Padilla, Alberto

Working Paper

An unbiased estimator of the variance of simple random sampling using mixed random-systematic sampling

Working Papers, No. 2009-13

Provided in Cooperation with:
Bank of Mexico, Mexico City

Suggested Citation: Padilla, Alberto (2009) : An unbiased estimator of the variance of simple random sampling using mixed random-systematic sampling, Working Papers, No. 2009-13, Banco de México, Ciudad de México

This Version is available at:
https://hdl.handle.net/10419/83776

Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2009-13

An Unbiased Estimator of the Variance of Simple Random Sampling Using Mixed Random-Systematic Sampling

Alberto Padilla
Banco de México

November, 2009

Documento de Investigación
2009-13

Working Paper
2009-13

An Unbiased Estimator of the Variance of Simple Random Sampling Using Mixed Random-Systematic Sampling [*]

Alberto Padilla[†]
Banco de México

Abstract
Systematic sampling is a commonly used technique due to its simplicity and ease of implementation. The drawback of this simplicity is that it is not possible to estimate the design variance without bias. There are several ways to circumvent this problem. One method is to suppose that the variable of interest has a random order in the population, so the sample variance of simple random sampling without replacement is used. By means of a mixed random - systematic sample, an unbiased estimator of the population variance for simple random sample is proposed without model assumptions. Some examples are given.
Keywords: Variance estimator; Systematic sampling; Simple random sampling; Random order.
JEL Classification: C80, C83.

Resumen
El muestreo sistemático es un método ampliamente usado en la práctica debido a su sencillez. Empero, tal sencillez tiene un costo, no es posible estimar insesgadamente la varianza de dicho diseño muestral. Hay varias formas de tratar este problema. Una de ellas consiste en suponer que la variable de interés tiene un orden aleatorio en la población, con lo cual puede emplearse el estimador de la varianza bajo muestreo aleatorio simple. En el presente trabajo se propone un estimador insesgado para la varianza poblacional del muestreo aleatorio simple sin suponer modelo alguno, empleando un muestreo mixto aleatorio-sistemático. Se ilustra el método con algunos ejemplos.
Palabras Clave: Estimador de varianza; Muestreo sistemático; Muestreo aleatorio simple; Orden aleatorio.

[*]The author would like to thank Ignacio Méndez from IIMAS-UNAM, research seminar participants at Banco de México and two reviewers from Banco de México for their useful comments and suggestions.
[†] Dirección General de Emisión. Email: ampadilla@banxico.org.mx

1. Introduction

Systematic sampling is a commonly used technique due to its simplicity and operational convenience. The main disadvantage is the non-existence of a design unbiased variance estimate of the sample mean with a single systematic sample. Several approaches have been proposed to overcome this difficulty. One of them treats the systematic sample as if it were drawn from a population in random order, so the formula of the variance estimator of the mean under simple random sampling without replacement, hereinafter *srswor*, applies, Cochran (1986). In another approach, a model is used for the variable of interest and, consequently, a specific formula for the estimator of the model variance has to be obtained. From the design perspective of a survey, one can also apply a random permutation to the elements of the population before the sample is drawn. With this method the variance estimator $\hat{v}_{srswor}(\hat{\bar{y}})$ is used, although this procedure is not feasible in many surveys. Another class of methods supplements the systematic sample with another systematic sample or a simple random sample. For a thorough discussion of these strategies see Wolter (1985) or Chaudhuri & Stenger (2005). In one of these methods a simple random sample is selected first, and in the remaining population a systematic sample is extracted, Leu & Tsui (1996) and Huang (2004). Other systematic sampling methods, called 'Markov sampling', have been proposed, see Sampath & Uthayakumaran (1998) and the references cited therein. Unfortunately, these methods cannot be applied to a population containing a large number of elements and the population size has to be a multiple of the sample size. In Sampath & Uthayakumaran (1998), for example, the sample size must be even. These are very stringent conditions in large surveys and have not been used extensively in applied work. All the methods above mentioned and its merits have been examined in detail in the literature and shall not be reviewed here.

A mixed random-systematic sampling method is proposed in which the population mean and variance of the mean, under *srswor*, are unbiasedly estimated by the sample

mean and a simple expression for the variance[1]. This last expression can be used without assuming that the sample was drawn from a population in random order or a random permutation has been applied to the population before the sample was extracted, preventing people to fall in *PISE*, an acronym coined by Valliant (2007), which stands for 'pretend it's something else'. It is worth mentioning that, compared to systematic sampling and similar methods, no gain in efficiency is expected with the proposed method, since it coincides with the population mean and variance of a *srswor*. A fair comparison of the proposed method is with the estimator of the variance between elements used under the random order approach in systematic sampling.

The article is organized as follows. Definitions, notation and a brief overview of finite population sampling are given in Section 2. Standard practices regarding the estimation of the design variance under systematic sampling are reviewed in Section 3. In this section, expressions for the bias and relative bias of the estimator of the variance between elements of the random order approach are given. To the author´s best knowledge, these expressions have not appeared previously in the literature. Section 4 contains the sampling procedure and an example. The estimators for the population mean and variance $v_{srswor}(\hat{\bar{y}})$ are presented in Section 5. Finally, the method is illustrated with numerical examples.

2. Finite population sampling

There are two types of surveys, descriptive or analytical. The former refers to the estimation of quantities such as totals, means, proportions and ratios, while the latter to the use of models based on the results of a survey. The formulas developed in this paper are of the descriptive type.

[1] This is an extended version of an article presented by the author in Puebla, Padilla (2009).

In this article it is assumed that all variability stems from sampling error, so any errors caused by faulty measurement, non-response and other nonsampling sources are ignored. It is also supposed that the design is noninformative. An informative design is one in which the probability of selection of the elements in the sample depends explicitly on the values of the study variables. As a matter of fact, the latter is an assumption made in almost all practical survey work not usually mentioned in books or articles.

It is also assumed that a frame exists from which a sample will be drawn.

2.1 Notation, population and sample

Let U denote a finite population of N elements labeled $k=1,\ldots,N$, $1<N$. It is customary to represent the finite population by its label k as: $U=\{1,2,\ldots,k,\ldots,N\}$. Moreover, there is a one to one correspondence between the labels of U and the labels of the frame.

The variable under study will be represented by y and y_k will be the value of y for the *kth* population element, $k \in U$.

The sample will be denoted by $\boldsymbol{s}$, a subset of U of size $1<n<N$, and will be represented by a column vector $I = (I_1,\ldots, I_k,\ldots, I_N)' \in \{0,1\}^N$. In this case, I_k is an indicator random variable and it is equal to *1* if the *kth* element is in the sample and *0* otherwise. It is worth mentioning that this indicator variable is the random element in finite population sampling and y_k is a number. So, the density function induced by the design is discrete. This approach is also known in the literature as design-based sampling.

2.2 Estimation

The objective is to estimate a function t that depends on the y_k, $t = t(y_1,...,y_k,...,y_N)$. For example, a total is written as $y_U = \sum_{k=1}^{N} y_k$. Since we are interested in estimating a total, from the design-based approach, it is customary to use the Horvitz-Thompson estimator, *HTE*, Horvitz & Thompson (1952). This estimator has the following expression: $\hat{y}_U = \sum_{k=1}^{N} I_k y_k \big/ \pi_k = \sum_{k=1}^{n} y_k \big/ \pi_k$, with $\pi_k > 0$. In this formula, $\pi_k = P(I_k = 1)$ is the first-order inclusion probability. For variance computation and estimation it is also necessary to determine the second-order inclusion probabilities, $\pi_{kl} = P(I_k I_l = 1)$.

The variance of a HTE is,

$$v(\hat{y}_U) = \sum\sum_U c(I_k, I_l)\hat{y}_k\hat{y}_l = \sum\sum_U (\pi_{kl} - \pi_k\pi_l)\hat{y}_k\hat{y}_l .$$

An unbiased estimator of this variance is, provided that $\pi_{kl} > 0$:

$$\hat{v}(\hat{y}_U) = \sum\sum_s \hat{c}(I_k, I_l)\hat{y}_k\hat{y}_l = \sum\sum_s \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

In these expressions, $c(I_k, I_l)$ and $\hat{c}(I_k, I_l)$ denote the population and estimated covariances respectively, between the sample indicator variables.

Remark 2.2.1: It is worth mentioning that in finite population sampling, the first two moments are well defined for designs used in practice, so there is no need to include this fact in the rest of the article.

Remark 2.2.2: Estimation in finite populations can also be made under a different approach known in the literature as model-based design in which it is supposed that the finite population is drawn from an infinite population (superpopulation), see Valliant et

al. (2000). The design and model based methods can be used together in what is denominated combined sampling, see Brewer (2002).

3. Standard practices in systematic sampling

As it was mentioned in the introduction, there is no design unbiased variance estimates of the variance of the sample mean with a single systematic sample, so in practice the following strategies, among others, are used.

3.1 During the design stage of a survey

D1) Supplement the systematic sample with another sample.

D2) Apply a random permutation to the elements of the population before the sample is extracted, so under all possible permutations of the population, the expectation of the design variance is the same as the variance under *srswor*. This result is due to Madow & Madow (1944).

Remark 3.1.1: A comparison of the efficiency of some designs of the *D1* type, can be found in Zinger (1980), Cochran (1986) and Wolter (1985).

3.2 Model for the structure of the variable of interest

Postulate a model for the structure of the variable under study before extracting the systematic sample and construct the variance estimator under this model. In this case, two models are routinely employed:

M_{sc}) Serial correlation: in some settings, there is evidence of similarities between neighboring elements in the population with respect to the variable of interest and this similarity diminishes as two elements are far apart from each other.

M_{ro}) Random order model in an infinite population: the finite population is considered as a random sample from an infinite population (superpopulation). If the variates $y_i, i = 1,..., N$, are drawn from a superpopulation in which $E_M(y_i) = \mu$, $E_M(y_i - \mu)^2 = \sigma_i^2$ and $E_M(y_i - \mu)(y_j - \mu) = 0, i \neq j$, it is known as a population in random order. In these expressions, E_M refers to expectation under the assumed model. The result of this is, see Cochran (1986), that $E_M(\hat{v}_{sys}(\hat{\bar{y}})) = E_M(\hat{v}_{srswor}(\hat{\bar{y}}))$, where *sys* refers to systematic sampling. Under this model, it is assumed that there is no relationship between the variable under study and the order of the elements in the frame, so one treats a systematic sample from a list, sorted in a specific order, as if the list were randomly ordered.

Remark 3.2.1: A comparison of the efficiency of models M_{sc} and M_{ro}, can be found in Wolter (1985) and Chaudhuri & Stenger (2005).

3.3 Bias of the random order approach (M_{ro})

Under the M_{ro} approach, the estimator of the variance of the mean under simple random sampling, $\hat{v}_{srswor}(\hat{\bar{y}}) = (1 - n/N)\hat{s}_{sys}^2 \big/ (n-1)$, is used. In this expression, $\hat{s}_{sys}^2$ stands for the variance between elements of the systematic sample. This is a reasonable strategy whenever there is information about the random order of the elements in the population. The problem is that it is easy to fall in PISE and work with a biased estimator of the variance or to routinely apply the simple random estimator without having enough information about the ordering of the elements in the population. To assess this approach, in the following theorem the bias and relative bias of the variance estimator are obtained. Suppose that $k = N/n$ is an integer and $\hat{s}_{sys,i}^2 = \sum_{j=1}^{n}(y_{ij} - \hat{\bar{y}}_i)^2 \big/ (n-1)$, $\hat{\bar{y}}_i = \sum_{j=1}^{n} y_{ij} \big/ n$, $S_U^2 = \sum_{j=1}^{N}(y_j - \bar{y}_U)^2 \big/ (N-1)$, $\bar{y}_U = \sum_{j=1}^{N} y_j \big/ N$ and $\rho = 2\sum_{l=1}^{k}\sum_{i=1}^{n-1}\sum_{j>i}^{n}(y_{li} - \bar{y}_U)(y_{lj} - \bar{y}_U) \Big/ (n-1)(N-1)S_U^2$, where ρ is the intraclass correlation coefficient, Cochran (1986).

Theorem 1: Under systematic sampling the expected value of the estimator $\hat{s}^2_{sys,i}$ is $\frac{N-1}{N}(1-\rho)S^2_U$.

Corollary 1.1: The relative bias of the estimator $\hat{s}^2_{sys,i}$ is $\frac{N-1}{N}(1-\rho)-1$.

Corollary 1.2: $E(\hat{s}^2_{sys,i})$ is a linear decreasing function of ρ, which achieves its maximum at $\rho=-1/(n-1)$, its minimum at $\rho=1$ and $E(\hat{s}^2_{sys,i})=S^2_U$ whenever $\rho=-1/(N-1)$. The maximum and minimum values of $E(\hat{s}^2_{sys,i})$ are $S^2_U\frac{n}{n-1}\frac{N-1}{N}$ and cero respectively.

Corollary 1.3: The expected value and relative bias of the estimator $\hat{s}^2_{sys,i}$ can also be expressed as $(1-\partial)S^2_U$ and $-\partial$, where ∂ is the measure of homogeneity proposed by Särndal et al. (1992).

Proof: see the Appendix.

Remark 3.3.1: It can be seen from corollary 1.2 that $E(\hat{s}^2_{sys,i})$ overestimates S^2_U for $\rho\in[\frac{-1}{n-1},\frac{-1}{N-1})$.

4. Design

4.1 Definition of mixed random-systematic sampling

Following the design based approach, we consider a population *U*, with *N* elements, $y_k, k = 1,...,N$. From this population a sample of size *n*, 1<*n*<*N*, is drawn by means of a mixed random-systematic sample, *mrss*. That is, a *srsrwor* of size 1 is first selected from the elements of *U* and then *m* elements, *m≥2*, are drawn from the *N-1* remaining elements of *U* using circular systematic sampling, Murty & Rao (1988). For brevity, this method shall be denoted by *mrss(1,m)*. The number of samples under this design is $N(N-1)$.

Remark 4.1.1: When $(N-1)/m$ is an integer, circular and linear systematic sampling coincide, Murty & Rao (1988), so the systematic sample can also be extracted by the latter method. In this case there are repeated circular systematic samples; nonetheless, the point estimators of the mean and element variance, which are built in the next section, continue to be unbiased after suppressing information.

Remark 4.1.2: The number of samples under a *mrss(1,m)* design, after eliminating repeated systematic samples, is $N(N-1)/m$ if $(N-1)/m$ is an integer and $N(N-1)$ in other case. For further details see Murthy & Rao (1988).

4.2 Circular systematic sampling

In order to obtain a circular systematic sample, *css*, of size *1*<*m*<*M* from a population with *M* elements, one proceeds as follows:

Step 1: compute $k_m = (N-1)/m$; if k_m is not an integer, round it to the nearest integer,

Step 2: select a random integer between *1* and *M*, say *r*, this is the first element in the *css*,

Step 3: determine the next numbers in the *css*, $r + jk_m$, for $j \in \{1,...,m-1\}$. If $r + jk_m > M$ consider the list as circular and assign the numbers until the sample size is achieved.

Remark 4.2.1: this procedure can be easily implemented in a spreadsheet or in the *R* system.

Example 1: let *U* be a population of size *N=7* and suppose a sample of size *n=3* is to be drawn using a *mrss(1,2)*. In this case *m=2* and there are *7(7-1)=42* samples. The indices for the possible samples are:

Table 1

1 2 5	2 1 5	3 1 5	4 1 5	5 1 4	6 1 4	7 1 4
1 3 6	2 3 6	3 2 6	4 2 6	5 2 6	6 2 5	7 2 5
1 4 7	2 4 7	3 4 7	4 3 7	5 3 7	6 3 7	7 3 6
1 5 2	2 5 1	3 5 1	4 5 1	5 4 1	6 4 1	7 4 1
1 6 3	2 6 3	3 6 2	4 6 2	5 6 2	6 5 2	7 5 2
1 7 4	2 7 4	3 7 4	4 7 3	5 7 3	6 7 3	7 6 3

The first number in each entry refers to the *srswor* selection and the following two correspond to the systematic sample.

5. Point estimators

As it was noted by Huang (2004), in mixed random systematic sampling the *HTE* $\hat{\bar{y}} = 1/N \sum_{k=1}^{n} y_k/\pi_k$, $\pi_k > 0$ can be used to estimate the population mean, provided that N is known. To compute this estimator, we only need to determine the first-order inclusion probabilities.

Theorem 2: Under *mrss(1,m)*, the first-order inclusion probabilities, π_k, are equal to n/N, for all $k = 1,..., N$.

Proof: see the Appendix.

Corollary: For an *mrss(1,m)* design, the *HTE* is the usual sample mean.

Proof: it follows immediately by substituting $\pi_k = n/N$ in the expression of the *HTE* of the mean.

Remark 5.1: The *mrss(1,m)* estimator of the mean can also be written as a weighted sum, $\hat{\bar{y}}_{r,s} = \beta y_r + \alpha \hat{\bar{y}}_s$, with $\beta = 1/n$, $\alpha = m/n$. The first term of the sum refers to the value of y obtained by *srswor*, while the second one is the sample mean of the systematic sample. This is also known as a Zinger estimator, Ruiz-Espejo (1997).

Remark 5.2: The *mrss(1,m)* estimator of the mean is unbiased because it is a *HTE*.

The most important result of this article is expressed in the next theorem.

Theorem 3: Under *mrss(1,m)*, an unbiased estimator of the population variance between elements, $s_U^2 = \sum_{k=1}^{N} (y_k - \bar{y}_U)^2 / (N-1)$, is:

$$\hat{s}_{r,s}^2 = \frac{\sum_{k=1}^{m} (y_r - y_{s,k})^2}{2m},$$

where y_r is the value of the variable selected by *srswor*, $y_{s,k}$ are the values of the elements selected by the circular systematic sample and $\bar{y}_U$ is the population mean.

Proof: see the Appendix.

Corollary: Under *mrss(1,m)*, an unbiased estimator or the variance of the mean of *srswor*, $v_{srswot}(\hat{\bar{y}})$, is given by the following expression,

$$\hat{v}_{srswor}(\hat{\bar{y}}_{r,s}) = \frac{(1-n/N)}{n} \hat{s}^2_{r,s}.$$

Proof: immediate from the property of expectations, $E(cX) = cE(X)$, where $c = (1-n/N)/n$.

Remark 5.3: There is no assumption about random order in the population and there was no need for applying a permutation before the sample was drawn. To put this briefly, the *mrss(1,m)* design provides a simple expression for the variance estimation without pretending it is something else, Valliant (2007).

Remark 5.4: In the expression $\hat{v}_{srswor}(\hat{\bar{y}}_{r,s})$ one can use a sample size *m* to estimate it.

Remark 5.5: Zinger (1980) proposed an unbiased estimator of the variance between elements using partially systematic sampling in which one first selects a systematic sample and then a *srswor* from the remaining population. Unfortunately, the formula proposed by Zinger is quite complex.

6. Numerical example

Example 2: let *U* be the population of example 3.4.2, pages 80-82, Särndal et al. (1992). This population has *N=100* elements and the variable *y* takes the values 1, 2,…,100. Using systematic sampling with *n=10* there are N/n=10 samples and the population mean $\bar{y}_U$ and variance between elements S_U^2 are 50.5 and 841.67 respectively. As simple random sampling does not take into account the ordering of the population, the variance of the mean estimator under this design is $v_{srswor}(\hat{\bar{y}}) = (1-n/N)\,S_U^2/n = 75.75$. In Tables 2 to 5 there are four orderings of the same population which have different values of the intraclass correlation coefficient. For each ordering and for all samples under systematic sampling, we present the values of the sample mean, $\hat{\bar{y}}_{sys}$, the estimator of the variance between elements, $\hat{s}_{sys}^2$, and the estimator of the variance of the sample mean under the random order assumption, $\hat{v}_{ro}(\hat{\bar{y}}_{sys})$. Under the random order assumption, the estimators for every systematic sample $\hat{s}_{sys}^2$ and $\hat{v}_{ro}$ were computed using the following expressions: $\hat{s}_{sys,i}^2 = \sum_{k=1}^{10}(y_k - \hat{\bar{y}}_{sys,i})^2/(10-1)^2$ and $\hat{v}_{ro}(\hat{\bar{y}}_{sys,j}) = (1-10/100)\,\hat{s}_{sys,j}^2/10$. The labels s-1, s-2,…,s-10 correspond to the results of sample 1 to sample 10. The last column has the expected values of the sample means and variances, $E(\hat{\bar{y}}_{sys})$, $E(\hat{s}_{sys}^2)$ and $E(\hat{v}_{ro})$ respectively.

Table 2

Population A: perfect linear trend in the values y_k, roh= -0.10.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{\bar{y}}_{sys}$	46.0	47.0	48.0	49.0	50.0	51.0	52.0	53.0	54.0	55.0	**50.5**
$\hat{s}_{sys}^2$	916.7	916.7	916.7	916.7	916.7	916.7	916.7	916.7	916.7	916.7	**916.7**
$\hat{v}_{ro}$	82.5	82.5	82.5	82.5	82.5	82.5	82.5	82.5	82.5	82.5	**82.5**

Table 3

Population B: a minimal variance ordering for systematic sampling, roh= -0.11.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{\bar{y}}_{sys}$	50.5	50.5	50.5	50.5	50.0	50.5	50.5	50.5	50.50	50.50	**50.5**
$\hat{s}^2_{sys}$	989.2	969.2	951.4	935.8	922.5	911.4	902.5	895.8	891.4	889.2	**925.8**
$\hat{v}_{ro}$	89.0	87.2	85.6	84.2	83.0	82.0	81.2	80.6	80.2	80.0	**83.3**

Table 4

Population C: a large positive roh value, roh= 0.989.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{\bar{y}}_{sys}$	5.5	15.5	25.5	35.5	45.5	55.5	65.5	75.5	85.5	95.5	**50.5**
$\hat{s}^2_{sys}$	9.2	9.2	9.2	9.2	9.2	9.2	9.2	9.2	9.2	9.2	**9.2**
$\hat{v}_{ro}$	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	**0.83**

Table 5

Population D: a random ordering, roh= -0.015.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{\bar{y}}_{sys}$	44.3	34.8	40.7	61.2	48.8	59.5	47.6	58.7	58.4	51.0	**50.5**
$\hat{s}^2_{sys}$	720.9	420.0	1014.7	948.2	494.4	948.7	1222.5	522.7	780.5	1388.4	**846.1**
$\hat{v}_{ro}$	64.9	37.8	91.3	85.3	44.5	85.4	110.0	47.0	70.2	125.0	**76.1**

In order to make a comparison between the strategy of estimating the variance between elements assuming random ordering of the population in systematic sampling and mixed random-systematic sampling, for populations A to D, a *mrss(1,9)* was used.

In this case, there are 100(100-1)=9,900 possible samples under mixed random-systematic sampling. For each population, the 9,900 samples were generated and the coefficient of variation of the variance between elements, $\hat{s}^2_{r,s}$, was computed to assess the performance of the estimator of the variance.

Table 6

	A	B	C	D
Population mean $\bar{y}_U$ =	50.5	50.5	50.5	50.5
S^2_U =	841.7	841.7	841.7	841.7
$v_{srswor}(\hat{\bar{y}})$ =	75.75	75.75	75.75	75.75
Systematic sampling:				
Intraclass correlation=	-0.10	-0.11	0.989	-0.015
Random order estimator $\hat{S}^2_{sys}$ =	916.7	925.8	9.2	846.1
Relative bias ($\hat{S}^2_{sys}$)=	8.9%	10.0%	98.9%	0.5%
Variance estimator $\hat{v}_{ro}(\hat{\bar{y}}_{sys})$ =	82.5	83.3	0.83	76.1
Coefficient of variation ($\hat{\bar{y}}_{sys}$)=	6.0%	0%	60.0%	17.7%
Coefficient of variation ($\hat{s}^2_{sys}$)=	0%	3.7%	0%	37.7%
Mixed random-systematic sampling:				
Variance estimator $\hat{v}_{srswor}(\hat{\bar{y}}_{r,s})$ =	75.75	75.75	75.75	75.75
Coefficient of variation ($\hat{\bar{y}}_{r,s}$)=	7.7%	7.7%	7.7%	21.3%
Coefficient of variation ($\hat{s}^2_{r,s}$)=	46.0%	46.3%	46.6%	60.9%

In Table 6, the letters at the top of each column correspond to populations from Tables 2 to 5. Comparing the variance estimators $\hat{v}_{ro}(\hat{\bar{y}}_{sys})$, $\hat{v}_{srswor}(\hat{\bar{y}}_{r,s})$ and the coefficients of variation of the estimators of the population mean and variance between elements for both designs, we can see that the estimators under the random order assumption used in systematic sampling, behave erratically and depend heavily on the order of the population. Mixed random-systematic sampling performs well for populations A through C; nevertheless, for population D the sampling distributions of

$\hat{\bar{y}}_{r,s}$ and $\hat{s}^2_{r,s}$ have more variation than their counterpart in systematic sampling. This is due to the presence of influential observations in the distribution of the $\hat{s}^2_{r,s}$.

7. Summary

By means of a mixed random-systematic sample, an unbiased estimator of the population variance for simple random sampling without replacement has been proposed. It was shown that there is no need to suppose random ordering of the population or to apply a permutation before a systematic sample is drawn in order to use the proposed estimator of the population variance between elements. It was also shown that the bias and relative bias of the estimator of the variance between elements under systematic sampling with the assumption of random ordering of the population depend on the intraclass correlation coefficient.

Appendix

Proof of Theorem 1:

Suppose that $N = nk$, $1 < n < N$ and k and n are integers.

Note that the variation between elements in the population can be decomposed as:

$$\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}k(\bar{y}_i - \bar{y}_U)^2 .$$

This is the decomposition of the total variation into the variation within systematic samples and the variation between systematic samples, as it is done in the standard one-way analysis of variance and can be expressed as:

$$SST = SSW + SSB$$

Here, *SS* represents sums of squares; *T*, total; *W*, within and *B*, between. The proof consists in computing the expectation of the sample variance between elements of the systematic sample, $s_{sys,i}^2 = \sum_{j=1}^{n}(y_{ij} - \hat{y}_i)^2 \Big/ (n-1)$.

$$E(\hat{s}_{sys,i}^2) = \frac{\sum_{i=1}^{k}\hat{s}_{sys,i}^2}{k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2 - n\sum_{i=1}^{k}\hat{\bar{y}}_i^2}{k(n-1)} = \frac{\sum_{i=1}^{N}y_i^2 - n\sum_{i=1}^{k}\hat{\bar{y}}_i^2}{k(n-1)}$$

We add $-N\bar{y}_U + kn\bar{y}_U$ in the last expression and noting that $k(n-1) = N-k$,

$$E(\hat{s}_{sys,i}^2) = \frac{N-1}{N-k}S_U^2 - \frac{1}{N-k}\sum_{i=1}^{k}k(\hat{\bar{y}}_i - \bar{y}_U)^2 = \frac{N-1}{N-k}S_U^2 - \frac{1}{N-k}(SST - SSW)$$

Recalling that $SST = (N-1)S_U^2$, we have $E(\hat{s}_{sys,i}^2) = \frac{SSW}{N-k}$.

This is the intra-sample variance proposed by Särndal et al. (1992, p. 79). This authors also showed that $\rho = 1 - \frac{n}{n-1}\frac{SSW}{SST}$. Solving this equation for SSW, substituting into $E(\hat{s}^2_{sys,i})$ and using the fact that $kn = N$, the result follows.

Proof of Corollary 1. 1:

It follows immediately by simplifying $\frac{E(\hat{s}^2_{sys,i}) - S^2_U}{S^2_U}$, provided that $S^2_U > 0$.

Proof of Corollary 1. 2:

Recall that in the design based approach, N and S^2_U are constants, so the expression $E(\hat{s}^2_{sys,i})$ is linear in ρ.

As it has been shown elsewhere, see for example Kish (1965), the minimum value of ρ is $-1/(n-1)$ and the maximum is *1*. Substitute this values in $E(\hat{s}^2_{sys,i})$ to obtain the maximum and minimum values. On the other hand, solving $E(\hat{s}^2_{sys,i}) = 0$, for ρ implies that $\rho = -1/(N-1)$.

Proof of Corollary 1. 3:

Särndal et al. (1992, p. 79) showed that $\partial = 1 - \frac{N-1}{N-k}\frac{SSW}{SST}$. Solving this equation for SSW we have that $SSW = (N-k)(1-\partial)S^2_U$. Substituting this expression into the formula for the intra-sample variance, the result follows from the expected value of $E(\hat{s}^2_{sys,i})$.

The formula for the relative bias in terms of the measure of homogeneity is obtained by computing $\frac{E(\hat{s}^2_{sys,i}) - S^2_U}{S^2_U}$ in terms of ∂.

Proof of Theorem 2:

Case 1: If $(N-1)/m$ is an integer.

The first element in the sample is selected with probability $1/N$ and an element is included in the circular systematic sample with probability $(N-1)m/N(N-1)$. The factor $(N-1)/N$ corresponds to those elements of the population not selected in the *srswor* of size *1*, and $m/(N-1)$ is the probability of inclusion of an element under *css*, see Murty & Rao (1988). It follows that for $k = 1,..., N$,

$$\pi_k = \frac{1}{N} + \frac{N-1}{N}\frac{m}{N-1} = \frac{1}{N} + \frac{n-1}{N} = \frac{n}{N}.$$

Case 2: If $(N-1)/m$ is not integer.

The proof is equal, since the first-order inclusion probability of an element under *css* is $m/(N-1)$ and the result follows.

Proof of Theorem 3:

Case 1: *N-1* even and eliminating duplicated systematic samples.

Let *ns* denote the number of possible samples under an *mrss(1,m)* design.

$$E(\hat{s}^2_{r,s}) = \frac{m}{N(N-1)} \frac{\sum_{j=1}^{ns}\sum_{k=1}^{m}(y_{r,j} - y_{s,k,j})^2}{2m}$$

Note that for every random selection between *1* and *N*, say *k*, there are *N(N-1)/m* systematic samples and all elements of population *U*, except the *k-th* random number, appear once (for brevity, this *N(N-1)/m* possible samples will be denominated as a *kth*-block). After doing some algebra, a *kth*-block has the following form:

$$\frac{my_k^2 + y_1^2 + ... + y_{k-1}^2 + y_{k+1}^2 + ... + y_N^2 - 2y_k \sum_{i=1}^{k-1} y_i - 2y_k \sum_{i=k+1}^{N} y_i}{2m}.$$

The sum of the *kth*-blocks from *1* to *N* is equal to:

$$\frac{N-1}{m} m(y_1^2 + ... + y_N^2) + (N-1)(y_1^2 + ... + y_N^2) - 4\sum_{I=1}^{N-1} \sum_{j>i}^{N} y_i y_j$$

We substitute this value in the expectation of the sample element variance:

$$E(\hat{s}_{r,s}^2) = \frac{(N-1)\sum_{k=1}^{N} y_k^2 - 2\sum_{i=1}^{N-1} \sum_{j>i}^{N} y_i y_j}{N(N-1)}$$

Using the identity, $(\sum_{k=1}^{N} y_k)^2 = \sum_{k=1}^{N} y_k^2 + 2\sum_{i=1}^{N-1} \sum_{j>i}^{N} y_i y_j$, the last expression turns out to be:

$$E(\hat{s}_{r,s}^2) = \frac{(N-1)\sum_{k=1}^{N} y_k^2 - (\sum_{k=1}^{N} y_k)^2 + \sum_{k=1}^{N} y_k^2}{N(N-1)} = \frac{\sum_{k=1}^{N} y_k^2 - N\bar{y}_U^2}{N-1},$$ which completes the proof.

Case 2: *N-1* odd.

Note that for every random selection between *1* and *N*, say *k*, there are (*N-1)* systematic samples and all elements of population *U*, except random number *k*, appear *m* times (for brevity, this (*N-1)* possible samples will be denominated as a *kth*-block). After doing some algebra, a *kth*-block has the following form:

$$\frac{(N-1)my_k^2 + m(y_1^2 + ... + y_{k-1}^2 + y_{k+1}^2 + ... + y_N^2) - 2my_k \sum_{i=1}^{k-1} y_i - 2my_k \sum_{i=k+1}^{N} y_i}{2m}$$

The sum of the *kth*-blocks from *1* to *N* is equal to:

$$(N-1)m(y_1^2 + ... + y_N^2) + (N-1)m(y_1^2 + ... + y_N^2) - 4m\sum_{I=1}^{N-1} \sum_{j>i}^{N} y_i y_j$$

Using the same identity for the square of a sum as in the previous case and replacing this value in the expectation of the sample element variance the result follows.

Case 3: *N-1* even and without eliminating duplicated systematic samples.

Same proof as case 2.

References

Brewer, K. (2002) *Combined Survey Sampling Inference: Weighing of Basu´s elephant*, London: Arnold.

Chaudhuri, A. & Stenger, H.(2005) *Survey Sampling: theory and methods*, 2nd ed., Chapman & Hall/CRC.

Cochran, W. (1986) *Técnicas de Muestreo*, Ed. CECSA, México.

Horvitz, D.G. & Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 663-685.

Huang, K. (2004) Mixed random systematic sampling designs, *Metrika*, 59, pp. 1-11.

Kish, L. (1965) *Survey sampling*, John Wiley & Sons, New York.

Leu, C. & Tsui, K. (1996) New partially systematic sampling, *Statistica Sinica*, 6, pp. 617-630.

Madow, G. W. & Madow, L. H. (1944) On the theory of systematic sampling, I, *Annals of Mathematical Statistics*, 25, pp. 1-24.

Murthy, M.N. & Rao, T.J. (1988) *Systematic Sampling,* Chapter 7 in Handbook of Statistics 6: Sampling, ed. by C.R. Rao, Amsterdam: North Holland.

Padilla, Terán, A. M. "*Un estimador insesgado de la varianza del muestreo aleatorio simple usando un diseño mixto aleatorio-sistemático*". Memorias electrónicas en extenso de la 2ª Semana Internacional de la Estadística y la Probabilidad, Puebla de Zaragoza, Puebla, México. Julio 2009, CD ISBN: 978-607-487-035-0.

Ruiz-Espejo, M. (1997) Uniqueness of the Zinger strategy with estimable variance: Rana-Singh estimator, *Sankhya*, Volume 59, Series B, pp. 76-83.

Sampath, S. & Uthayakumaran, N. (1998) Markov systematic sampling, *Biometrical Journal*, Vol. 40, Issue 7, pp. 883-895.

Särndal, C.E., Swensson, B. & Wretman, J.H. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Tillé, Y. (2006) *Sampling Algorithms*, Springer-Verlag, New York.

Valliant, R., *An Overview of the Pros and Cons of Linearization versus Replication in Establishment Surveys*, 2007 International Conference on Establishment Surveys, CD-ROM, Alexandria, VA: American Statistical Association: 929-940.

Valliant, R., Dorfman, A. and Royall, R. (2000) *Finite Population Sampling and Inference: a prediction approach*, John Wiley and Sons, New York.

Wolter, K.M. (1985) *Introduction to Variance Estimation*, Springer-Verlag, New York.

Zinger, A. (1980) Variance estimation in partially systematic sampling, *Journal of the American Statistical Association*, Vol. 75, No. 369, pp. 206-211.