Some Methods of Model-Based Sampling

SungJoon Hong
SoHyung Park
SunWoong Kim
HongYup Ahn
Steven G. Heeringa

Dongguk University
&
Survey Research Center, University of Michigan

Overview

- **Background**
- **Mechanism of Conventional Design-Based πPS Sampling**
- **Mechanism of Model-Based πPS Sampling**
- **The Previous Studies of Model-Based πPS Sampling**
- **New Model-Based πPS Sampling**
- **Empirical Studies**
- **Concluding Remarks**

Background

- Since Hansen and Hurwitz (1943), a large number of sampling techniques with unequal probabilities have been developed.

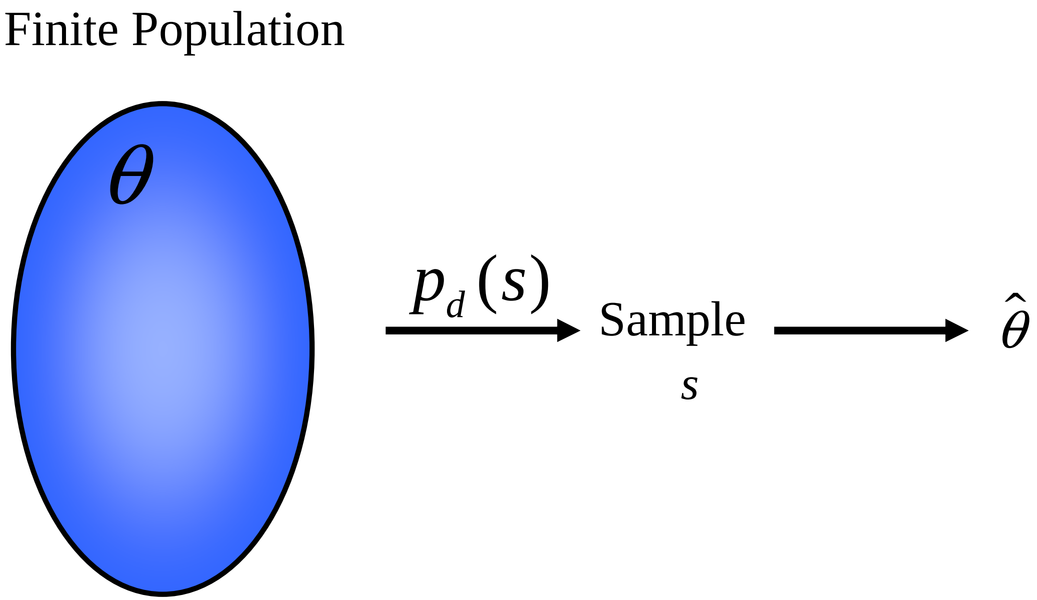
- Any number of methods are preferred in the literature for comparison purposes, whereas only a few methods, including inclusion probability proportional to size (πPS) sampling methods, are widely used for practical uses among the samplers, and none of them have yet had the general acceptance.

- As shown in many empirical studies, this may be due to the fact that the variance of estimates of interest according to sampling methods is quite sensitive to population characteristics, especially in the case of selecting a small sample from a small population.

- The selection of a small sample from a small population is not unusual in practice. In many national surveys there are many strata, and a few units are sampled in a stratum. For example, two per stratum is a common situation in nationwide samples.

- Accordingly, it is desirable to develop the sampling methods whose accuracy is less sensitive, possibly consistent, to population characteristics in the case of a small sample.

Mechanism of Conventional Design-Based πPS Sampling (One-Step Sampling)



Finite Population

θ

$p_d(s)$

Sample s

$\hat{\theta}$

Estimators

A generalized regression (GREG) estimator may be one of the useful estimators. But it is well-known that it might be appreciably biased for a small sample, although the bias is in modest for large samples.

As an alternative, the Horvitz-Thompson (H-T) (1952) estimator, which is unbiased for the population total, is acceptable. It is highly efficient under a good πPS sampling scheme.

$\theta = Y$: Population Total

$\hat{\theta} = \hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$: H-T estimator

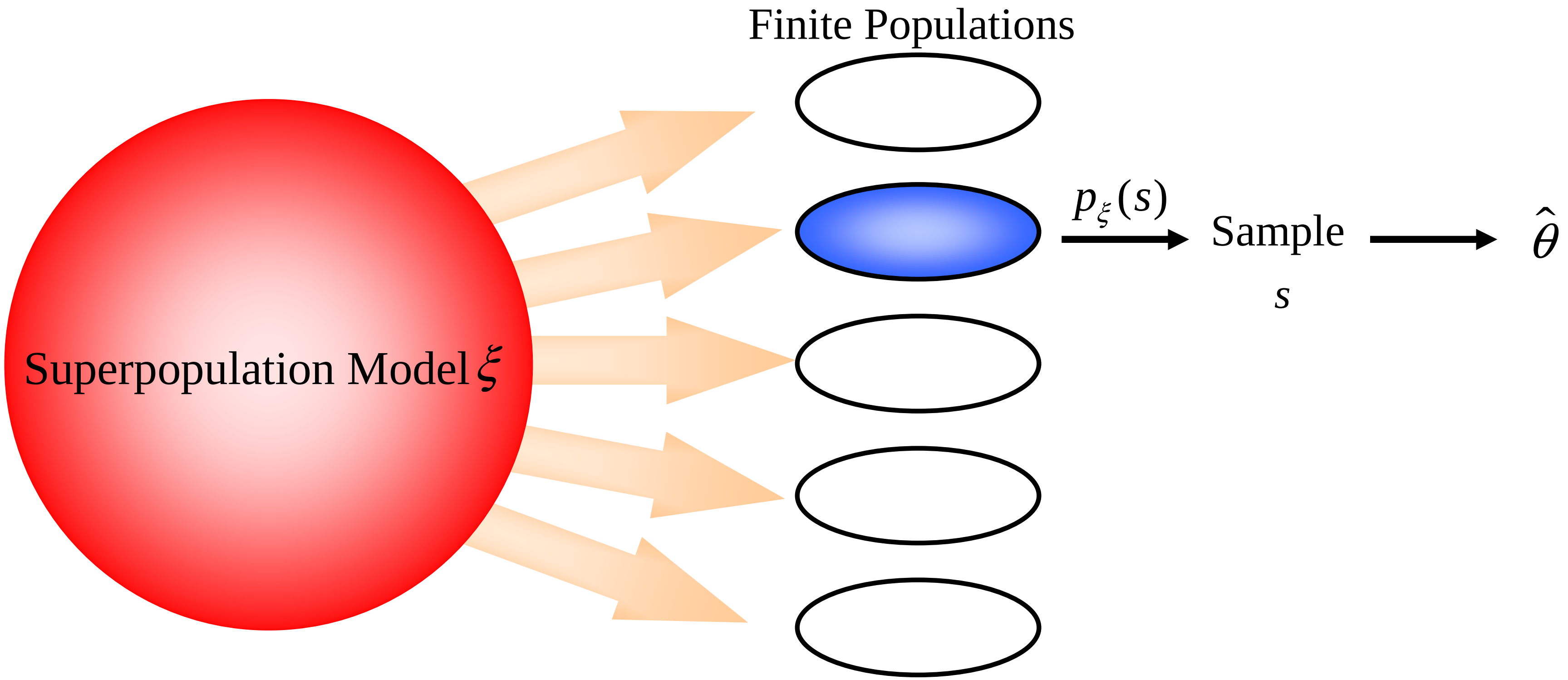
Sampling Design $p_d(s)$

The samplers often prefer $p_d(s)$ yielding a smaller design variance than in other sampling methods, as given by $Var_d(\hat{\theta}) < Var_{d^*}(\hat{\theta})$

One may choose the sampling methods of Midzuno (1952), Murthy (1957), Brewer (1963), and Sampford (1967), based on the empirical studies in the past. The first, third, and fourth are design-based πPS sampling methods. The second, third, and forth are especially available in SPSS and SAS.

But it is questionable if the sampling designs implemented by those will always lead to a lower design variance for a population of interest.

Mechanism of Model-Based πPS Sampling (Two-Step Sampling)



Why superpopulation model in the actual selection of a sample?

At the design stage, we consider the available knowledge, the assumed superpopulation model ξ, from the point of view of the strategy of reducing the design variance.

With respect to inference, the anticipated variance (AV), introduced by Isaki and Fuller (1982), is used as a measure describing the variability between the total and the estimator of the total under both the design and superpopulation model. If the H-T estimator is used, it becomes simply

$$E_{\xi}E_{d}\left[\left(\hat{Y}_{HT}-Y\right)^{2}\right]=E_{\xi}\left[Var_{d}\left(\hat{Y}_{HT}\right)\right]$$

where both Y and $\hat{Y}_{HT}$ are random variables. The AV is the model expectation of the design variance, or simply the average variance.

Optimal Sampling Design $p_{\xi}(s)$

Now we would like to decide an optimal sampling design in a set of possible πPS sampling designs to achieve:

$$\textbf{Minimize } E_{\xi}\left[Var_{d}\left(\hat{Y}_{HT}\right)\right]$$

Note that we focus on the design stage for the selection of a sample from a finite population, not the estimation stage, and the H-T estimator does not involve a superpopulation model, different from the GREG estimator. Thus, although the superpopulation model is assumed, interest still lies in reducing $Var_{d}\left(\hat{Y}_{HT}\right)$ in practice.

The Previous Studies of Model-Based πPS Sampling

Kim, Heeringa, and Solenberger (2006) first introduced the theory of model-based πPS methods based on the average variance of the H-T estimator.

Superpopulation Model

We assume that the finite population is a random sample from an infinite superpopulation in which

$$y_i = \alpha + \beta x_i + \varepsilon_i \,,$$

where $E_\xi(\varepsilon_i) = 0$, $V_\xi(\varepsilon_i) = \sigma^2 x_i^\gamma$, and $E_\xi(\varepsilon_i \varepsilon_j) = 0$

Here E_ξ denotes the model expectation over all the finite populations that can be drawn from the superpopulation.

Average Variance (AV)

Under a form of the variance of the H-T estimator expressed as

$$Var_I(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{y_i^2(1-\pi_i)}{\pi_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - 2\sum_{i=1}^{N}\sum_{j>i}^{N} y_i y_j \,,$$

the AV is given by

$$\begin{aligned} E_\xi\left(Var_I(\hat{Y}_{HT})\right) = {} & \frac{\left(\sum_i x_i\right)^2}{n}\left[\frac{2\alpha}{n}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\alpha+\beta(x_i+x_j)}{x_i x_j}\pi_{ij} + \beta^2(n-1)\right] \\ & + \sum_{i=1}^{N}\left(\sum_i x_i / n x_i - 1\right)\left(\sigma^2 x_i^\gamma + \alpha^2 + \beta^2 x_i^\gamma + 2\alpha\beta x_i\right) \\ & - 2\sum_{i}^{N}\sum_{j>i}^{N}\left(\alpha^2 + \alpha\beta(x_i+x_j) + \beta^2 x_i x_j\right) \end{aligned}$$

With respect to a different form of the variance of the H-T estimator denoted by

$$Var_{II}(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2,$$

the AV is given by

$$E_\xi\left(Var_{II}(\hat{Y}_{HT})\right) = \frac{\sigma^2\left(\sum_i x_i\right)^\gamma}{n}\sum_{i=1}^{N}(1 - n\frac{x_i}{\sum_i x_i})\left(\frac{x_i}{\sum_i x_i}\right)^{\gamma-1}$$

$$+\ 2\alpha\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j - x_i\right)\left(\alpha x_i^{-1} + \beta\right)\right)$$

$$+\ \frac{2\alpha\left(\sum_i x_i\right)^2}{n^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j^{-1} - x_i^{-1}\right)\left(\alpha x_i^{-1} + \beta\right)\pi_{ij}$$

,

Optimization Problems (OP)

Since some terms in each AV are the known values, the minimization of each AV is equivalent to the following optimization problems.

Note that the linear constraints are for the πPS property, the nonnegativity, and the stability of estimated variances.

OP 1:

$$\text{Minimize} \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{\alpha + \beta(x_i + x_j)}{x_i x_j} \sum_{i,j \in s} p_\xi(s)$$

subject to

$$\pi_i = \sum_{i \in s} p_\xi(s), \quad i = 1, \cdots, N$$

$$c\pi_i\pi_j \le \sum_{i,j \in s} p_\xi(s) \le \pi_i\pi_j,$$

where c is a real number between 0 and 1

OP 2:

$$\text{Minimize } \sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j^{-1}-x_i^{-1}\right)\left(\alpha x_i^{-1}+\beta\right)\sum_{i,j\in s} p_\xi(s)$$

subject to

$$\pi_i = \sum_{i\in s} p_\xi(s), \quad i=1,\cdots,N$$

$$c\pi_i\pi_j \le \sum_{i,j\in s} p_\xi(s) \le \pi_i\pi_j,$$

The sampling design $p_\xi(s)$ is the solution to each optimization problem. A sample of n distinct units is selected with the probability of $p_\xi(s)$, and it is called the whole sample procedure.

New Model-Based πPS Sampling

Derive more factors not depending on the joint probabilities, π_{ij}, by extending the terms in the AV considered in the previous study, and simplify or elaborate optimization problems, relative to OP 1 and OP 2.

Alternative Forms of AV

$$
\begin{aligned}
AV_{III} = E_{\xi}\left(Var_{I}(\hat{Y}_{HT})\right) = {} & \frac{2\alpha^{2}\left(\sum_{i} x_{i}\right)^{2}}{n^{2}} \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{x_{i}x_{j}}\pi_{ij} \\
& +2\alpha\beta N\frac{n-1}{n}\sum_{i} x_{i} + \beta^{2}\frac{n-1}{n}\left(\sum_{i} x_{i}\right)^{2} \\
& + \sum_{i=1}^{N}\left(\sum_{i} x_{i}/nx_{i} - 1\right)\left(\sigma^{2}x_{i}^{\gamma} + \alpha^{2} + \beta^{2}x_{i}^{\gamma} + 2\alpha\beta x_{i}\right) \\
& - 2\sum_{i}^{N}\sum_{j>i}^{N}\left(\alpha^{2} + \alpha\beta(x_{i}+x_{j}) + \beta^{2}x_{i}x_{j}\right)
\end{aligned}
$$

Note that AV_{I} is a form of $A+B+C+D$, whereas AV_{III} is a form of $A^{'}+A^{''}+B+C+D$

$$AV_{IV} = E_{\xi}\left(Var_{II}(\hat{Y}_{HT})\right)$$

$$= \frac{\sigma^{2}\left(\sum_{i} x_i\right)^{\gamma}}{n} \sum_{i=1}^{N}(1 - n\frac{x_i}{\sum_{i} x_i})\left(\frac{x_i}{\sum_{i} x_i}\right)^{\gamma-1} + 2\alpha\left(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(x_j - x_i\right)\left(\alpha x_i^{-1} + \beta\right)\right)$$

$$+ \frac{2\alpha\left(\sum_{i} x_i\right)^{2}}{n^{2}}\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\left\{\alpha\left((x_i x_j)^{-1} - x_i^{-2}\right) - \left(2\beta x_i^{-1}\right)\right\}\pi_{ij} + \beta Nn(n-1)(\sum_{i} x_i)^{-1}\right]$$

,

Note that AV_{II} is a form of $E + F + G$, whereas AV_{IV} is a form of $E + F + G^{'} + G^{"}$.

New Optimization Problems

OP 3:

$$\text{Minimize } \sum_{i=1}^{N}\sum_{j>i}^{N}(x_i x_j)^{-1}\sum_{i,j\in s} p_\xi(s)$$

subject to

$$nx_i / \sum x_i = \sum_{i\in s} p_\xi(s), \;\; i = 1, \cdots, N$$

$$cx_i x_j /(\sum x_i)^2 \le \sum_{i,j\in s} p_\xi(s) \le x_i x_j /(\sum x_i)^2,$$

where c is a real number between 0 and 1

The objective function is a simple form, which does not depend on α or β, as well as σ^2 or γ. With the model without any assumption on the error term, Raj (1956) derived the same form.

OP 4:

$$\text{Minimize} \quad \sum_{i=1}^{N}\sum_{j>i}^{N}\left\{\alpha\left((x_i x_j)^{-1} - x_i^{-2}\right) - \left(2\beta x_i^{-1}\right)\right\}\sum_{i,j\in s} p_\xi(s)$$

subject to

$$nx_i / \sum x_i = \sum_{i\in s} p_\xi(s), \quad i = 1,\cdots,N$$

$$cx_i x_j /(\sum x_i)^2 \le \sum_{i,j\in s} p_\xi(s) \le x_i x_j /(\sum x_i)^2$$

The objective function, which is a bit complicated form involving $x_i x_j$, depends on α or β.

Empirical Studies

18 Natural Populations in Rao and Bayless (1969) were used in the evaluation of unequal probability sampling methods for the case of $n = 2$.

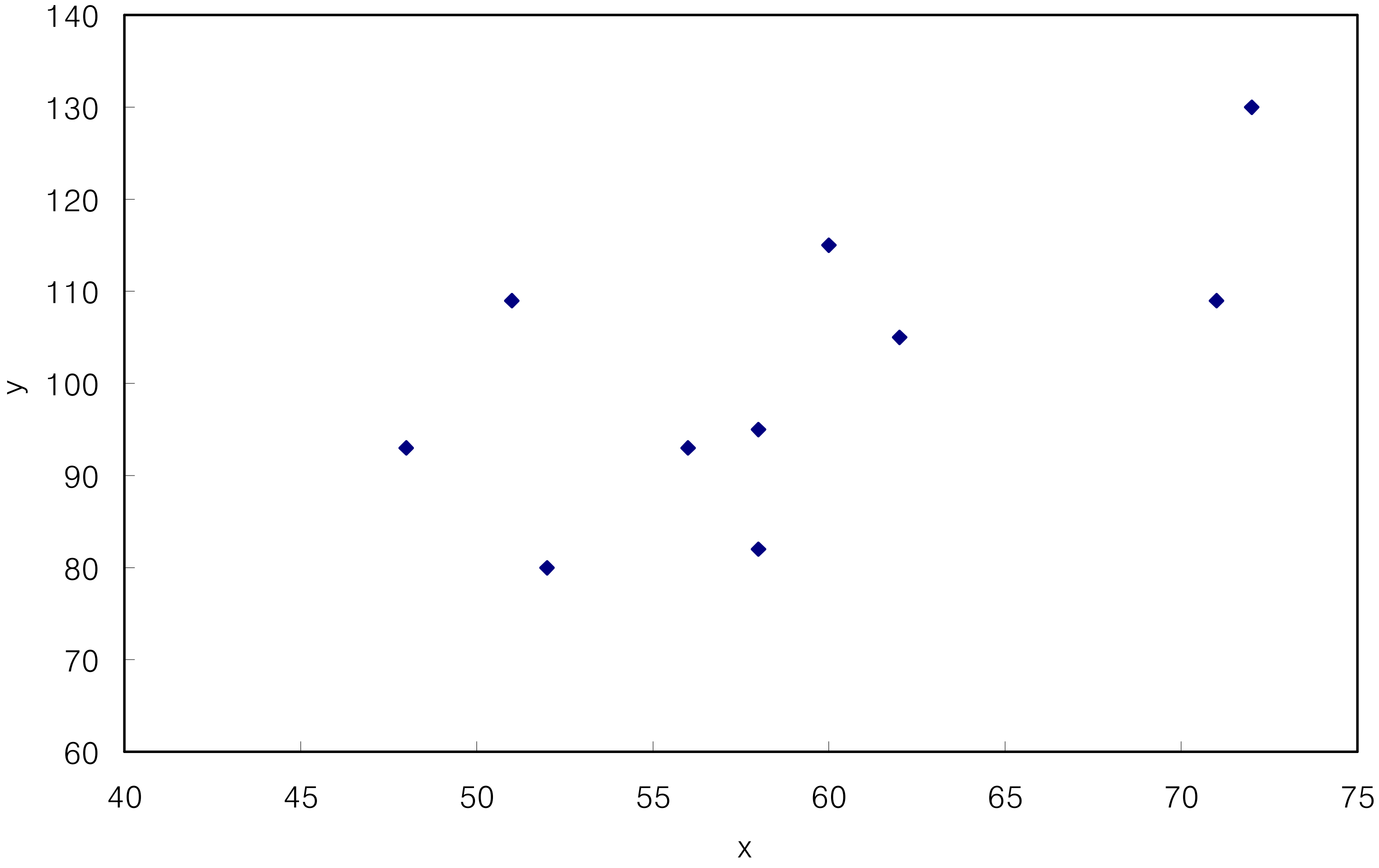
No	Source	y	x	N
1	Horvitz and Thompson (1952)	No. of Households	Eye-estimated no. of Households	20
2	Des Raj (1965)	No. of Households	Eye-estimated no. of Households	20
3	Rao (1963)	Corn acreage in 1960	Corn acreage in 1958	14
4	Kish (1965)	No. of rented dwelling units	Total no. of dwelling units	10
5	Kish (1965)	No. of rented dwelling units	Total no. of dwelling units	10
6	Hanurav (1967)	Population in 1967	Population in 1957	20

No	Source	y	x	N
7	Hanurav (1967)	Population in 1967	Population in 1957	16
8	Hanurav (1967)	Population in 1967	Population in 1957	17
9	Cochran (1963)	No. of persons per block	No. of rooms per block	10
10	Cochran (1963)	No. of people In 1930	No. of people In 1920	16
11	Cochran (1963)	No. of people In 1930	No. of people In 1920	16
12	Cochran (1963)	No. of people In 1930	No. of people In 1920	17
13	Sukhatme (1954)	No. of wheat acres in 1937	No. of wheat acres in 1936	10
14	Sukhatme (1954)	No. of wheat acres in 1937	No. of wheat acres in 1936	10

No	Source	y	x	N
15	Sampford (1962)	Oats acreage in 1957	Oats acreage in 1947	35
16	Sukhatme (1954)	Wheat acreage	No. of villages	20
17	Sukhatme (1954)	Wheat acreage	No. of villages	20
18	Sukhatme (1954)	Wheat acreage	No. of villages	9

The following chart is the scatter plot for Population 9 (Cochran (1963)).

Population 9 (Cochran (1963))



Estimation of Model Parameters

Two approaches for the estimation of model parameters (α , β , σ^2 , γ) are considered:

- Maximum Likelihood (ML) Estimation
- Restricted Maximum Likelihood (REML) Estimation

For the method of ML, as discussed by Godfrey, Roshwalb and Wright (1984) and Särndal and Wright (1984), the Harvey (1976) algorithm can be used. His algorithm uses the ordinary least squares (OLS) estimates as the starting values of α and β and in each iteration the values of α and β depend on σ^2 and γ , or the reverse.

The REML estimation was developed by Patterson and Thompson (1971), and Harville (1977). The values of α and β only depend on σ^2 and γ.

Model-Based πPS Sampling vs Conventional πPS Sampling

As presented by Rao and Bayless (1969), Murthy’s method is one of the most efficient πPS methods.

With respect to $n = 2$, we compare the four model-based sampling methods and Murthy’s method.

We used “LP procedure” in the SAS software to solve the optimization problems.

Comparison of Efficiency between Model-based Sampling and Murthy's Method

Pop.	Size	1st best		2nd best		3rd best	
		Method (c)	$V < V_{Murthy}$	Method (c)	$V < V_{Murthy}$	Method (c)	$V < V_{Murthy}$
1	20	OP4 (0.2)	O	OP1 (0.2)	O	OP1 (0.1) OP1 (0.4)	O
2	20	OP2 (0.4)	O	OP4 (0.3) OP4 (0.5)	O	OP2 (0.2) OP2 (0.5)	O
3	14	OP1 (0.2)	O	OP3 (0.2)	O	OP3 (0.3)	O
4	10	OP3 (0.1)	O	OP3 (0.2)	O	OP3 (0.3)	O
5	10	OP1 (0.1)	X	OP1 (0.2)	X	OP3 (0.2)	X
6	20	OP4 (0.1)	O	OP2 (0.1)	O	OP3 (0.1)	O
7	16	OP2 (0.1)	O	OP1 (0.1)	O	OP4 (0.1)	O
8	17	OP1 (0.3)	O	OP1 (0.2)	O	OP1 (0.1)	O

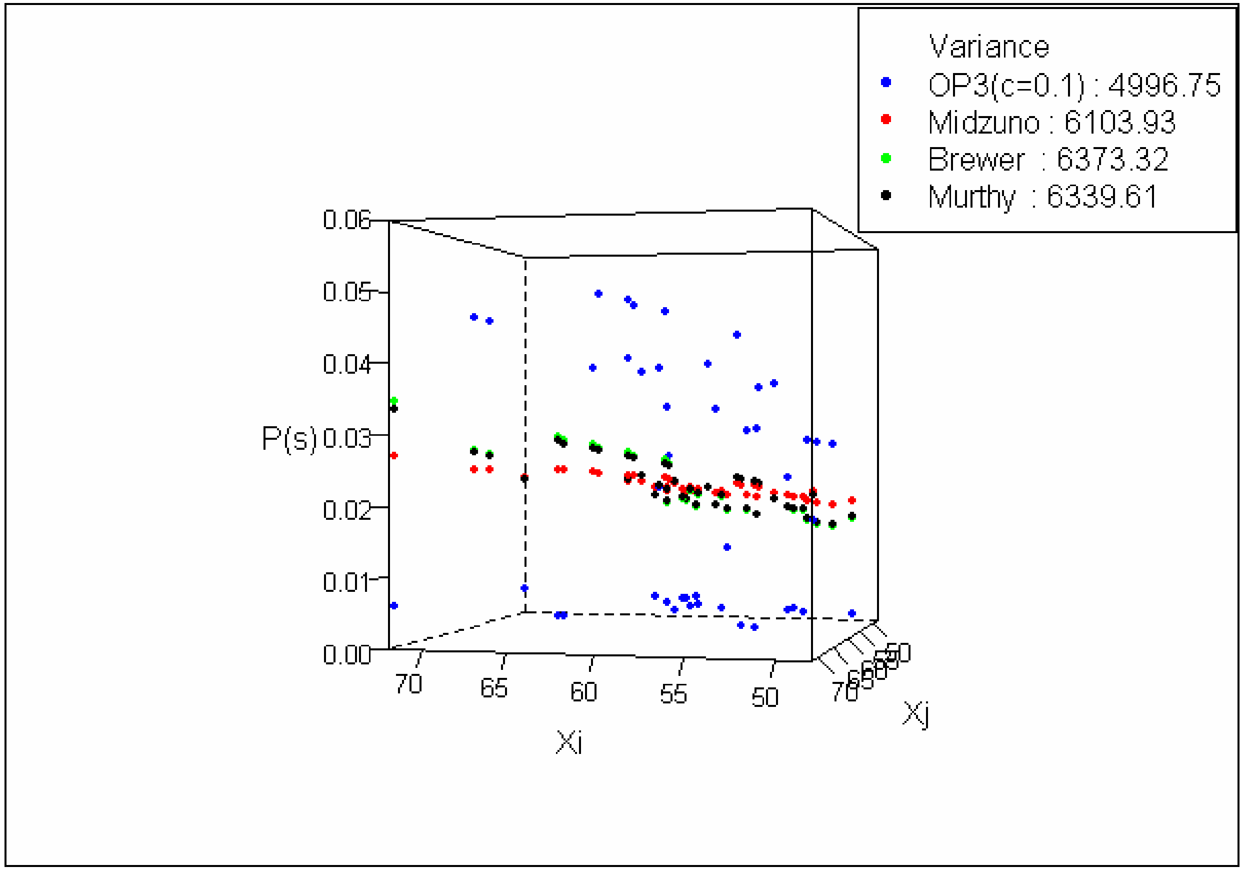
9	10	OP1 (0.1) OP3 (0.1)	O	OP1 (0.3) OP3 (0.3)	O	OP3 (0.2)	O
10	16	OP2 (0.4)	X	OP4 (0.4)	X	OP2 (0.1) OP4 (0.3)	X
11	16	OP1 (0.3)	X	OP2 (0.1)	X	OP2 (0.2)	X
12	17	OP3 (0.3)	O	OP3 (0.1)	X	OP1 (0.4)	X
13	10	OP1 (0.1)	O	OP1 (0.3)	O	OP3 (0.4)	O
14	10	OP1 (0.1)	O	OP3 (0.1)	O	OP3 (0.3)	O
15	35	OP4 (0.1)	O	OP3 (0.2)	O	OP2 (0.1) OP2 (0.2)	O
16	20	OP1 (0.1)	O	OP1 (0.2)	O	OP3 (0.2)	O
17	20	OP4 (0.4)	O	OP2 (0.4)	O	OP2 (0.2)	O
18	9	OP2 (0.1)	O	OP4 (0.1)	O	OP1 (0.1)	O

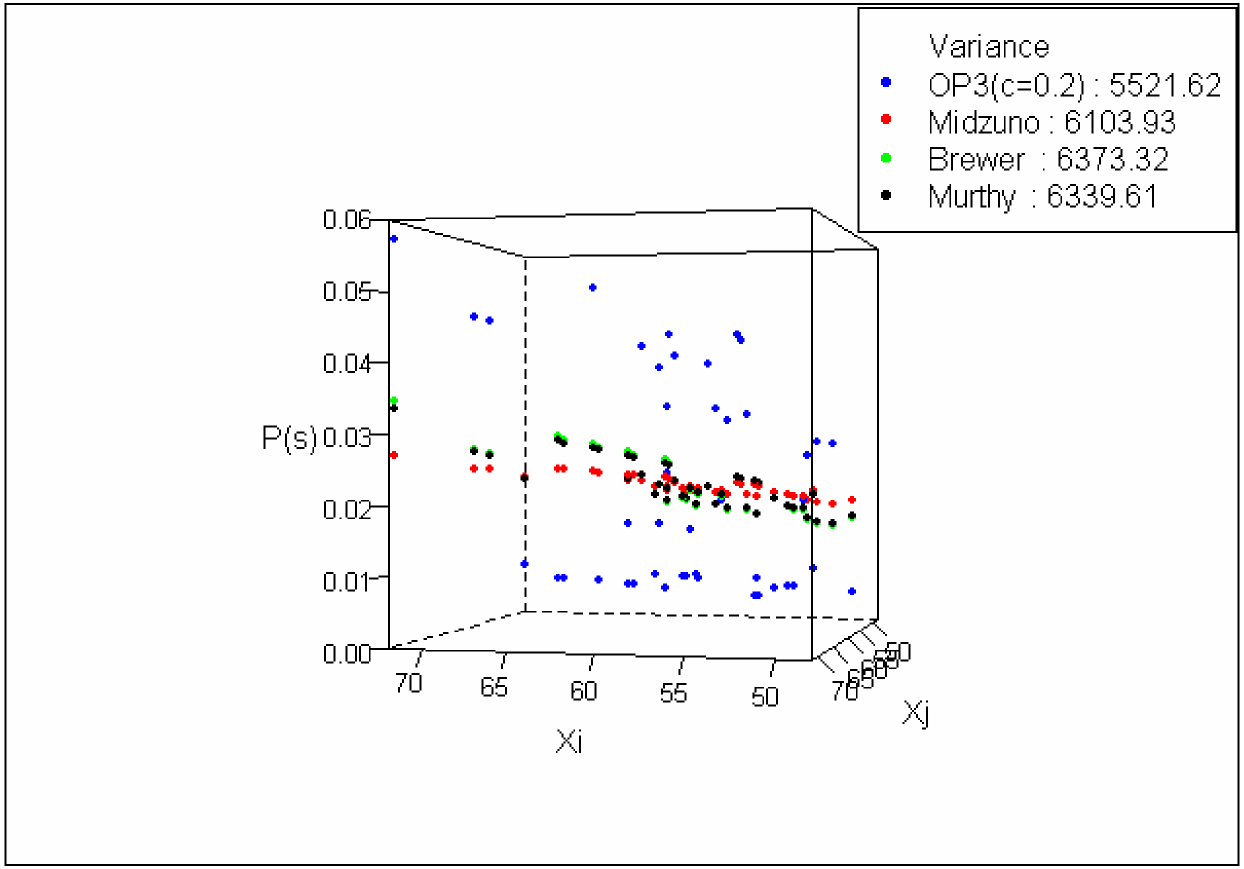
Note. For the population 1, for example, “1st best” means that the method using OP4 when $c = 0.2$ has the smallest variance among those using OP1, OP2, OP3, and OP4.

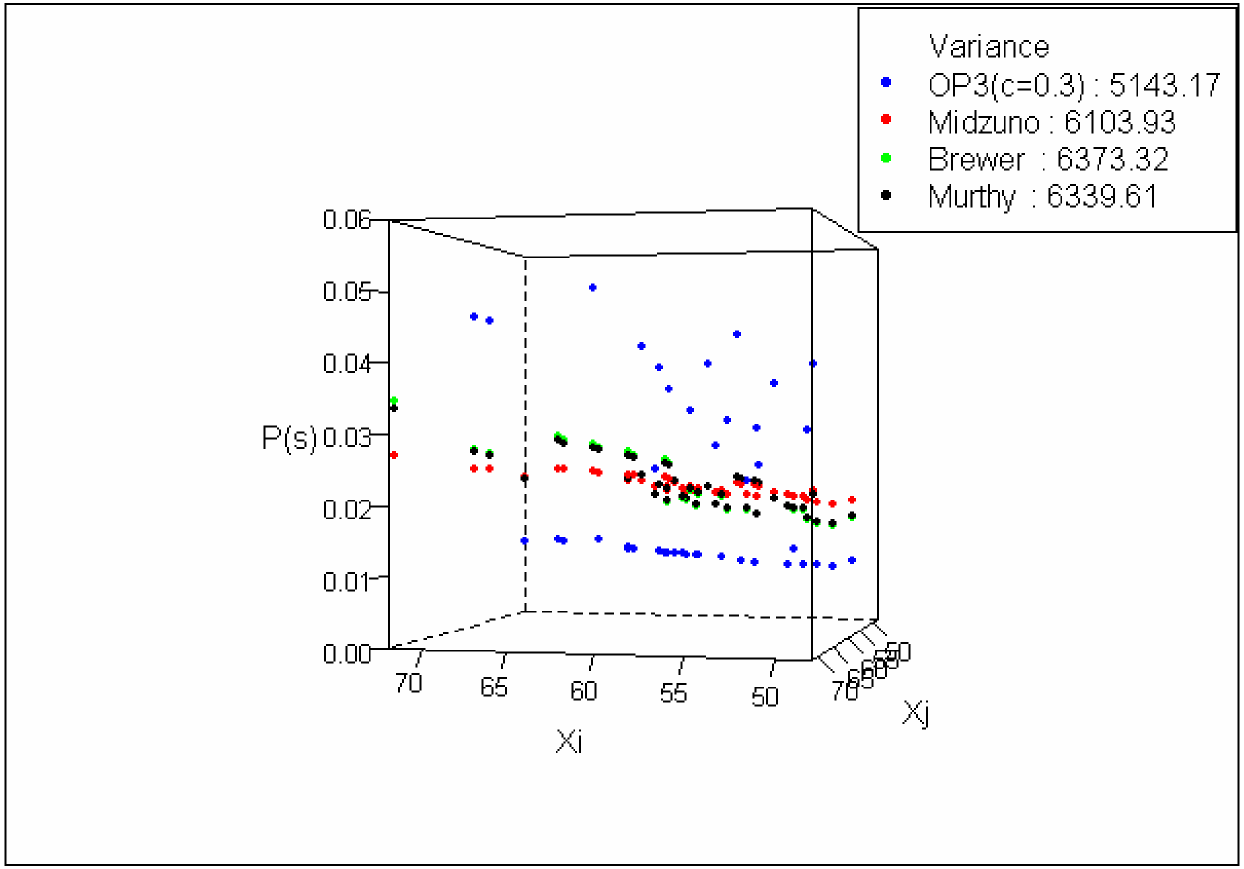
▶ Model-based sampling methods are preferable to Murthy’s method with respect to the efficiency.

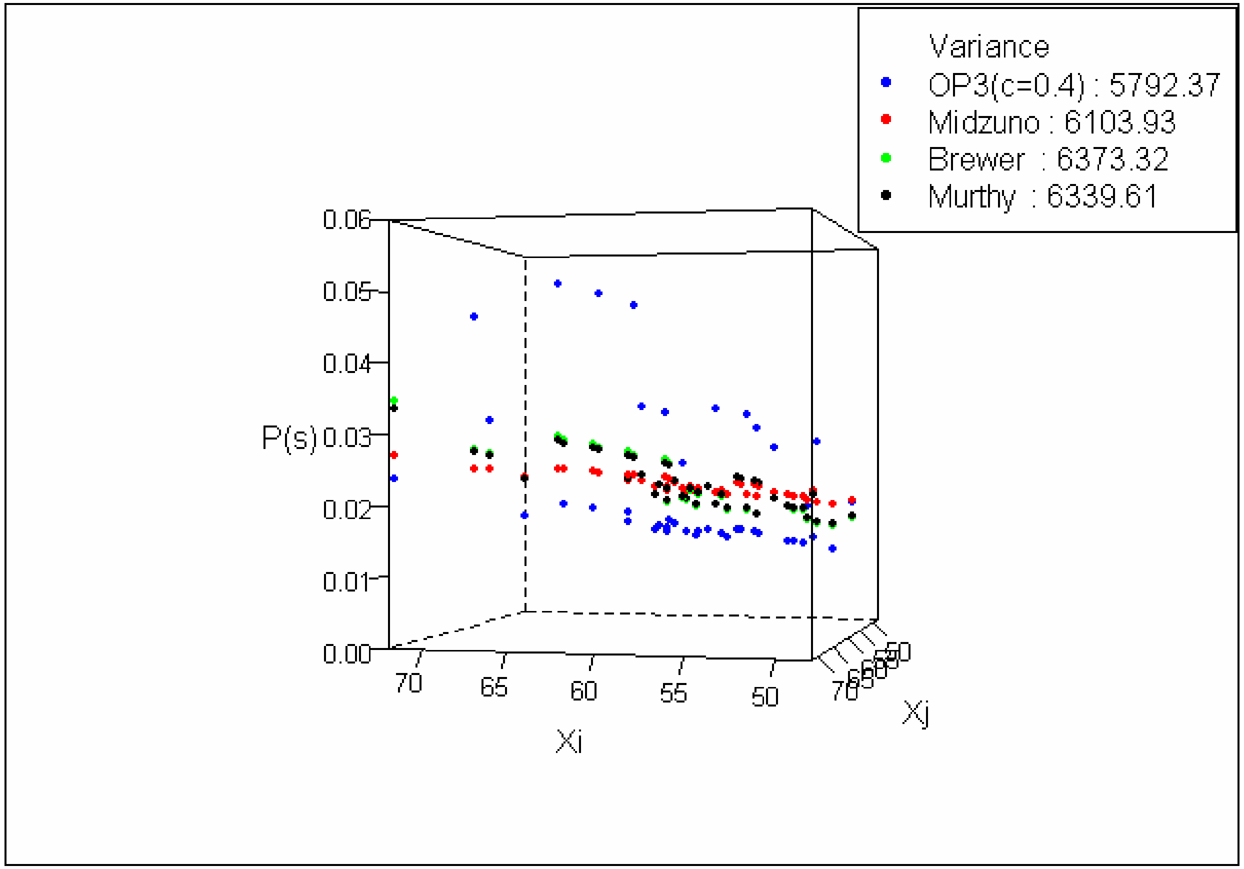
Comparison of Distributions of Sampling Designs

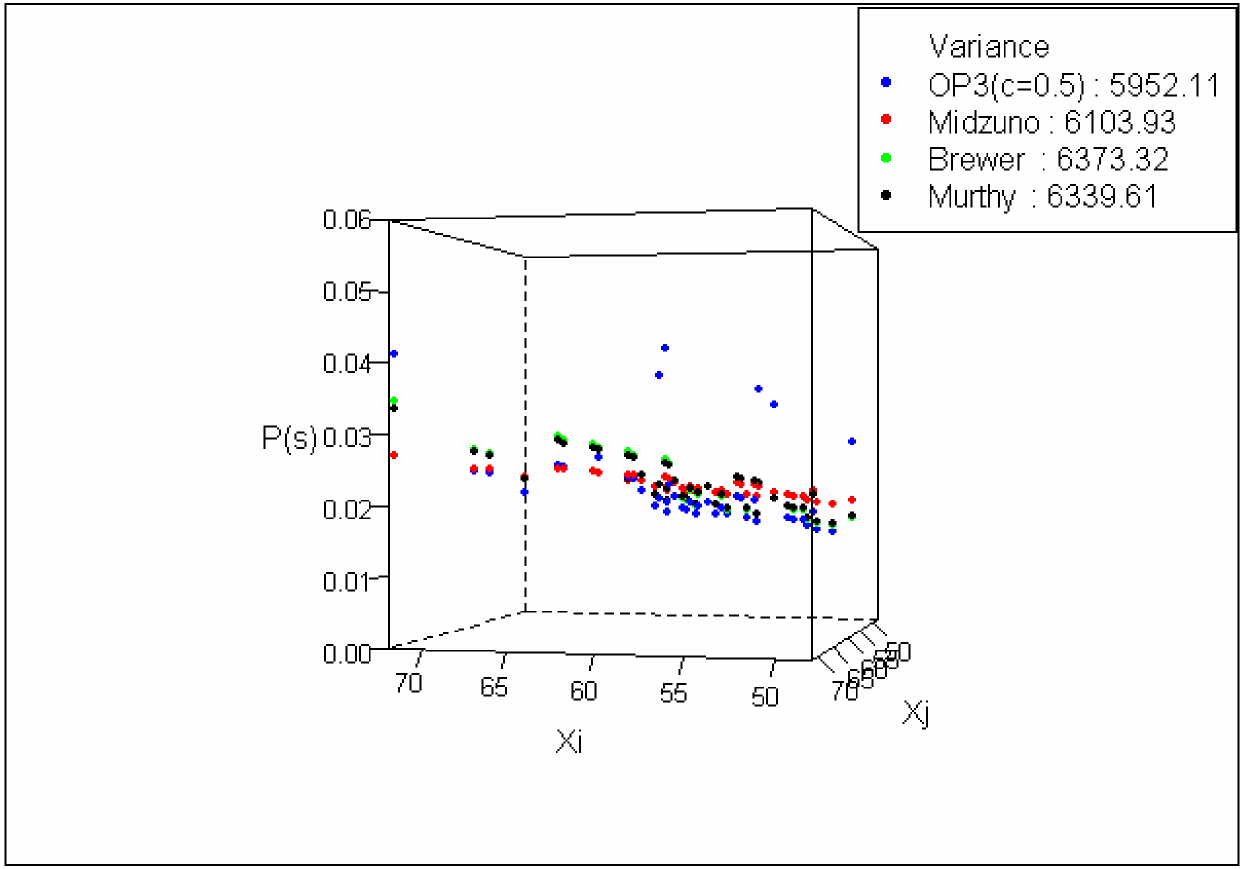
Population 9 (Cochran (1963))

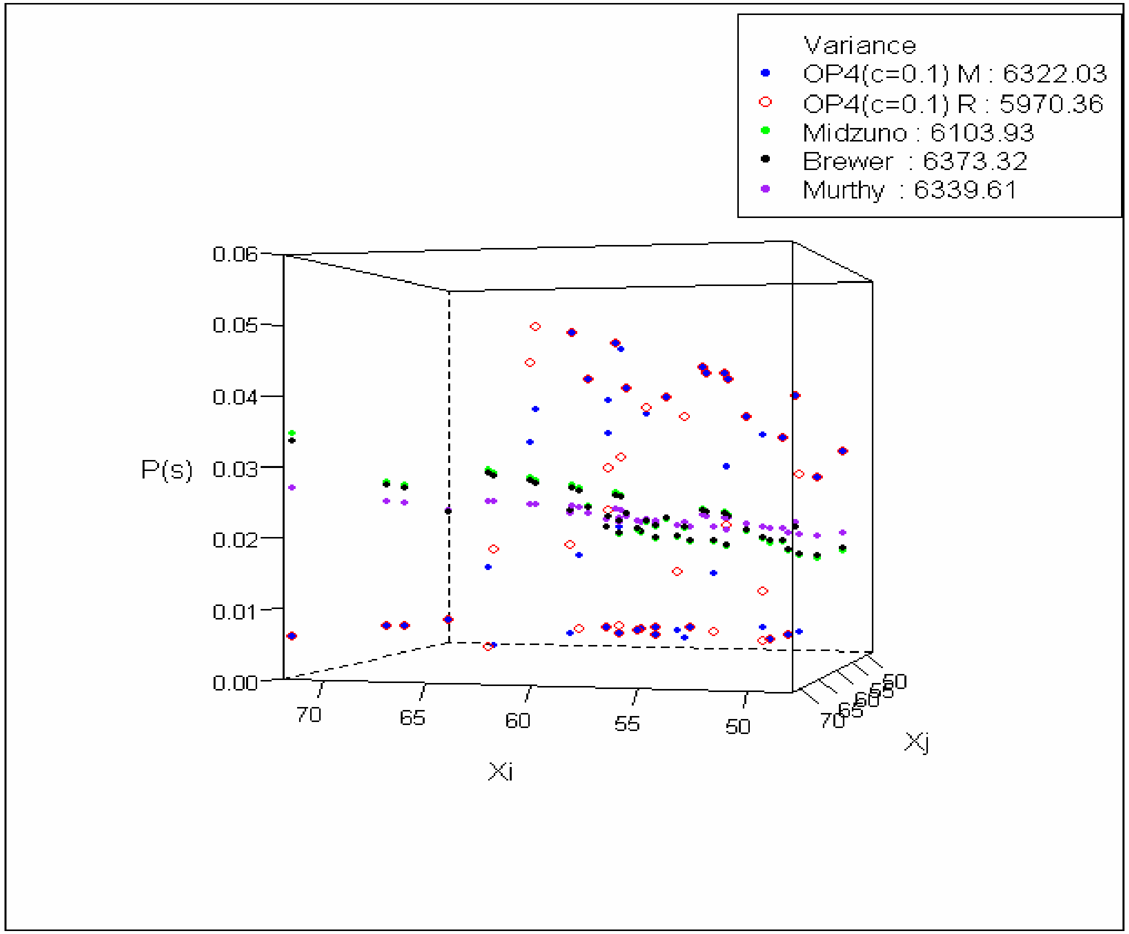



Variance
● OP3(c=0.2) : 5521.62
● Midzuno : 6103.93
● Brewer : 6373.32
● Murthy : 6339.61
P(s)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
70
65
60
55
50
Xi
70
65
60
55
50
Xj

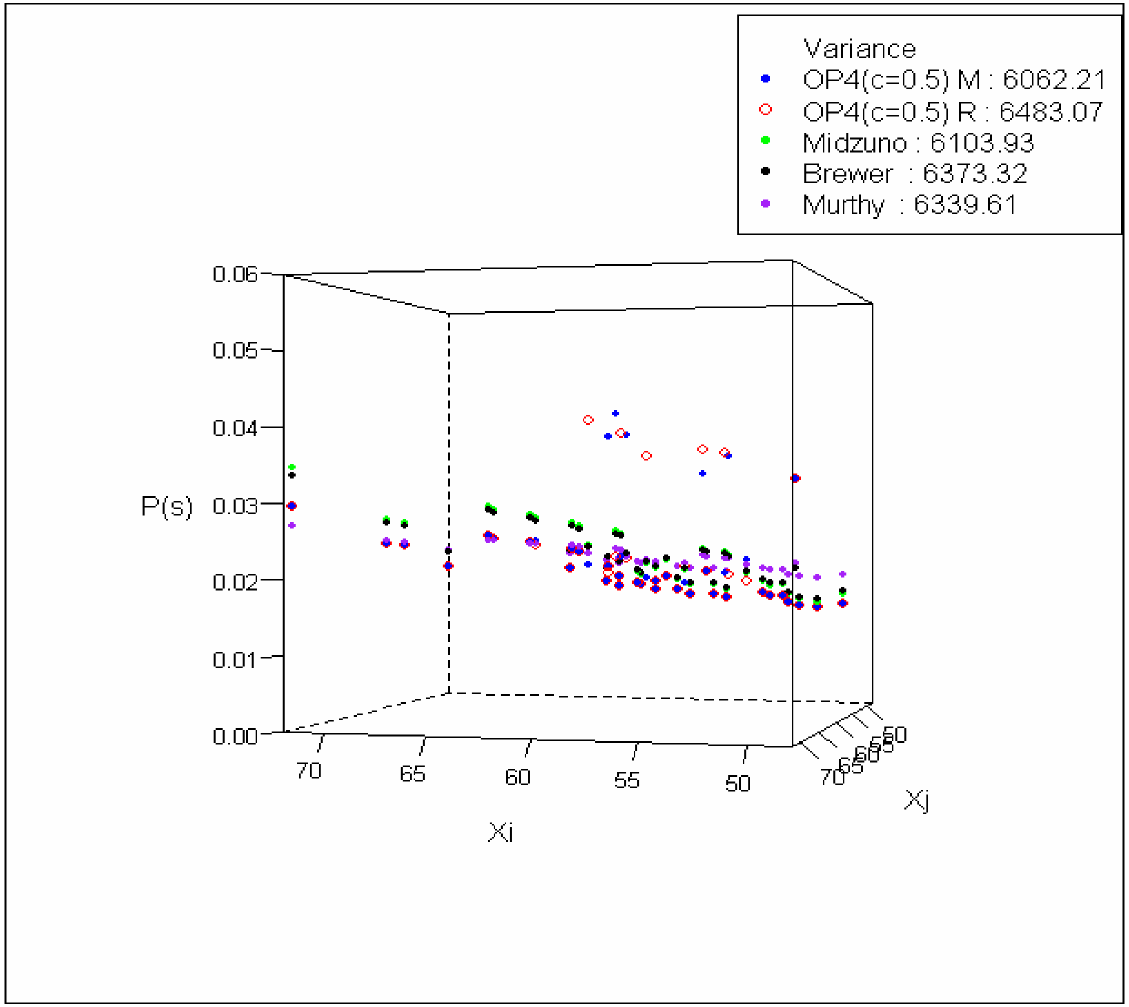

Variance
● OP3(c=0.3) : 5143.17
● Midzuno : 6103.93
● Brewer : 6373.32
● Murthy : 6339.61
P(s)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
70
65
60
55
50
Xi
70
65
60
55
50
Xj


Variance
● OP3(c=0.4) : 5792.37
● Midzuno : 6103.93
● Brewer : 6373.32
● Murthy : 6339.61
P(s)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
70
65
60
55
50
Xi
70
65
60
55
50
Xj


Variance
● OP3(c=0.5) : 5952.11
● Midzuno : 6103.93
● Brewer : 6373.32
● Murthy : 6339.61
P(s)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
70
65
60
55
50
Xi
70
65
60
55
50
Xj



Note. The estimates of α and β from the ML and REML yield the different sampling design, resulting in the different variances.


Variance
OP4(c=0.5) M : 6062.21
OP4(c=0.5) R : 6483.07
Midzuno : 6103.93
Brewer : 6373.32
Murthy : 6339.61
P(s)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
70
65
60
55
50
Xi
50
55
60
65
70
Xj

- With respect to the sampling designs, model-based sampling methods are very flexible according to the changes of value of *c* considered for the purpose of increasing the stability of the estimated variances.

An Illustration: Comparison of Efficiency

Population	Estimation	*C*	Relative Efficiency			
			OP1	OP2	OP3	OP4
9	ML	0.1	142.4	98.2	142.4	112.6
	REML		142.4	98.2		119.2
	ML	0.2	122.7	107.5	128.9	105.9
	REML		122.7	107.5		110.7
	ML	0.3	138.4	113.9	138.4	116.5
	REML		138.4	113.9		109.7
	ML	0.4	122.9	119.2	122.9	126.2
	REML		122.9	119.2		118.7
	ML	0.5	119.7	117.5	119.6	117.4
	REML		119.7	117.5		109.8

Midzuno	Brewer	Murthy	PPS_with
116.6	111.7	112.2	100.0

- In many populations, OP3 is consistently more efficient than OP1, OP2 and OP4.

- Compared to the others, OP3 leads to the higher c values, which may yield the gains in stability of the variance estimators.

Concluding Remarks

✓ We have presented πPS sampling strategy consisting of model-based sampling designs and the H-T estimator.

✓ Empirical studies paid attention to the sample size of $n = 2$, which is one of the most important cases for the practical uses.

✓ Model-based sampling methods are flexible in terms of the sampling designs, and are preferable to the conventional πPS sampling such as the methods of Murthy and Brewer.

✓ Of model-based sampling methods, the method using OP3, which has the simplest form among optimization problems, seems to perform best. Note that OP3 does not depend α and β in the superpopulation model.

- ✓ We should be careful in choosing the estimation method of model parameters.
- ✓ We should note that there might be an infeasible problem when solving optimization problems.
- ✓ The comparison of the efficiencies of the H-T estimator in the model-based sampling and the GREG estimator in the conventional sampling method may be one of the interesting issues.
- ✓ A study on the efficiency of model-based sampling methods in the larger sample size will be continued.
- ✓ The polynomial models might be adopted to increase the efficiency of model-based sampling.

Contact Information

Sung-Joon Hong
hsj8129@dongguk.edu

So-Hyung Park
12astro@naver.com

Sun-Woong Kim
sunwk@dongguk.edu

Hong-Yup Ahn
ahn@dongguk.edu

Steven G. Heeringa
sheering@isr.umich.edu