Comparison of sample set curtailment procedures

Jason Legg
Center for Survey Statistics and Methodology
Iowa State University

Abstract

For many designs, there is a nonzero probability of selecting a sample that provides poor estimates for known quantities. Stratified random sampling reduces the set of such possible samples by fixing the sample size within each stratum. However, undesirable samples are still possible with stratification. Rejective sampling removes unwanted samples by only retaining a sample if specified functions of sample estimates are within a tolerance of known values. The resulting samples are often said to be balanced on the function of the variables used in the rejection procedure. Cube sampling, an alternative to rejective sampling, attempts to select a balanced sample with the same first-order inclusion probabilities as the original design. Through simulation, we compare estimation properties of a rejective sampling procedure to those of cube sampling for estimating a mean.

KEY WORDS: rejection sampling, cube sampling, stratification, balanced sampling

1. Introduction

When auxiliary data are known for the entire population, properties of estimators can be improved by incorporating the auxiliary information in the sample design. In one classic case, the model

$$\begin{aligned} y_i &= \beta x_i + e_i \\ e_i &\sim ind(0, \sigma^2 x_i) \end{aligned} \tag{1}$$

is assumed for the population, where y is an analysis variable and x is a known auxiliary variable. To estimate the total of y, the design of choice is probability proportional to size (PPS) sampling with first-order inclusion probabilities

$$\pi_i = \left(\sum_{i=1}^{N} x_i \right)^{-1} n x_i, \tag{2}$$

where n is the sample size and N is the population size. A second common case uses the model

$$\begin{aligned} y_{ij} &= \tau_i + e_{ij} \\ e_{ij} &\sim ind(0, \sigma_i^2), \end{aligned} \tag{3}$$

for $i = 1, \ldots, I$. When indicators for the I groups are known, stratified sampling with an optimal allocation sample size is used. For PPS sampling and stratified sampling with known x_i, samples that produce poor estimators have nontrivial positive probabilities. Stratification greatly reduces the sample space from simple random sampling. However, an excessive number of strata can be needed when the groups are defined from continuous auxiliary variables or a large number of grouping variables are crossed. For a continuous auxiliary variable satisfying a linear model, stratification loses some information.

Another way to incorporate information from an auxiliary variable in a design is balancing. A sample is balanced for x if

$$\widehat{T}_x = \sum_{i=1}^{n} \pi_i^{-1} x_i = \sum_{i=1}^{N} x_i = T_x. \tag{4}$$

A design is balanced for x if every sample with positive probability is balanced for x. Balancing for x gives a regression property to Horvitz-Thompson estimators. That is,

$$\begin{aligned} \widehat{T}_{y,reg} &= \textstyle\sum_{i=1}^{n} \pi_i^{-1} y_i + (T_x - \widehat{T}_x)\widehat{\beta} \\ &= \textstyle\sum_{i=1}^{n} \pi_i^{-1} y_i \\ &= \widehat{T}_y, \end{aligned} \tag{5}$$

where $\widehat{\beta}$ is an estimator, such as

$$\widehat{\beta} = \widehat{V}(\widehat{T}_x)^{-1} \widehat{Cov}(\widehat{T}_x, \widehat{T}_y). \tag{6}$$

If

$$\begin{aligned} y_i &= \boldsymbol{x}_i' \boldsymbol{\beta} + e_i \\ e_i &\sim ind(0, \sigma_i^2), \end{aligned} \tag{7}$$

the design variance of $\widehat{T}_y$ is the design variance of $\widehat{T}_e$ under perfect balance on $\boldsymbol{x}$. For highly correlated variables x and y, the design variance of $\widehat{T}_e$ can be much smaller than the design variance of $\widehat{T}_y$. Balancing can be thought of as calibration by design. Balanced sampling has been proposed by Royall and Cumberland (1981) as a way to reduce model bias from incorrectly specified polynomial superpopulation models. Tillé and Deville (2004) investigated methods of selecting balanced samples.

In practice, finding a balanced design is infeasible for most data. Very tight balancing can lead to a design with some joint inclusion probabilities that are zero, called an unmeasurable design. If a design is unmeasurable, there does not exist a design unbiased variance estimator. Even if the design is measurable, some joint probabilities can be very small due to balancing. Some balancing procedures can lead to small first-order inclusion probabilities and high variability in weights. Rather than reduce the sample space to few possible samples, partial balancing

is done in practice (Tillé 2006 and Fuller 2007). Two easily implementable methods have been developed to select a nearly balanced sample: rejection sampling and cube sampling. We compare design properties of these two methods using simulations.

2. Balanced Sampling Procedures

Rejection sampling involves discarding any sample that does not meet a specified balancing tolerance and is the method used by Royall and Herson (1973). Fuller (2007) presents a condition for rejecting a sample. A sample is selected under any design and retained if

$$(\widehat{\boldsymbol{T}}_{\boldsymbol{x}} - \boldsymbol{T}_{\boldsymbol{x}})'[V(\widehat{\boldsymbol{T}}_{\boldsymbol{x}})]^{-1}(\widehat{\boldsymbol{T}}_{\boldsymbol{x}} - \boldsymbol{T}_{\boldsymbol{x}}) < \gamma^2 \qquad (8)$$

for some constant $\gamma > 0$, where $\widehat{\boldsymbol{T}}_{\boldsymbol{x}}$ is the Horvitz-Thompson total estimator for $\boldsymbol{x}$,

$$V(\widehat{\boldsymbol{T}}_{\boldsymbol{x}}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\boldsymbol{x}_i\boldsymbol{x}_i'\pi_i^{-1}\pi_j^{-1}, \qquad (9)$$

and π_{ij} is the joint inclusion probability of unit i and unit j. Otherwise, the sample is rejected, a new sample is selected under the initial design, and condition (8) is checked for the new sample. If the original design has a central limit theorem, the left side of (8) is asymptotically a χ^2 random variable with degrees of freedom equal to the number of auxiliary variables. For application, an approximate rejection rate can be set using the quantiles of a χ^2 distribution for γ^2.

The cube method was developed by Tillé and Deville and is described in Tillé (2006). The cube method attempts to select a balanced sample with predetermined first-order inclusion probabilities. If the first-order inclusion vector does not lead to a balanced design, an additional step of minimizing a cost constraint is used. Unlike the rejection procedure, higher order initial inclusion probabilities are not prespecified. The cost minimization step maintains the specified initial first-order inclusion probabilities.

As a way to understand the cube procedure, Tillé (2006) describes sampling geometrically. The set of all possible samples is defined to be the set of vectors for vertices of an N dimensional unit cube. For example, if $N = 3$, the vertex $(0, 1, 1)$ denotes a sample containing units two and three. Using the balancing equation (4) and desired π_i for $i = 1, \ldots, N$, a balancing plane is created. Any sample where the balancing plane intersects a vertex of the unit N dimensional cube is a balanced sample. The design is balanced if every point of intersection between the balancing plane and the unit cube is a vertex of the unit cube.

The cube sampling procedure begins by selecting a vector on the balancing plane. Using a balancing martingale, a random walk from the initial point to an edge of the unit cube is done. Tillé refers to the random walk step as the flight phase. If the edge point at the end of the random walk is a vertex of the unit cube, the sample is selected. Otherwise, a cost minimization procedure is used to convert the fractional components of the edge vector to integers. The integer components of the edge vector are not changed in the cost minimization step. Tillé refers to the cost minimization step as the landing phase.

Other procedures besides rejection and cube sampling can be used to obtain nearly balanced samples. Stratification can use boundaries determined by the size of the x variables (Fuller 1981). A time consuming procedure of enumerating all possible samples and assigning probabilities could be implemented for small populations. The main drawback with sample enumeration and balancing through stratification is the lack of clear extensions to a large number of auxiliary variables. The decision for the number of variables to use in the rejection and cube sampling procedures is the same as deciding how many variables to include in a regression estimator.

Software has been developed for selecting cube samples. For rejection sampling, standard software packages can be used to select a sample and compute (8). A loop needs to be written to complete the procedure. Programs for selecting cube samples have been written for SAS and R. See Rousseau and Tardieu, 2004 for SAS and Tillé and Matei (2005) for R. A faster SAS-IML macro has been developed and is available at http://www2.unine.ch/Jahia/site/statistics/op/edit/pid/10891. The R program available in the *sampling* library was used in the simulations in this paper. Because the cost minimization step of cube sampling is computationally intensive, for more than 20 balancing variables, a variable suppression step is recommended for the landing phase.

3. Inclusion Probabilities

Both rejection sampling and cube sampling require initial first-order inclusion probabilities as inputs. The actual first-order inclusion probabilities are different than the initial values for rejection sampling. For rejection sampling, units closer to the population mean will have a higher selection probability than units far from the mean. Cube sampling has explicit control on the first-order inclusion probabilities.

To illustrate differences between initial and final inclusion probabilities, samples of size 20 from a population of 100 units were simulated. The population of x-values was generated as random variables from a standard normal distribution. The rejection procedure used simple random sampling as the initial design. First-order inclusion probabilities were estimated using a Monte Carlo simulation of size 100,000 (Figure 1). The lines are smoothing

spline fits. An approximate 90% rejection rate was used for the rejection sampling. From rejection sampling theory, first-order inclusion probabilities are approximately a quadratic function of the distance $x_i - \overline{x}_N$ for an equal probability initial sample design. Changes to the first-order inclusion probabilities from 0.2 are not detected for the cube sample design.

The joint inclusion probabilities for the rejection sampling procedure differ from those of the initial design. A pair of units i and j are likely to have a high joint inclusion probability if $x_i + x_j - 2\overline{x}_N$ is close to zero for an equal probability initial sample design. The joint inclusion probabilities were estimated for the simulation of samples of size 20 from 100 (Figure 2). The initial joint inclusion probability for simple random sampling is 0.038. The rejection sampling joint inclusion probabilities are approximately a quadratic function of $x_i + x_j$. The cube sampling joint inclusion probability field appears to have sharper angles than the rejection joint inclusion probabilities. The high joint inclusion probabilities for the cube are associated with pairs of units that are on the far opposite sides of $\overline{x}_N$.

Given the first and second-order inclusion probabilities, the Horvitz-Thompson estimator using the initial inclusion probabilities under rejection sampling is biased while the Horvitz-Thompson estimator under cube sampling is unbiased. However, the standard Horvitz-Thompson variance estimator is biased for both procedures. Using Monte Carlo methods, the inclusion probabilities can be estimated so that nearly unbiased Horvitz-Thompson estimators can be used. However, for a large population size, simulating enough samples to give a precise estimate of the joint inclusion probability for each pair of units can be impractical. A small sample and a model for the first and second-order inclusion probabilities can be used instead of directly estimated inclusion probabilities. An alternative approach is to use a regression estimator and the variance of the regression estimator since balancing is similar to regression through design. Fuller (2007) gives conditions for the consistency of the regression estimator and variance estimator for rejection sampling. We investigate the regression estimator with another simulation.

4. Simulation

A population of size 100 was generated from the model

$$\begin{array}{rcl} y_i & = & x_i + 0.55x_i^2 + x_i e_i \\ e_i & \sim & iidN(0, 0.4), \end{array} \tag{10}$$

where the x_i are fixed values in the range of 0 to 4 (Figure 3). The population was held fixed after initial selection.

The regression estimator is

$$\overline{y}_{reg} = \overline{\boldsymbol{z}}_N \widehat{\boldsymbol{\beta}}, \tag{11}$$

where $\boldsymbol{z}$ is a vector of auxiliary variables containing the design variables and x_i,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{z}_i \phi_i \pi_i^{-2} \boldsymbol{z}_i' \right)^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i \phi_i \pi_i^{-2} y_i, \tag{12}$$

$\overline{\boldsymbol{z}}_N$ is the population mean of $\boldsymbol{z}_i$, and ϕ_i are constants similar to the finite population correction factors. The $\boldsymbol{z}_i$ depend on the initial design used for rejection sampling.

We consider the cases of Poisson sampling and stratified random sampling with two units per stratum selected as initial designs. Strata were determined by setting the boundary so that the within stratum sum of sorted x_i was roughly equal for all strata. The sample size was set to be 20, so ten strata were formed. The stratum sizes were 35, 15, 11, 9, 8, 7, 5, 4, 3, and 3. The rejection procedure uses a stratified two-per stratum sample selection with equal inclusion probability within a stratum. Only the x variable is used in the rejection balancing step. For cube sampling, the vector of balancing variables is the vector of stratum indicators and x. The $\boldsymbol{z}$ vector for stratification contains $H = 10$ stratum indicator variables,

$$z_{hi} = \begin{cases} 1 & \text{unit } i \text{ in stratum } h \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

for $h = 1, 2, \ldots, 10$, and x_i.

Initial selection probabilities for Poisson sampling with sample size 20 were set to the values in the initial stratified sampling design. The vector of balancing variables is $(1, \pi_i, x_i, (1-\pi_i)^{-1}\pi_i)$. The first variable controls population size estimators, the second variable controls sample size, the third variable balances on x, and the fourth variable is necessary so that the regression estimator is design consistent. The rejection rate was set at 90% for stratified sampling and 90% and set at 95% for Poisson initial sampling.

For Poisson sampling, $\phi_i = (1 - \pi_i)$. The variance estimator for $\overline{y}_{reg}$ is the variance estimator of $\overline{e}$, the mean of $\widehat{e}_i = y_i - \boldsymbol{z}_i' \widehat{\boldsymbol{\beta}}$, based on the initial design. For Poisson sampling, the variance estimator used is

$$\begin{array}{rcl} \widehat{V}(\overline{y}_{reg}) & = & (n-k)^{-1} n \overline{\boldsymbol{z}}_N' \boldsymbol{M}_{zz}^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i \pi_i^{-4} \\ & & \times (1-\pi_i)^3 \widehat{e}_i^2 \boldsymbol{z}_i' \boldsymbol{M}_{zz}^{-1} \overline{\boldsymbol{z}}_N, \end{array} \tag{14}$$

where

$$\boldsymbol{M}_{zz} = \sum_{i=1}^{n} \boldsymbol{z}_i \pi_i^{-2} (1-\pi_i) \boldsymbol{z}_i', \tag{15}$$

and k is the number of variables in $\boldsymbol{z}$.

For stratified random sampling with two-per stratum, $\phi_i = (N_h - 1)^{-1}(N_h - 2)$. The variance estimator used for stratified sampling is

$$\begin{array}{rcl} \widehat{V}(\overline{y}_{reg}) & = & (H-1)^{-1} H \sum_{h=1}^{H} [(1-f_h)^{1/2} \{0.5 W_h \\ & & + (\overline{\boldsymbol{z}}_N - \overline{\boldsymbol{z}}) \boldsymbol{M}_{zz}^{-1} \phi_h W_h^2 (\boldsymbol{z}_{h1} - \boldsymbol{z}_{h2})\} \\ & & \times (\widehat{e}_{h1} - \widehat{e}_{h2})]^2, \end{array} \tag{16}$$

where $W_h = n_h/N_h$, ϕ_h is ϕ_i for units in stratum h, $\boldsymbol{z}_{hi}$ is the auxiliary variable vector i in stratum h,

$$\widehat{e}_{hi} = y_{hi} - \overline{y}_h - (\boldsymbol{z}_{hi} - \overline{\boldsymbol{z}}_h)\widehat{\boldsymbol{\beta}}, \tag{17}$$

$\overline{y}_h$ and $\overline{\boldsymbol{z}}_h$ are stratum means of y_{hi} and $\boldsymbol{z}_{hi}$, respectively, and $H = 10$ is the number of strata.

Horvitz-Thompson estimators were constructed using initial inclusion probabilities. The variance estimators used for cube sampling were the same as for rejection sampling. Confidence intervals were constructed using a normal approximation for the distribution of the regression estimator. The number of samples selected was 30,000 for each Monte Carlo simulation.

The variance of the Horvitz-Thompson mean under Poisson sampling with a sample size of 20 with no balancing is 0.080. Results reported in Table 1 are standardized by the Horvitz-Thompson variance. Thus the variance of the Horvitz-Thompson estimator for cube sampling is 0.080(0.142) and the bias of the Horvitz-Thompson estimator is $\sqrt{0.080}(-0.0020)$. The regression estimator is superior to the Horvitz-Thompson estimator for both rejection and cube sampling. The gain from using the regression estimator is larger for rejection than for the cube method, likely due to the cube method achieving tighter balance than the rejection method. The biases in the regression estimators are negligible relative to the variances.

The cube sampling procedure has a slightly smaller variance than that of the rejection design with a 90% rejection rule. The rejection procedure with an approximate 95% rejection rate has a slightly smaller variance than the cube design. Increasing the rejection rate increased the bias of the regression estimator. The bias of the regression estimator for 95% rejection is significant, but negligible relative to the variance.

Hajek ratio means were also computed using the initial first-order inclusion probabilities. The Hajek means differed little from the Horvitz-Thompson means due to high control on estimated population sizes and, hence are not reported.

The Monte Carlo average of the Horvitz-Thompson variance estimator, $\widehat{V}(\overline{y}_{HT})$, is close to the variance under the initial design. Tillé (2006) recommends several variance estimators based on a Poisson sampling approximation with corrections for known constraints in the design variance. We found the regression variance estimators in equations (14) and (16) performed better for this simulation than the third estimator in Tillé (2006, p. 171). Deville and Tillé (2005) recommend the fourth estimator on page 171 in Tillé (2006), but that estimator requires solving a nonlinear equation system, which would have been computationally expensive to add to the simulation. Matei and Tillé (2005) recommend variance estimator 1, which is a regression variance estimator using a Poisson sampling design variance as an approximation. However, estimator 1 also performs poorly in this simulation setting due to the sample size and number of auxiliary variables. Given the strong performance of the regression variance estimators, we prefer the regression variance estimator to the Tillé suggested variance estimators.

Table 1: Properties of samples based on Poisson sampling of size 20

	Cube	Rej. 90%	Rej. 95%
$bias(\overline{y}_{HT})$	-0.002	-0.016	-0.007
$bias(\overline{y}_{reg})$	-0.002	0.002	0.005
$V(\overline{y}_{HT})$	0.142	0.270	0.220
$V(\overline{y}_{reg})$	0.131	0.136	0.129
$\widehat{V}(\overline{y}_{HT})$	0.995	0.989	0.991
$\widehat{V}(\overline{y}_{reg})$	0.122	0.123	0.121
95% CI rate	94.5%	93.9%	95.5%

Table 2: Properties of samples based on stratified sampling of size 20

	Cube	Rej. 90%
$bias(\overline{y}_{HT})$	-0.028	0.014
$bias(\overline{y}_{reg})$	-0.013	0.014
$V(\overline{y}_{HT})$	0.910	0.866
$V(\overline{y}_{reg})$	0.929	0.865
$\widehat{V}(\overline{y}_{HT})$	1.022	0.998
$\widehat{V}(\overline{y}_{reg})$	0.907	0.881
95% CI rate	94.5%	93.7%

The variance of the Horvitz-Thompson mean under the initial stratified design is 0.011. Estimates of Table 2 are standardized by the Horvitz-Thompson variance. The gain from balancing on x is not large since stratification has curtailed the sample space. Likewise, the regression estimator is not a noticeable improvement over the Horvitz-Thompson estimator. The rejection sampling design slightly outperforms the cube sampling design in terms of variance, likely due to the joint inclusion probabilities in the initial design for rejection sampling.

To assess large sample properties of the balancing procedures, the size of the Poisson simulation was quadrupled. The population was replicated four times and a sample of size 80 was selected. The Horvitz-Thompson variance of a mean under the Poisson design is 0.020. The resulting relative variances and biases were close to the results for samples of size 20 (Table 3).

The simulation results agree with the theoretical result that the regression estimator is an $O_p(n^{-1/2})$ estimator after rejection of the type used in this paper. From the simulations, it appears that the regression estimator after cube sampling possesses similar properties.

Table 3: Properties of samples based on Poisson sampling of size 80

	Cube	Rej. 90%
$bias(\overline{y}_{HT})$	0.002	-0.006
$bias(\overline{y}_{reg})$	0.002	0.000
$V(\overline{y}_{HT})$	0.127	0.267
$V(\overline{y}_{reg})$	0.122	0.124
$\widehat{V}(\overline{y}_{HT})$	0.996	0.994
$\widehat{V}(\overline{y}_{reg})$	0.121	0.121
95% CI rate	96.0%	94.8%

5. Discussion

Rejection sampling and cube sampling produced roughly equally performing regression estimators. Balancing provides major gains when the initial design provides little control on the auxiliary values entering samples. A well stratified sample design provides many of the benefits of balancing on a continuous variable. However, further balancing after stratification can lead to smaller variances for estimators. In practice, the additional gains from balancing after large amounts of stratification are likely not large.

For the simulations, the rejection rate was fixed at 90% for the larger population. When the population and sample sizes are increased, the rejection rate can be increased while still maintaining a large set of possible samples. From further analysis, the bias of the regression estimator remain negligible for rejection rates near 99%. The marginal variance reduction due to balancing decreases as the balancing condition is tightened. It is possible to have populations where balancing procedures result in some loss in estimator precision. Therefore, choice of balancing variables remains important.

Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Deville, J.-C., and Tillé,Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference* 128, 569-591.

Fuller, W. A. (1981). An empirical Study of the ratio estimator and estimators of its variance: comment. *Journal of the American Statistical Association* 76: 78-80.

Fuller, W. A. (2007). Some design properties of a rejective sampling procedure. *Unpublished*.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics* 21: 543-570.

Royall, R. M. and Cumberland, W. G.(1981). The finite-population linear regression estimator and estimators of its variance–An empirical study. *Journal of the American Statistical Association* 76: 924-930.

Rousseau, S. and Tardieu,F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. *Technical report, INSEE, PARIS*.

Royall, R. M. and Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association* 68: 880-889.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer Science+Business Media, Inc.

Tillé, Y. and Matei, A. (2005). The R package Sampling. *The Comprehensive R Archive Network, http://cran.r-project.org/, Manual of the Contributed Packages*.

Figure 1: Simulated first-order inclusion probabilities

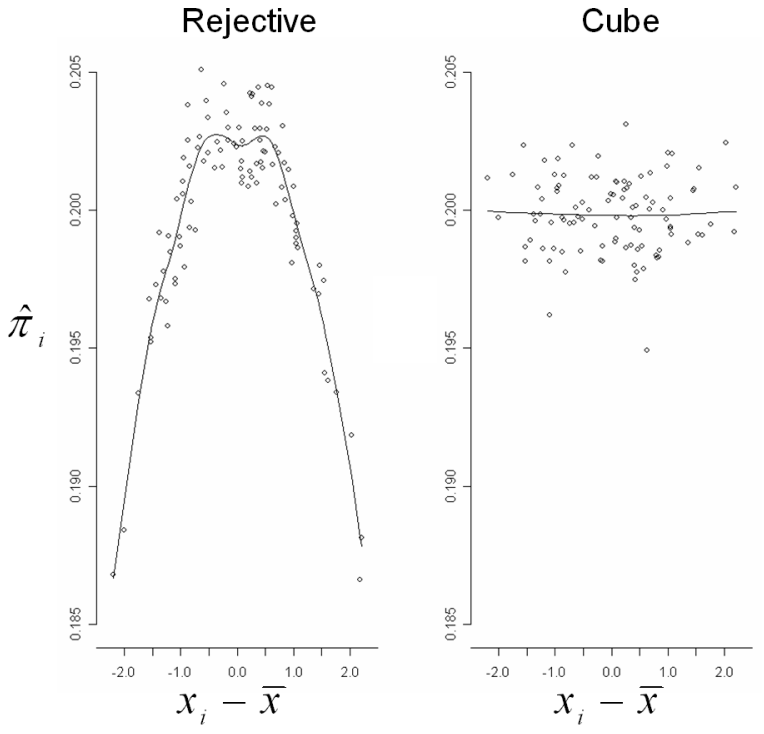

Figure 2: Simulated second-order inclusion probabilities

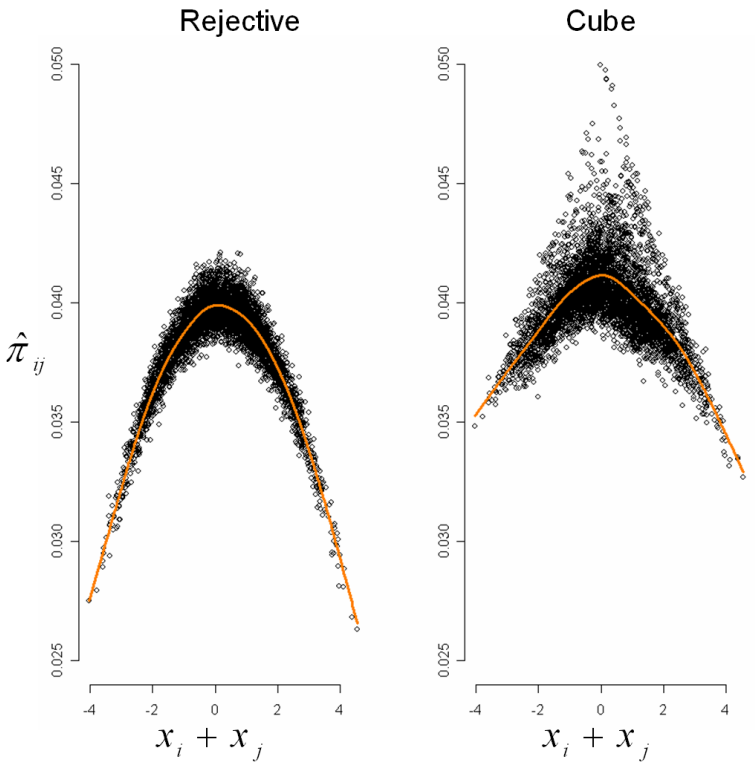

Figure 3: Simulation population