step involves two-way Kolmogorov-Smirnov tests for continuous variables and χ^2 tests (or Fisher's exact) for discrete variables.

For each continuous variable X_i, at each case selection stage, two-way Kolmogorov-Smirnov tests are performed across all eligible cases. Assuming we have already sampled m cases (randomly and/or deterministically) from population N, the Kolmogorov-Smirnov statistic calculated for the inclusion of the $m + 1$ case is:

$$D_{m+1,N} = \sup_x |F_{1,m+1}(x) - F_{2,N}(x)|. \tag{1}$$

In (1), $F_{1,m+1}$ and $F_{2,N}$ are the empirical cumulative distribution functions of the sample and population respectively, for variable X_i. The expression

$$\sqrt{\frac{N(m+1)}{N + (m+1)}} D_{m+1,N}$$

converges to the Kolmogorov distribution and using the respective table we can test the hypothesis of common sample and population distributions for the sample and obtain an asymptotic p value. If the user requests asymptotic tests a correction is applied to the p value by using a numerical approximation technique (StataCorp. 2011a). If exact tests are requested the exact p value is calculated by a counting algorithm (Gibbons and Chakraborti 2011). In all cases, the p value obtained gives the probability of the two distributions being as different as observed (or more), under the assumption of a common distribution.

For each binary or categorical variable Y_i, at each case selection stage, χ^2 or Fisher's exact tests are performed across all eligible cases. Assuming we have already sampled m cases (randomly and/or deterministically) from population N, we evaluate the homogeneity between population and sample (including case $m+1$ in the latter) through the χ^2 statistic:

$$X^2 = \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{2}$$

where K is the number of categories for variable Y_i, O_{ij} is the number of observations for the ith row and jth column of the K by K table and

$$E_{ij} = \frac{\sum_{j=1}^{K} O_{ij} \sum_{i=1}^{K} O_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} O_{ij}}.$$

The statistic described in (2) asymptotically follows a χ^2 distribution with $(K-1)^2$ degrees of freedom and a p value is obtained from the relevant table. At the user's request, Fisher's exact test and its extension for $r \times c$ tables is alternatively used. The test is a permutation test that uses every possible table to calculate a probability of observing a table that gives at least as much evidence of heterogeneity as the one observed under the homogeneity assumption. The version implemented in Stata is the algorithm proposed by Mehta and Patel (1983).

2.2. Theoretical sampling

Alternatively, the user might not wish to treat the dataset as the whole population but as a pool from which to draw a sample that is representative in terms of one or more theoretical distributions. In that case, the second step of the algorithm involves one-way Kolmogorov-Smirnov tests for continuous variables and one-sample tests of proportion (or binomial probability tests) for discrete variables.

Similarly to the population sampling scenario, for each continuous variable X_i, at each case selection stage, one-way Kolmogorov-Smirnov tests are performed across all eligible cases. The Kolmogorov-Smirnov statistic, after the inclusion of the $m+1$ case would be:

$$D_{m+1} = \sup_x |F_{m+1}(x) - F(x)| \tag{3}$$

In (3), F_{m+1} is the empirical cumulative distribution function of the sample and F the theoretic cumulative distribution function, e.g., a normal cumulative distirbution function with mean and standard deviation provided by the user, for variable X_i. In this case, expression $\sqrt{m+1}D_{m+1}$ converges to the Kolmogorov distribution and we test the hypothesis the sample is normally distributed with the specified parameters and obtain a corrected asymptotic p value (StataCorp. 2011a).

For each binary variable Y_i, at each case selection stage, one-sample tests of proportions or binomial probability tests are performed across all eligible cases. For the inclusion of the

$m{+}1$ case, for example, we compare the proportion observed in the sample to the hypothesized proportion through the statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/(m+1)}}, \tag{4}$$

where $\hat{p}$ and p_0 are the observed and hypothesized proportions respectively. The statistic described in (4) asymptotically follows a normal distribution, which can be used to obtain an asymptotic p value. Alternatively, an exact p value is calculated using the binomial distribution. In that case the upper and lower one-sided p values are provided by (5) and (6):

$$\mathsf{P}(\kappa \geq \kappa') = \sum_{i=\kappa'}^{m+1} \binom{m+1}{i} {p_0}^i (1-p_0)^{m+1-i} \tag{5}$$

$$\mathsf{P}(\kappa \leq \kappa') = \sum_{i=0}^{\kappa'} \binom{m+1}{i} {p_0}^i (1-p_0)^{m+1-i} \tag{6}$$

where κ' is the observed number of successes. The two-sided p value is $\mathsf{P}(\kappa \leq \kappa_a$ or $\kappa \geq \kappa')$ if $\kappa' \geq (m+1)p_0$ and $\mathsf{P}(\kappa \leq \kappa'$ or $\kappa \geq \kappa_b)$ otherwise, with κ_a the largest number $\leq (m+1)p_0$ such that $\mathsf{P}(\kappa = \kappa_a) \leq \mathsf{P}(\kappa = \kappa')$ and κ_b the smallest number $\geq (m+1)p_0$ such that $\mathsf{P}(\kappa = \kappa_b) \leq \mathsf{P}(\kappa = \kappa')$ (Zar 1999).

The p value obtained from all tests in this section gives the probability of the sample distribution being as different to the theoretical as observed (or more), under the assumption the sample is drawn from that reference distribution.

2.3. Overall distance measure

Assuming the performed tests, n_1 for continuous and n_2 for discrete variables, are independent, we can combine their results using Fisher's combined probability test (Fisher 1932). The test uses the p values from independent tests of the same null hypothesis and calculates the test statistic described in (7):

$$\chi_F^2 = -2 \sum_{i=1}^{n_1+n_2} \ln(p_i). \tag{7}$$

When all null hypotheses are true, χ_F^2 will have a χ^2 distribution with $2(n_1+n_2)$ degrees of freedom. Following from that, a p value can be determined to inform the decision of rejection or not of the overall null hypothesis: in our case, common sample-population distributions (population sampling) or the sample being drawn from hypothesized distributions (theoretical sampling). Although the role of the procedure as a composite test has been criticized due to the fact that results consistent with the alternative hypothesis influence the test statistic disproportionately more than the results consistent with the null (Rice 1990; Whitlock 2005), this property is arguably desirable in representative sampling. Rice (1990) provides an example to illustrate the issue; if two studies with p values of 0.999 and 0.001 were combined the resulting p value under Fisher's method would be 0.008. However, this asymmetric sensitivity to small p values can be beneficial when we wish to maximize representativeness across all provided distributions and not 'on average'. In other words, we wish for the algorithm to select the path that leads as far from rejection of any of the null hypotheses as possible: at each selection step it includes in the sample the case that leads to the lowest χ^2_F and the highest p value. Therefore, the algorithm, at each selection step, selects the sample with the highest probability of being the least different to the population (observed or hypothesized), under the assumption of common distributions. The formula in (7) can easily be edited to account for unequal weights across the performed tests.

3. The repsample command

3.1. Syntax

`repsample` # [*if*] [*in*] [, `cont(`*varlist*`)` `bincat(`*varlist*`)` `mean(`*numlist*`)` `sd(`*numlist*`)` `perc(`*numlist*`)` `seednum(`#`)` `randomperc(`#`)` `srule(`#`)` `wght(`*numlist*`)` `retain(`*varname*`)` `exact` `force`]

3.2. Options

`cont`(*varlist*) Continuous variable(s) on which 'representativeness' will be based.

`bincat`(*varlist*) Binary and categorical variable(s) on which 'representativeness' will be based. For theoretical sampling only binary variables are allowed.

`mean`(*numlist*) List of means for continuous variables. Order must correspond to order in option `cont`. Only required for sampling using one or more theoretical distributions.

`sd`(*numlist*) List of standard deviations for continuous variables. Order must correspond to order in option `cont`. Only required for sampling using one or more theoretical distributions.

`perc`(*numlist*) List of percentages for continuous variables. Order must correspond to order in option `bincat` and percentages need to be in the $(0, 100)$ range. Only required for sampling using one or more theoretical distributions.

`seednum`(#) Random seed number; the default is 7.

`randomperc`(#) The percentage of cases that will be randomly selected at the start of the algorithm. The percentage must be in the $[0, 100]$ range and the default value is 10. Setting to zero will provide a completely deterministic sample and to 100 a completely random sample.

`srule`(#) Early stopping rule that speeds up the process, using the p value for Fisher's combined probability (χ^2) test. It must be in the [0.5,1) range and once a sub-sample is identified for which the p value is above the one specified, that sub-sample is selected and the search is stopped early (without going through all cases). Then the algorithm proceeds to select the next case in the sample using the same decision rule. This option is a compromise and the smaller the threshold value, the less likely that the resulting sample will be a close match.

`wght`(*numlist*) Variable weighting, to give greater importance to some variables in the process. User needs to provide as many numbers as there are variables and the total weight needs to add up to 100. Assigning order is fixed and must correspond to the variables provided under the `cont` and `bincat` options, with all continuous variables (if any) prioritized, followed by binary and categorical variables (if any). Note that the overall matching measures reported in `r(chi2)` and `r(p)` are using the weighted scores and therefore quantify the matching under the provided weights assumption. But the individual variable matching measures are unaffected.

General		

Fisher's combined probability test	`r(p)`
	`r(chi2)`
	`r(df)`
		p value
		χ^2_F statistic
		degrees of freedom

Population sampling		

Two-sample Kolmogorov-Smirnov test for continuous variable *var*	`r(`*var*`_p)`
	`r(`*var*`_D)`
		p value, corrected or exact
		combined D

χ^2 or Fisher's exact test for binary or categorical variable *var*	`r(`*var*`_p)`
	`r(`*var*`_chi2)`
		p value
		χ^2 test statistic (asymptotic test only)

Theoretical sampling		

One-sample Kolmogorov-Smirnov test for continuous variable *var*	`r(`*var*`_p)`
	`r(`*var*`_D)`
		p value, corrected
		combined D

One-sample test of proportions or binomial probability test for binary variable *var*	`r(`*var*`_p)`
	`r(`*var*`_z)`
		p value
		z statistic (asymptotic test only)

Table 1: Command results, scalars in `r()`.

`retain(`*varname*`)` If a sampling variable to be retained is provided the program will sample on top of the current sample. This option is provided for batched sampling (running time can be very long for large samples and populations), and replacing cases that have withdrawn or become unavailable. For example, if the idea is to sample 100 representative patients to be enrolled in a study the researcher might wish to replace patients who did not agree to participate in the first instance. Cases that are not eligible for selection and cannot be dropped since they define the population should be set to missing in variable *varname* prior to executing the `repsample` command.

`exact` Use exact tests instead of asymptotic approximations. For population sampling, this option increases computation time considerably since it calculates exact p values in the two-sample Kolmogorov-Smirnov tests (for continuous variables) and Fisher's exact test (for categorical and binary variables). For theoretical sampling, the increase in computation time is not as dramatic since only binary variables are affected with the use of the binomial probability test.

`force` Force replace sample information variable *repsample*, if present in the dataset. Cannot be used along with option `retain(`*repsample*`)`, but can be used with that option if another variable name is specified.

3.3. Saved results

The command creates binary variable *repsample* which contains the sample information. In addition it saves the results for the tests used (for the final sample only) in scalars in `r()` (Table 1).

4. Motivational example

In England, primary care is delivered by over 8,200 family practices (called general practices).

Although they share many contract characteristics and they uniformly participate in the Quality and Outcomes Framework, a remuneration scheme that rewards practitioners for good clinical practice, they vary greatly in practice list size and characteristics of the area they serve. Therefore, primary care studies that wish to investigate practices attempt to recruit a sample that is as representative as possible of all English practices on these key characteristics. This is usually a two step process: (i) a large pool of practices (five to ten times the size of the required sample) is asked to be considered for participation and (ii) a sample representative of all English practices is selected from those that agreed to participate. However, considering that sample sizes are usually small, stratified random sampling approaches are limited in what they can deliver. Usually, two variables are the practical limit on which to stratify and variables need to be reduced to two or three categories with loss of information. On occasion, practices that withdraw from a study might need to be replaced, an issue which leads to complications if the respective stratified random sampling cell from which to replace is empty.

As an example we apply the `repsample` command to a dataset of 2,474 practices in the North of England (North-West, North-East and Yorkshire & the Humber), that contains practice unique National Health Service identifiers and location, organizational and structural characteristics. Location characteristics are measured at the super-output area level, a low level geographical area categorization with unchanged boundaries over time (Communities and Local Government 2011; Bibby and Shepherd 2004). We focus on list size (variable `listsize`, continuous in the 518–36,884 range), area deprivation as measured by the 2007 Index of Multiple Deprivation (variable `soaimd07`, continuous in the 1.53–85.46 range) and rurality (variable `ruralvar`, binary with areas labeled as urban if they contain populations of 10,000 or more). The details of the dataset and the distributions of the variables of interest are provided below:

```
. use repsample_example.dta,

. describe

Contains data from repsample_example.dta
  obs:         2,474
 vars:            10                          3 May 2013 15:37
 size:       227,608
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
practicecode    str6   %9s                    Practice code
postcode        str8   %9s                    Postcode
shacode         str3   %9s                    Strategic Health Authority code
shaname         str24  %24s                   Strategic Health Authority name
pctcode         str3   %9s                    Primary Care Trust code
pctname         str35  %35s                   Primary Care Trust name
listsize        long   %12.0g                 Practice list size
ftes            float  %9.0g                  Doctors´ Full Time Equivalence
                                                in practice
```

```
soaimd07        float  %9.0g                  Practice location Index of
                                                Multiple Deprivation
ruralvar        byte   %11.0g      rurallb    Practice location rurality
-------------------------------------------------------------------------------

Sorted by:

. sum listsize soaimd07, detail

                     Practice list size
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1273            518
 5%         1833            538
10%         2125            628       Obs                2474
25%         2985            630       Sum of Wgt.        2474

50%         5339                      Mean           6100.614
                        Largest       Std. Dev.      3848.635
75%         8263          29195
90%        11355          29748       Variance       1.48e+07
95%        13098          31135       Skewness       1.418106
99%        17524          36884       Kurtosis       6.943647

       Practice location Index of Multiple Deprivation
-------------------------------------------------------------
      Percentiles      Smallest
 1%         4.22           1.53
 5%         6.97            1.6
10%         9.66           2.05       Obs                2474
25%        16.03           2.22       Sum of Wgt.        2474

50%        30.46                      Mean           33.58333
                        Largest       Std. Dev.      19.96176
75%        49.43          80.69
90%         63.1          85.46       Variance       398.4717
95%        68.89          85.46       Skewness       .4262682
99%         77.4          85.46       Kurtosis       2.101225


. tab ruralvar

   Practice |
   location |
   rurality |      Freq.     Percent        Cum.
------------+-----------------------------------
Urban > 10k |      2,181       88.16       88.16
      Other |        293       11.84      100.00
------------+-----------------------------------
      Total |      2,474      100.00
```

Let us assume we wish to recruit practices only from the North-East of England for logistical reasons. However, the distributions for the area differ somewhat from the distributions for the whole of the North of England. Although, on average, practices are located in areas of similar deprivation, they tend to be larger and are more often located in rural area compared to all Northern English practices:

```
. sum listsize soaimd07 if shaname=="North East", detail

                     Practice list size
-------------------------------------------------------------

      Percentiles      Smallest
 1%          843            699
 5%         1874            724
10%         2165            809       Obs                 394
25%         3344            843       Sum of Wgt.         394

50%         5805                      Mean           6657.558
                        Largest       Std. Dev.      4111.979
75%         8993          20992
90%        11894          21554       Variance       1.69e+07
95%        14180          22904       Skewness       1.305007
99%        20992          29748       Kurtosis       6.008121

       Practice location Index of Multiple Deprivation
-------------------------------------------------------------

      Percentiles      Smallest
 1%         3.82           2.77
 5%         8.01           2.97
10%        11.06           3.65       Obs                 394
25%        17.06           3.82       Sum of Wgt.         394

50%        34.04                      Mean           34.86195
                        Largest       Std. Dev.      19.16166
75%        50.89          76.07
90%        61.68          79.05       Variance       367.1692
95%        65.15          80.62       Skewness         .26784
99%        76.07          80.62       Kurtosis       2.014374


. tab ruralvar if shaname=="North East"

   Practice |
   location |
   rurality |      Freq.     Percent        Cum.
------------+-----------------------------------
Urban > 10k |        325       82.49       82.49
      Other |         69       17.51      100.00
------------+-----------------------------------
      Total |        394      100.00
```

If we wish to recruit a sample of 20 practices from the North-East with distributions that match those of all Northern English practices in terms of deprivation, list size and rurality, as closely as possible, we can do so with the following code:

```
. qui gen repsample=.

. qui replace repsample=0 if shaname=="North East"

. repsample 20, cont(listsize soaimd07) bincat(ruralvar) randomperc(30)
> seednum(16) retain(repsample)

Representative sample of 20 cases requested, sampling using a population.
Asymptotic approximations selected.
....................
. return list

scalars:
         r(ruralvar_p) =  .7993279153781966
      r(ruralvar_chi2) =  .0646262847022684
         r(listsize_p) =  1
         r(listsize_D) =  .0784155214227971
         r(soaimd07_p) =  1
         r(soaimd07_D) =  .0669361358124495
                 r(df) =  6
               r(chi2) =  .4479680203561796
                  r(p) =  .9984152631862227
```

Through variable `repsample` we have informed the command that only practices in the North-East are eligible for selection (set to zero when all other practices are set to missing) but all available practices are included in the comparisons with the various methods. The *p* value for Fisher's combined probability test provides a measure of the sample's overall representativeness which seems to be quite strong. We can see that the command selected only practices from the requested locality and the distributions for the sample more closely match distributions for Northern English practices, than those for the North-East:

```
. tab shaname if repsample==1

    Strategic Health |
      Authority name |      Freq.     Percent        Cum.
---------------------+-----------------------------------
          North East |         20      100.00      100.00
---------------------+-----------------------------------
               Total |         20      100.00

. sum listsize soaimd07 if repsample==1, detail

                      Practice list size
-------------------------------------------------------------
      Percentiles      Smallest
```

```
 1%         2029           2029
 5%         2135           2241
10%         2297           2353       Obs                  20
25%       3024.5           2829       Sum of Wgt.          20

50%       5358.5                      Mean            6327.15
                        Largest       Std. Dev.      4244.435
75%       8182.5          10011
90%        11468          10311       Variance       1.80e+07
95%        15912          12625       Skewness       1.519511
99%        19199          19199       Kurtosis       5.262171

       Practice location Index of Multiple Deprivation
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3.65           3.65
 5%        5.865           8.08
10%       10.415          12.75       Obs                  20
25%        16.34          13.62       Sum of Wgt.          20

50%        31.21                      Mean             34.267
                        Largest       Std. Dev.      21.36601
75%       50.585          57.01
90%       62.735          61.68       Variance       456.5062
95%       72.205          63.79       Skewness       .4694546
99%        80.62          80.62       Kurtosis        2.24604
```

```
. tab ruralvar if repsample==1

   Practice |
   location |
   rurality |      Freq.     Percent        Cum.
------------+-----------------------------------
Urban > 10k |         18       90.00       90.00
      Other |          2       10.00      100.00
------------+-----------------------------------
      Total |         20      100.00
```

If we assume one practice withdrew and we need to provide a replacement we can set the practice to missing in variable `repsample` and re-run the command:

```
. gsort -repsample

. qui replace repsample=. in 1

. qui replace repsample=. if shaname!="North East"

. repsample 20, cont(listsize soaimd07) bincat(ruralvar) randomperc(0)
> seednum(16) retain(repsample)
```

```
Representative sample of 20 cases requested, sampling using a population. 19
> cases carried forward.
Asymptotic approximations selected.
.
. return list

scalars:
          r(ruralvar_p) =  .7993279153781966
       r(ruralvar_chi2) =  .0646262847022684
          r(listsize_p) =  1
          r(listsize_D) =  .0784155214227971
          r(soaimd07_p) =  1
          r(soaimd07_D) =  .0669361358124495
                  r(df) =  6
                r(chi2) =  .4479680203561796
                   r(p) =  .9984152631862227
```

The p value for Fisher's method has not changed and the sample, again, seems to be representative:

```
. tab shaname if repsample==1

        Strategic Health |
          Authority name |      Freq.     Percent        Cum.
-------------------------+-----------------------------------
              North East |         20      100.00      100.00
-------------------------+-----------------------------------
                   Total |         20      100.00

. sum listsize soaimd07 if repsample==1, detail

                      Practice list size
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2029           2029
 5%         2135           2241
10%         2297           2353       Obs                  20
25%       3024.5           2829       Sum of Wgt.          20

50%         5489                      Mean            6498.45
                        Largest       Std. Dev.      4277.536
75%       8703.5          10011
90%        11468          10311       Variance       1.83e+07
95%        15912          12625       Skewness       1.373455
99%        19199          19199       Kurtosis       4.872573

       Practice location Index of Multiple Deprivation
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3.65           3.65
 5%        5.865           8.08
10%       10.415          12.75       Obs                  20
25%        16.34          13.62       Sum of Wgt.          20

50%        31.21                      Mean            34.3755
                        Largest       Std. Dev.      21.28607
75%       50.585          57.01
90%       62.735          61.68       Variance       453.0967
95%       72.205          63.79       Skewness       .4671501
99%        80.62          80.62       Kurtosis       2.262287

. tab ruralvar if repsample==1

   Practice |
   location |
   rurality |      Freq.     Percent        Cum.
------------+-----------------------------------
Urban > 10k |         18       90.00       90.00
      Other |          2       10.00      100.00
------------+-----------------------------------
      Total |         20      100.00
```

Alternatively, we may wish to sample using theoretical distributions. In that case we can select the sub-pool we wish to sample from (North-East) and provide the details for the theoretical distributions in the `repsample` syntax:

```
. qui keep if shaname=="North East"

. repsample 20, cont(listsize soaimd07) bincat(ruralvar) mean(5000 20)
> sd(3000 15) perc(30) randomperc(30) seednum(16) exact

Representative sample of 20 cases requested, sampling using theoretical distr
> ibutions.
Exact tests selected; only binary variables affected in theoretical sampling.
....................
. return list

scalars:
         r(ruralvar_p) =  .9999999999999993
         r(listsize_p) =  .9426655665292563
         r(listsize_D) =  .1110056095020276
         r(soaimd07_p) =  .8647982952848431
         r(soaimd07_D) =  .1253467276053295
                 r(df) =  6
               r(chi2) =  .4086053831329215
                  r(p) =  .9987796929701271
```

Distributions for the sample are similar to the theoretical distributions specified:

```
. tab shaname if repsample==1
```

```
        Strategic Health |
          Authority name |      Freq.     Percent        Cum.
-------------------------+-----------------------------------
              North East |         20      100.00      100.00
-------------------------+-----------------------------------
                   Total |         20      100.00

. sum listsize soaimd07 if repsample==1, detail

                     Practice list size
-------------------------------------------------------------
      Percentiles      Smallest
 1%          886            886
 5%       1457.5           2029
10%         2135           2241       Obs                  20
25%         3012           2536       Sum of Wgt.          20

50%         5118                      Mean            5166.55
                        Largest       Std. Dev.      2626.369
75%         6838           7338
90%         9002           8908       Variance        6897814
95%         9881           9096       Skewness       .3627777
99%        10666          10666       Kurtosis       2.339546

       Practice location Index of Multiple Deprivation
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2.77           2.77
 5%         2.87           2.97
10%         3.31           3.65       Obs                  20
25%        10.03           6.58       Sum of Wgt.          20

50%       19.395                      Mean            23.0005
                        Largest       Std. Dev.      17.20247
75%       31.845          34.25
90%       50.605           44.2       Variance        295.925
95%         60.4          57.01       Skewness       .9054984
99%        63.79          63.79       Kurtosis       3.145318

. tab ruralvar if repsample==1

   Practice |
   location |
   rurality |      Freq.     Percent        Cum.
------------+-----------------------------------
Urban > 10k |         14       70.00       70.00
      Other |          6       30.00      100.00
```

```
-------------+-----------------------------------
       Total |         20      100.00
```

5. Performance

A small simulation was performed to investigate the performance of the algorithm. We generated data for 200 observations on two normally distributed continuous variables (Mean = 0; SD = 1) and one binary variable (p = 0.15). Performance of `repsample` under the default options (asymptotic, `randomperc(0)` and population sampling) was compared to a computationally inexpensive random sampling approach, for samples of 10, 30 and 50 cases and matching on one continuous, two continuous or two continuous & one binary variable. The measure of comparison was the mean absolute difference between sample and population means and (for the two continuous variables) standard deviations, over 1000 repetitions (Table 2). As expected, samples with the purely random approach match the population better as the size of sample increases. The trend is the same with the `repsample` algorithm but the absolute differences in all scenarios are much lower, compared to the random approach.

Unfortunately the command can be computationally very expensive, especially when the pool from which to sample is large. The command was executed fully deterministically (i.e., with `randomperc(0)`) in various scenarios on an Intel Core i5-2500 CPU (3.30GHz) computer with 8GB of RAM, running Windows 7 64bit and Stata 12.1 SE and the completion times are presented in Table 3. More specifically, we varied the sampling approach (population or theoretical; exact or asymptotic), the pool size from which to sample (100, 1000 or 10000) and the number of variables used in the process (1 to 3). Reported times are approximate since they can be greatly affected by other processes executed in parallel. Note that we did not execute a population-exact sampling for a pool of size 10000 since the running time would likely be a few months.

Sample size	Matching	Absolute mean differences		
		Var. #1	Var. #2	Var. #3
10	repsample, 1 var	0.063 (0.118)	.	.
	repsample, 2 vars	0.060 (0.123)	0.068 (0.113)	.
	repsample, 3 vars	0.065 (0.115)	0.064 (0.111)	2.5%
	random	0.216 (0.117)	0.232 (0.173)	8.7%
30	repsample, 1 var	0.031 (0.060)	.	.
	repsample, 2 vars	0.031 (0.060)	0.032 (0.055)	.
	repsample, 3 vars	0.033 (0.063)	0.032 (0.054)	0.8%
	random	0.118 (0.098)	0.123 (0.093)	4.4%
50	repsample, 1 var	0.026 (0.047)	.	.
	repsample, 2 vars	0.026 (0.047)	0.027 (0.043)	.
	repsample, 3 vars	0.027 (0.047)	0.026 (0.040)	0.5%
	random	0.088 (0.071)	0.089 (0.066)	3.2%

Table 2: Mean absolute difference between population and sample for means (standard deviations) when sampling using one continuous, two continuous, or two continuous & one binary variable.

Sampling	Pool size	Sample size								
		10			30			50		
		1 var	2 vars	3 vars	1 var	2 vars	3 vars	1 var	2 vars	3 vars
Population, asymptotic	100	1	2	3	5	7	8	6	11	12
	1000	78	155	157	230	456	465	380	748	765
	10000	9055	17851	18054	28099	54473	55352	45590	90011	104235
Population, exact	100	12	24	24	27	56	55	33	67	67
	1000	6896	13820	14243	19974	37915	36581	27323	57846	58902
	10000	.	.	.	.	.	.	.	.	.
Theoret., asymptotic	100	1	1	2	2	4	5	3	5	8
	1000	25	49	58	76	148	175	126	245	290
	10000	2380	4710	5408	7476	14570	16662	12686	24682	27115
Theoret., exact	100	3	5	9	7	14	24	10	20	35
	1000	44	87	129	134	262	386	221	433	635
	10000	2605	5048	5712	7870	15412	17489	13365	25896	28671

Table 3: Approximate running time (in seconds) for sampling using one continuous, two continuous, or two continuous & one binary variable.

6. Conclusions

We have presented a simple greedy algorithm for representative sampling from a population or theoretical distributions. Although computational time can be very large for large samples and populations, the algorithm was created having small samples in mind. For small samples, stratified random sampling has practical limitations that are very difficult to overcome and the suggested deterministic method might be a better alternative in terms of:

- Complexity: Stratified random sampling might involve numerous trial and error steps to identify the practical limits of the method and choose the best possible approach.
- Controlling for more variables: Stratified random sampling rarely involves more than two variables, while there is no limit under the `repsample` algorithm (of course, the more variables are included the worse the matching will be for each variable individually).
- Outcome, i.e., 'representativeness': For small samples a random selection often fails to deliver a representative sample on all important variables, while the more deterministic repsample algorithm is more likely to find a better solutions for the problem and provides a sample that is closer to the population or hypothesized distributions.

The algorithm has certain limitations. First, users must realize that `repsample` does not guarantee a sample that is representative but one that is as representative as possible under the parameters specified. Second, the algorithm only indirectly takes into account relationships between the included variables, through closely matching each univariate distribution. Third, when quantifying overall 'representativeness' under Fisher's combined probability test the combined tests are assumed to be independent. Although, dependence can introduce anti-conservative bias to the test (i.e., p values for χ^2_F are biased towards zero) and methods have been developed to deal with the issue (Brown 1975; Kost and McDermott 2002), this is less of a problem in the context of a minimization algorithm. In `repsample` we are mainly interested in the relative ranking of the p values and not their absolute values; assuming the introduced bias is uniform across cases, rankings should be unaffected. Fisher's method does not even seem to be the best approach for combining independent p values (Davidov 2011) but its simplicity and computational speed make it attractive in this ranking context.

Acknowledgments

Many thanks to the two anonymous reviewers whose comments improved both the manuscript and the command significantly. The author was partly supported by a NIHR School for Primary Care Research fellowship in primary health care. The views expressed are those of the author and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

Begg CB, Iglewicz B (1980). "A Treatment Allocation Procedure for Sequential Clinical Trials." *Biometrics*, **36**(1), 81–90.

Bibby P, Shepherd J (2004). "Developing a New Classification of Urban and Rural Areas for Policy Purposes – The Methodology." *Technical report*, Office of National Statistics. URL `http://archive.defra.gov.uk/evidence/statistics/rural/documents/rural-defn/rural-urban-method.pdf`.

Blackwell M, Iacus S, King G, Porro G (2009). "**cem**: Coarsened Exact Matching in Stata." *Stata Journal*, **9**(4), 524–546.

Brown MB (1975). "Method for Combining Non-Independent, One-Sided Tests of Significance." *Biometrics*, **31**(4), 987–992.

Communities and Local Government (2011). "The English Indices of Deprivation 2010." *Technical report*, Department for Communities and Local Government. URL `http://www.communities.gov.uk/publications/corporate/statistics/indices2010technicalreport`.

Davidov O (2011). "Combining P-Values Using Order-Based Methods." *Computational Statistics & Data Analysis*, **55**(7), 2433–2444.

Fisher RA (1932). *Statistical Methods for Research Workers*. 4th edition. Oliver and Boyd, Edinburgh.

Gibbons JD, Chakraborti S (2011). *Nonparametric Statistical Inference*. 5th edition. Champman & Hall, Boca Raton.

Iacus SM, King G, Porro G (2009). "**cem**: Software for Coarsened Exact Matching." *Journal of Statistical Software*, **30**(9), 1–27. URL `http://www.jstatsoft.org/v30/i09/`.

Kost JT, McDermott MP (2002). "Combining Dependent P-Values." *Statistics & Probability Letters*, **60**(2), 183–190.

Mehta CR, Patel NR (1983). "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables." *Journal of the American Statistical Association*, **78**(382), 427–434.

Pocock SJ, Simon R (1975). "Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial." *Biometrics*, **31**(1), 103–115.

Rice WR (1990). "A Consensus Combined P-Value Test and the Family-Wide Significance of Component Tests." *Biometrics*, **46**(2), 303–308.

Rosenbaum PR, Ross RN, Silber JH (2007). "Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer." *Journal of the American Statistical Association*, **102**(477), 75–83.

Rosenbaum PR, Rubin DB (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, **70**(1), 41–55.

StataCorp (2011a). *Stata 12 Base Reference Manual.* StataCorp LP, College Station, TX.

StataCorp (2011b). *Stata Data Analysis Statistical Software: Release 12.* StataCorp LP, College Station, TX. URL http://www.stata.com/.

Whitlock MC (2005). "Combining Probability from Independent Tests: The Weighted Z-Method Is Superior to Fisher's Approach." *Journal of Evolutionary Biology*, **18**(5), 1368–1373.

Zar JH (1999). *Biostatistical Analysis.* 4th edition. Prentice Hall, Upper Saddle River.

Affiliation:

Evangelos Kontopantelis
Centre for Biostatistics & Centre for Primary Care
NIHR School for Primary Care Research
Institute of Population Health
University of Manchester
Williamson building 5th floor
M13 9PL, United Kingdom
E-mail: e.kontopantelis@manchester.ac.uk
URL: http://www.medicine.manchester.ac.uk/staff/EvanKontopantelis

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 55, Code Snippet 1 *Submitted:* 2012-06-26
November 2013 *Accepted:* 2013-06-08