Banco de México
Documentos de Investigación

Banco de México
Working Papers

N° 2009-13

**Un Estimador Insesgado de la Varianza del Muestreo
Aleatorio Simple Usando un Muestreo Mixto
Aleatorio-Sistemático**

Alberto Padilla
Banco de México

Noviembre 2009

La serie de Documentos de Investigación del Banco de México divulga resultados preliminares de trabajos de investigación económica realizados en el Banco de México con la finalidad de propiciar el intercambio y debate de ideas. El contenido de los Documentos de Investigación, así como las conclusiones que de ellos se derivan, son responsabilidad exclusiva de los autores y no reflejan necesariamente las del Banco de México.

The Working Papers series of Banco de México disseminates preliminary results of economic research conducted at Banco de México in order to promote the exchange and debate of ideas. The views and conclusions presented in the Working Papers are exclusively the responsibility of the authors and do not necessarily reflect those of Banco de México.

Un Estimador Insesgado de la Varianza del Muestreo Aleatorio Simple Usando un Muestreo Mixto Aleatorio-Sistemático^{*}

Alberto Padilla[†]
Banco de México

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.

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.

^{*}El autor agradece al Dr. Ignacio Méndez del IIMAS-UNAM, a los participantes del seminario del Banco de México, así como a dos revisores del Banco de México por sus comentarios y sugerencias.

[†] Dirección General de Emisión. Email: ampadilla@banxico.org.mx

1. Introducción

El muestreo sistemático es una técnica empleada comúnmente en la práctica debido a su simplicidad y facilidad de aplicación. La principal desventaja de este método es que no existe un estimador insesgado de la varianza poblacional bajo muestreo sistemático con una sola muestra. Este problema se ha abordado de diferentes maneras. En una de ellas se supone que la muestra sistemática se extrajo de una población en orden aleatorio, de tal suerte que puede emplearse la fórmula del estimador de la varianza del estimador de la media bajo muestreo aleatorio simple sin reemplazo, en lo sucesivo $massr$, $\hat{v}_{massr}(\hat{y})$, Cochran (1986). Otro enfoque consiste en suponer un modelo para la variable de interés, por lo cual es necesario desarrollar una fórmula para el estimador de la varianza bajo dicho modelo. Bajo el enfoque del muestreo probabilístico, también puede aplicarse una permutación a los elementos de la población previo a la extracción de la muestra. Con este método, se usa como estimador de varianza a $\hat{v}_{massr}(\hat{\hat{y}})$; empero, este procedimiento no es factible en muchas encuestas. En otra clase de métodos, los denominados mixtos, se suplementa la muestra sistemática con otra muestra sistemática o con una muestra aleatoria simple. Para una amplia discusión de estas estrategias véase Wolter (1985) o Chaudhuri & Stenger (2005). En uno de los métodos mixtos, primero se extrae una muestra aleatoria simple y en el resto de la población se selecciona una muestra sistemática, véase Leu & Tsui (1996) y Huang (2004). También se ha propuesto otra clase de métodos sistemáticos denominada ‘muestreo de Markov’, véase Sampath & Uthayakumaran (1998) y las

referencias ahí citadas. Infortunadamente, estos métodos no pueden emplearse en poblaciones grandes y se requiere, además, que el tamaño de la población sea un múltiplo del tamaño de muestra. En Sampath & Uthayakumaran (1998), por ejemplo, el tamaño de muestra debe ser par, lo cual es una condición difícil de satisfacer en encuestas grandes y casi no han sido utilizados en la práctica. Los métodos arriba mencionados y sus propiedades, han sido estudiados con detalle en la literatura y no serán revisados en este artículo.

En este artículo se propone un diseño mixto aleatorio-sistemático en el que la media poblacional y la varianza poblacional de la media, bajo *mmas*, se estiman insesgadamente con la media muestral y una expresión simple para la varianza¹. Esta última expresión puede ser usada sin suponer que la muestra fue extraída de una población en orden aleatorio o que se aplicó una permutación a la población previo a la selección de la muestra, previniendo al usuario del método de caer en PISE. Este es un acrónimo en inglés acuñado por Valliant (2007) y significa “pretender que es otra cosa” (pretend it’s something else), refiriéndose al uso sin validar supuestos de las fórmulas del *massr* con muestras sistemáticas. Es importante mencionar que no se esperan ganancias en eficiencia del método propuesto comparado con el muestreo sistemático y similares, ya que se tiene un estimador insesgado de la varianza bajo *massr*. Una comparación justa del método propuesto es con el estimador de la varianza entre elementos usando el enfoque de orden aleatorio en el muestreo sistemático.

¹ Este trabajo es una versión extendida de un artículo presentado por el autor en Puebla, Padilla (2009).

El artículo está organizado de la siguiente manera. En la sección 2 se presenta la notación, definiciones y una breve descripción del muestreo de poblaciones finitas. Las prácticas usuales para la estimación de la varianza del diseño bajo el muestreo sistemático se encuentran en la sección 3. En esta sección, también se exhiben las expresiones para el sesgo y sesgo relativo del estimador de la varianza entre elementos bajo el enfoque de orden aleatorio en la población. En la literatura revisada a la fecha, estas expresiones no han sido publicadas. En la sección 4 se presenta el esquema de muestreo junto con un ejemplo. Los estimadores de la media poblacional y varianza $v_{massr}(\hat{y})$ se encuentran en la sección 5. Finalmente, el método se ilustra con un par de ejemplos.

2. Muestreo de poblaciones finitas

Hay dos tipos de encuestas, descriptivas y analíticas. La primera se refiere a la estimación de cantidades poblacionales como promedios, totales, proporciones y razones, en tanto que la segunda al uso de modelos con base en los resultados de una encuesta. Las fórmulas que se desarrollan en este artículo son del tipo descriptivo.

En este artículo se supone que toda la variabilidad proviene del error de muestreo, de tal forma que cualquier error causado por mediciones incorrectas, no respuesta u otro error no de muestreo, son ignorados. También se supone que el diseño no es informativo. Un diseño informativo es aquel en el que la probabilidad de selección de los elementos en la muestra depende explícitamente de los valores de la variable bajo

estudio. De hecho, este último supuesto se emplea en la mayor parte de encuestas y no se menciona explícitamente en libros o artículos.

También se supone que se cuenta con un marco muestral del cual se extrae la muestra.

2.1 Notación, población y muestra

Sea U una población finita con N elementos etiquetados como $k=1, \dots, N$, $1 \leq N$. Es común representar a la población finita por su etiqueta k como: $U=\{1, 2, \dots, k, \dots, N\}$. Más aún, hay una correspondencia uno a uno entre las etiquetas de U y las etiquetas del marco muestral.

La variable bajo estudio se representará como y y y_k será el valor de y para el k -ésimo elemento de la población, $k \in U$.

La muestra se denotará con s , un subconjunto de U de tamaño, n , $1 \leq n \leq N$, y será representado por un vector columna $I = (I_1, \dots, I_k, \dots, I_N)' \in \{0,1\}^N$. En este caso, I_k es una variable indicadora que es igual a 1 si el k -ésimo elemento está en la muestra y 0 en otro caso. Es necesario mencionar que esta variable indicadora es el elemento aleatorio en el muestreo de poblaciones finitas y y_k es un número real. Así, la función de densidad inducida por el diseño es discreta. Este enfoque también se conoce en la literatura como muestreo basado en el diseño.

2.2 Estimación

El objetivo es estimar una función t que depende de y_k , $t = t(y_1, \dots, y_k, \dots, y_N)$. Por ejemplo, un total se escribe como $y_U = \sum_{k=1}^N y_k$. Debido a que se desea estimar un total, bajo el enfoque del diseño, es usual emplear el estimador de Horvitz-Thompson, *EHT*, Horvitz & Thompson (1952). Este estimador tiene la siguiente forma:

$$\hat{y}_U = \sum_{k=1}^N I_k y_k / \pi_k = \sum_{k=1}^n y_k / \pi_k, \text{ con } \pi_k > 0.$$
 En esta fórmula, $\pi_k = P(I_k = 1)$ es la probabilidad de inclusión de primer orden. Para el cálculo y estimación de la varianza es necesario determinar las probabilidades de inclusión de segundo orden, $\pi_{kl} = P(I_k I_l = 1)$.

La varianza de un *EHT* es,

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

Un estimador insesgado de esta varianza está dado por la siguiente expresión, siempre que $\pi_{kl} > 0$:

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

En estas expresiones, $c(I_k, I_l)$ y $\hat{c}(I_k, I_l)$ denotan las covarianzas poblacionales y muestrales respectivamente, entre las variables indicadoras muestrales.

Observación 2.2.1: Cabe mencionar que el muestreo de poblaciones finitas, los dos primeros momentos están bien definidos para los principales diseños usados en la práctica, de tal forma que no hay necesidad de mencionar esto en el resto del documento.

Observación 2.2.2: La estimación en el muestreo de poblaciones finitas también puede hacerse bajo un enfoque diferente conocido en la literatura como muestreo basado en modelo, en el cual se supone que la población finita es una muestra de una población infinita (superpoblación), véase Valliant et al. (2000). Los enfoques de muestreo basado en el diseño y en el modelo pueden emplearse juntos en lo que se conoce como muestreo combinado, ver Brewer (2002).

3. Prácticas comunes en el muestreo sistemático

Como se mencionó en la introducción, no existe un estimador insesgado de la varianza del promedio, bajo el enfoque del diseño, con una muestra sistemática. Por lo anterior, en la práctica se emplean las siguientes estrategias, entre otras.

3.1 Durante la etapa del diseño de la encuesta

D1) Suplementar la muestra sistemática con otra muestra.

D2) Aplicar una permutación aleatoria a los elementos de la población previo a la extracción de la muestra, de tal forma que bajo todas las posibles permutaciones de

la población, la esperanza de la varianza bajo el enfoque del diseño es la misma que la varianza bajo *massr*. Este resultado se debe a Madow & Madow (1944).

Observación 3.1.1: Una comparación de la eficiencia entre algunos diseños del tipo *DI*, pueden encontrarse en Zinger (1980), Cochran (1986) y Wolter (1985).

3.2 Modelo para la estructura de la variable de interés

Se postula un modelo para la estructura de la variable bajo estudio previo a la extracción de la muestra sistemática y se construye un estimador de la varianza según el modelo. En este caso se emplean regularmente dos modelos:

M_{sc}) Correlación serial: este modelo resultará adecuado en aquellas situaciones en las que se tiene evidencia de similitud entre elementos vecinos de la población para la variable de interés y, además, esta similitud disminuye para elementos lejanos entre sí.

M_{ro}) Modelo de orden aleatorio en una población infinita: la población finita se considera como una muestra aleatoria de una población infinita (superpoblación). Se dice que se tiene una población en orden aleatorio, si las $y_i, i = 1, \dots, N$, son extraídas de una superpoblación en la que se cumple que: $E_M(y_i) = \mu$, $E_M(y_i - \mu)^2 = \sigma_i^2$ y $E_M(y_i - \mu)(y_j - \mu) = 0, i \neq j$. En estas expresiones, E_M se refiere a la esperanza bajo el modelo supuesto. El resultado de esto es que $E_M(\hat{v}_{sis}(\hat{y})) = E_M(\hat{v}_{massr}(\hat{y}))$, donde *sis* se refiere al muestreo sistemático, véase

Cochran (1986). En este modelo se supone que no hay relación entre la variable de estudio y el orden de los elementos en el marco muestral, de tal forma que una muestra sistemática extraída de este marco, ordenado con algún criterio específico, se trata como si el marco hubiera sido ordenado aleatoriamente.

Observación 3.2.1: Una comparación de la eficiencia de los modelos M_{sc} y M_{ro} , se encuentra en Wolter (1985) y Chaudhuri & Stenger (2005).

3.3 Sesgo del enfoque de orden aleatorio (M_{ro})

Bajo el enfoque M_{ro} , se emplea el estimador de la varianza de la media bajo muestreo aleatorio simple, $\hat{v}_{massr}(\hat{y}) = (1 - n/N)\hat{s}_{sis}^2/n$. En esta expresión, $\hat{s}_{sis}^2$ se refiere a la varianza entre elementos de la muestra sistemática. Esta es una estrategia adecuada siempre que se tenga información acerca del orden aleatorio de los elementos de la población. El problema es que fácilmente puede uno caer en PISE y trabajar con un estimador sesgado de la varianza o usar mecánicamente el estimador bajo $massr$ sin contar con información suficiente acerca del orden de los elementos en la población. Con el fin de evaluar este enfoque, en el siguiente teorema se exhiben el sesgo y sesgo relativo del estimador de la varianza. Supóngase que $k = N/n$ es un entero y

$$\hat{s}_{sis,i}^2 = \sum_{j=1}^n (y_{ij} - \hat{y}_i)^2 / (n-1), \quad \hat{y}_i = \sum_{j=1}^n y_{ij} / n, \quad S_U^2 = \sum_{j=1}^N (y_j - \bar{y}_U)^2 / (N-1),$$

$$\bar{y}_U = \sum_{j=1}^N y_j / N \text{ y } \rho = 2 \sum_{l=1}^k \sum_{i=1}^{n-1} \sum_{j>i}^n (y_{li} - \bar{y}_U)(y_{lj} - \bar{y}_U) / ((n-1)(N-1)S_U^2), \text{ donde } \rho \text{ es}$$

el coeficiente de correlación intraclase, Cochran (1986).

Teorema 1: Bajo muestreo sistemático, el valor esperado del estimador $\hat{s}_{sis,i}^2$ es

$$\frac{N-1}{N}(1-\rho)S_U^2.$$

Corollary 1.1: El sesgo relativo del estimador $\hat{s}_{sis,i}^2$ es $\frac{N-1}{N}(1-\rho)-1$.

Corollary 1.2: $E(\hat{s}_{sys,i}^2)$ es una función lineal decreciente de ρ , que se maximiza en

$\rho = -1/(n-1)$, minimiza en $\rho = 1$ y $E(\hat{s}_{sis,i}^2) = S_U^2$, siempre que $\rho = -1/(N-1)$. Los

valores máximos y mínimos de $E(\hat{s}_{sis,i}^2)$ son $S_U^2 \frac{n}{n-1} \frac{N-1}{N}$ y cero respectivamente.

Corollary 1.3: El valor esperado y sesgo relativo del estimador $\hat{s}_{sis,i}^2$ puede ser

expresado como $(1-\partial)S_U^2$ y $-\partial$, donde ∂ es una medida de homogeneidad propuesta

por Särndal et al (1992).

Prueba: véase el Apéndice.

Observación 3.3.1: Del corolario 1.2 se observa que $E(\hat{s}_{sis,i}^2)$ sobrestima a S_U^2 para

$$\rho \in \left[\frac{-1}{n-1}, \frac{-1}{N-1} \right).$$

4. Diseño

4.1 Definición del muestreo mixto aleatorio-sistemático

Usando el enfoque del diseño o muestreo probabilístico, se supondrá que se tiene una población U , con N elementos, $y_k, k = 1, \dots, N$. De esta población se extrae una muestra de tamaño n , $1 < n < N$, por medio de un muestreo mixto aleatorio-sistemático, $mmas$. Es decir, primero se selecciona una muestra $massr$ de tamaño l de los elementos de U y después se extraen m elementos, $m \geq 2$, de los $N-l$ elementos restantes de U usando muestreo sistemático circular, Murty & Rao (1988). Este método se denotará como $mmas(l, m)$. El número de muestras bajo este diseño es $N(N-1)$.

Observación 4.1.1: Cuando $(N-1)/m$ es un entero, el muestreo sistemático circular y lineal coinciden, Murty & Rao (1988), de tal forma que una muestra sistemática puede extraerse con este último esquema. En este caso, hay muestras circulares que se repiten; sin embargo, los estimadores puntuales de la media poblacional y la varianza poblacional entre elementos, que se construyen en la sección siguiente, continúan siendo insesgados aún después de suprimir las muestras repetidas.

Observación 4.1.2: El número de muestras bajo un diseño $mmas(l, m)$, después de eliminar las muestras repetidas es $N(N-1)/m$ si $(N-1)/m$ es un entero y $N(N-1)$ en otro caso. Para mayores detalles véase Murthy & Rao (1988).

4.2 Muestreo sistemático circular

Para obtener una muestra sistemática circular, msc , de tamaño m , $1 < m < M$ de una población con M elementos, se procede de la manera siguiente:

Paso 1: calcule $k_m = (N - 1)/m$; si k_m no es entero, redondee al entero siguiente,

Paso 2: seleccione un número aleatorio entre 1 y M , digamos r , este es el primer elemento de la msc ,

Paso 3: calcule los siguientes números en la msc , $r + jk_m$, para $j \in \{1, \dots, m-1\}$. Si $r + jk_m > M$ considere la lista como circular y asigne los números hasta obtener el tamaño de muestra requerido.

Observación 4.2.1: este procedimiento se puede implementar fácilmente en una hoja de cálculo o en el sistema R .

Ejemplo 1: sea U una población de tamaño $N=7$ y supóngase que se desea extraer una muestra de tamaño $n=3$ usando un diseño $mmas(1,2)$. En este caso $m=2$ y se tienen $7(7-1)=42$ muestras. Los índices de las posibles muestras son:

Tabla 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

El primer número en cada entrada se refiere a la selección por *massr*, en tanto que los otros dos corresponden a la muestra sistemática.

5. Estimadores puntuales

Huang (2004) observó que en el muestreo mixto aleatorio-sistemático el estimador de Horvitz-Thompson, *EHT*, $\hat{y} = 1/N \sum_{k=1}^n y_k / \pi_k$, $\pi_k > 0$ puede usarse para estimar la media poblacional, siempre que N sea conocida. Para calcular este estimador, solo se requiere conocer las probabilidades de inclusión de primer orden.

Teorema 2: Bajo *mmas*(1, m), las probabilidades de inclusión de primer orden, π_k , son iguales a n/N , para toda $k = 1, \dots, N$.

Prueba: véase el Apéndice.

Corolario: Para un diseño *mmas*(1, m), el *EHT* es el promedio aritmético.

Prueba: se obtiene al sustituir $\pi_k = n/N$ en la expresión del *EHT* de la media.

Observación 5.1: Con un diseño *mrss*(1, m) el estimador de la media también puede ser expresado como una suma ponderada, $\hat{y}_{r,s} = \beta y_r + \alpha \hat{y}_s$, con $\beta = 1/n$, $\alpha = m/n$. El primer término de esta suma se refiere al valor de y obtenido por *massr*, en tanto que el

segundo es la media muestral del muestro sistemático. Esto se conoce como el estimador Zinger, Ruiz-Espejo (1997).

Observación 5.2: El estimador de la media bajo un diseño $mmas(l, m)$ es insesgado porque se trata de un *EHT*.

El resultado más importante de este artículo se encuentra en el siguiente teorema.

Teorema 3: Bajo $mmas(l, m)$, un estimador insesgado de la varianza poblacional entre

elementos, $s_U^2 = \sum_{k=1}^N (y_k - \bar{y}_U)^2 / (N - 1)$, es:

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

donde y_r es el valor de la variable seleccionada por *massr*; $y_{s,k}$ son los valores de los elementos seleccionados por la muestra sistemática circular y $\bar{y}_U$ es la media poblacional.

Prueba: véase el Apéndice.

Corolario: Con un diseño $mmas(l, m)$, un estimador insesgado de la varianza de la media de un *massr*, $v_{massr}(\hat{y})$, está dado por la siguiente expresión,

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

Prueba: es inmediata por la linealidad del operador esperanza, $E(cX) = cE(X)$, donde

$$c = (1 - n/N)/n.$$

Observación 5.3: No se hicieron supuestos acerca del orden aleatorio en la población ni fue necesario permutar los elementos de la población previo a la extracción de la muestra. En resumen, el $mmas(I, m)$ proporciona una expresión simple para la estimación de la varianza sin pretender que es otra cosa, Valliant (2007).

Observación 5.4: En la expresión $\hat{v}_{massr}(\hat{y}_{r,s})$ puede usarse un tamaño de muestra m para la estimación.

Observación 5.5: Zinger (1980) propuso un estimador insesgado de la varianza entre elementos usando un muestreo parcialmente sistemático en el que primero seleccionaba una muestra sistemática y después una *massr* de la población restante. Infortunadamente, la fórmula propuesta por Zinger es compleja.

6. Ejemplo numérico

Ejemplo 2: sea U la población del ejemplo 3.4.2, páginas 80-82, de Särndal et al. (1992). Esta población tiene $N=100$ elementos y la variable y toma los valores $1, 2, \dots, 100$. Usando muestreo sistemático con $n=10$, hay $N/n=10$ muestras y la media poblacional $\bar{y}_U$ y la varianza entre elementos S_U^2 son 50.5 y 841.67 respectivamente. Debido a que el muestreo aleatorio simple no se ve afectado por el orden de los elementos en la población la varianza del estimador de la media bajo este diseño es $v_{massr}(\hat{y}) = (1 - n/N) S_U^2 / n = 75.75$. En las tablas 2 a 5 hay cuatro ordenamientos de la misma población con diferentes valores del coeficiente de correlación intraclase. Para

cada ordenamiento y para todas las muestras posibles bajo muestreo sistemático, se presentan los valores de la media muestral, $\hat{y}_{sis}$, el estimador de la varianza entre elementos, $\hat{s}_{sis}^2$, y el estimador de la varianza de la media muestral con el supuesto de orden aleatorio, $\hat{v}_{ro}(\hat{y}_{sis})$. Bajo el supuesto de orden aleatorio, se calcularon los estimadores de cada muestra sistemática $\hat{s}_{sis}^2$ y $\hat{v}_{ro}$ usando las siguientes fórmulas:

$$\hat{s}_{sis,i}^2 = \sum_{k=1}^{10} (y_k - \hat{y}_{sis,i})^2 / (10-1)^2 \text{ y } \hat{v}_{ro}(\hat{y}_{sis,j}) = (1-10/100) \hat{s}_{sis,j}^2 / 10.$$

Las etiquetas s-1, s-2,...,s-10 corresponden a los resultados de las muestras 1 a la 10. La última columna tiene los valores esperados de las medias muestrales y las varianzas, $E(\hat{y}_{sis})$, $E(\hat{s}_{sis}^2)$ y $E(\hat{v}_{ro})$ respectivamente.

Tabla 2

Población A: tendencia lineal perfecta para los valores y_k , $\rho_{oh} = -0.10$.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{y}_{sis}$	46.0	47.0	48.0	49.0	50.0	51.0	52.0	53.0	54.0	55.0	50.5
$\hat{s}_{sis}^2$	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

Tabla 3

Población B: un ordenamiento con varianza mínima para muestreo sistemático, $\rho_{oh} = -0.11$.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{y}_{sis}$	50.5	50.5	50.5	50.5	50.0	50.5	50.5	50.5	50.50	50.50	50.5
$\hat{s}_{sis}^2$	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

Tabla 4

Población C: un valor grande de roh, roh= 0.989.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{y}_{sis}$	5.5	15.5	25.5	35.5	45.5	55.5	65.5	75.5	85.5	95.5	50.5
$\hat{s}_{sis}^2$	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

Tabla 5

Población D: un ordenamiento aleatorio, roh= -0.015.											
	s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	
$\hat{y}_{sis}$	44.3	34.8	40.7	61.2	48.8	59.5	47.6	58.7	58.4	51.0	50.5
$\hat{s}_{sis}^2$	720.9	420.0	1014.7	948.2	494.4	948.7	1222.5	522.7	780.5	1388.4	916.7
$\hat{v}_{ro}$	64.9	37.8	91.3	85.3	44.5	85.4	110.0	47.0	70.2	125.0	76.1

Con el fin de comparar las estrategias de estimación de la varianza entre elementos suponiendo orden aleatorio de la población en el muestreo sistemático y el muestreo mixto aleatorio-sistemático, para las poblaciones A a D, se usó un diseño $mmas(1,9)$. En este caso, hay $100(100-1)=9,900$ muestras posibles bajo el muestreo mixto aleatorio-sistemático. Se generaron para cada población, las 9,900 muestras y se calculó el coeficiente de variación de la varianza entre elementos, $\hat{s}_{r,s}^2$, para evaluar al estimador de la varianza.

Tabla 6

	A	B	C	D
Media poblacional $\bar{y}_U =$	50.5	50.5	50.5	50.5
$S_U^2 =$	841.7	841.7	841.7	841.7
$v_{massr}(\hat{\hat{y}}) =$	75.75	75.75	75.75	75.75
Muestreo sistemático:				
Correlación intraclase=	-0.10	-0.11	0.989	-0.015
Estimador orden aleatorio $\hat{S}_{sis}^2 =$	916.7	925.8	9.2	846.1
Sesgo relativo ($\hat{S}_{sis}^2$)=	8.9%	10.0%	98.9%	0.5%
Estimador de varianza $\hat{v}_{ro}(\hat{\hat{y}}_{sis}) =$	82.5	83.3	0.83	76.1
Coefficiente de variación ($\hat{\hat{y}}_{sis}$)=	6.0%	0%	60.0%	17.7%
Coefficiente de variación ($\hat{S}_{sis}^2$)=	0%	3.7%	0%	37.7%
Muestreo mixto aleatorio-sistemático:				
Estimador de varianza $\hat{v}_{massr}(\hat{\hat{y}}_{r,s}) =$	75.75	75.75	75.75	75.75
Coefficiente de variación ($\hat{\hat{y}}_{r,s}$)=	7.7%	7.7%	7.7%	21.3%
Coefficiente de variación ($\hat{S}_{r,s}^2$)=	46.0%	46.3%	46.6%	60.9%

En la tabla 6, las letras en la parte superior de las columnas 2 al 5, corresponden a las poblaciones de las tablas 2 al 5. Al comparar los estimadores de varianza $\hat{v}_{ro}(\hat{\hat{y}}_{sis})$, $\hat{v}_{massr}(\hat{\hat{y}}_{r,s})$ y los coeficientes de variación de los estimadores de la media poblacional y la varianza entre elementos para ambos diseños, se observa que los estimadores bajo el supuesto de orden aleatorio en el muestreo sistemático se comportan erráticamente y dependen fuertemente del orden en la población. Con el muestreo mixto aleatorio-sistemático no se presentaron estimaciones erráticas para las poblaciones A a C; empero, para la población D, las distribuciones muestrales de $\hat{\hat{y}}_{r,s}$ y $\hat{S}_{r,s}^2$ tuvieron mayor variación que sus contrapartes en el muestreo sistemático. Esto se debió a la presencia de observaciones influyentes en la distribución de $\hat{S}_{r,s}^2$.

7. Conclusiones

Se construyó un estimador insesgado de la varianza poblacional entre elementos bajo muestreo aleatorio simple sin reemplazo usando un diseño mixto aleatorio-sistemático. Se demostró que no es necesario suponer orden aleatorio de los elementos de la población ni permutar las unidades de la población previo a la extracción de la muestra sistemática para usar el estimador insesgado de la varianza poblacional entre elementos. También se mostró que el sesgo y sesgo relativo del estimador de la varianza poblacional entre elementos con muestreo sistemático con el supuesto de orden aleatorio de unidades de la población depende del coeficiente de correlación intraclase.

Apéndice

Prueba del Teorema 1:

Supóngase que $N = nk$, $1 < n < N$, k y n son enteros.

Observe que la variación entre elementos en la población puede descomponerse como:

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

Esta es la descomposición de la variación total en la suma de la variación dentro de muestras sistemáticas y la variación entre muestras sistemáticas, como se efectúa en el análisis de la varianza y se puede escribir como:

$$SST = SSW + SSB$$

Aquí, SS representa la suma de cuadrados; T , el total; W , dentro y B , entre. La prueba consiste en calcular la esperanza de la varianza muestral entre elementos de la muestra

sistemática, $s_{sis,i}^2 = \sum_{j=1}^n (y_{ij} - \hat{y}_i)^2 / (n-1)$.

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

Sumamos $-N\bar{y}_U + kn\bar{y}_U$ en la última expresión y observamos que $k(n-1) = N-k$,

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

Como $SST = (N-1)S_U^2$, se tiene que $E(\hat{s}_{sis,i}^2) = \frac{SSW}{N-k}$.

Esta es la varianza intramuestral propuesta por Särndal et al. (1992, p. 79). Estos autores mostraron que $\rho = 1 - \frac{n}{n-1} \frac{SSW}{SST}$. Al resolver esta ecuación para SSW, sustituyendo en

$E(\hat{s}_{sis,i}^2)$ y usando el hecho de que $kn = N$, se obtiene el resultado deseado.

Prueba del Corolario 1. 1:

El resultado se obtiene al simplificar $\frac{E(\hat{s}_{sis,i}^2) - S_U^2}{S_U^2}$, siempre que $S_U^2 > 0$.

Prueba del Corolario 1. 2:

Recordemos que en el enfoque del diseño, N y S_U^2 son constantes; así, la expresión $E(\hat{s}_{sis,i}^2)$ es lineal en ρ .

Como ya ha sido demostrado, véase por ejemplo Kish (1965), el valor mínimo de ρ es $-1/(n-1)$ y el máximo $\rho = 1$. Al sustituir estos valores en $E(\hat{s}_{sis,i}^2)$ se obtienen los valores máximo y mínimo respectivamente. Por otro lado, al resolver $E(\hat{s}_{sis,i}^2) = 0$, para ρ se tiene que $\rho = -1/(N-1)$.

Prueba del Corolario 1. 3:

Särndal et al. (1992, p. 79) mostró que $\partial = 1 - \frac{N-1}{N-k} \frac{SSW}{SST}$. Al resolver esta ecuación para

SSW se tiene que $SSW = (N-k)(1-\partial)S_U^2$ y sustituirla en la fórmula para la varianza intramuestral, se obtiene el resultado deseado del valor esperado de $E(\hat{s}_{sis,i}^2)$.

La fórmula para el sesgo relativo en términos de la medida de homogeneidad, se obtiene al

calcular $\frac{E(\hat{s}_{sis,i}^2) - S_U^2}{S_U^2}$ en términos de ∂ .

Prueba del Teorema 2:

Caso 1: $(N-1)/m$ es un entero.

El primer elemento de la muestra se selecciona con probabilidad $1/N$ y un elemento se incluye con probabilidad $(N-1)m/N(N-1)$ en la muestra sistemática circular. El factor $(N-1)/N$ corresponde a aquellos elementos de la población que no fueron seleccionados en la *massr* de tamaño 1 , y $m/(N-1)$ es la probabilidad de inclusión de un elemento en la muestra sistemática circular, véase Murty & Rao (1988). De lo anterior se tiene que, para cada $k = 1, \dots, N$,

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

Case 2: $(N-1)/m$ no es entero.

La prueba es igual, ya que la probabilidad de inclusion de primer orden de un elemento bajo una muestra sistemática circular es $m/(N-1)$.

Prueba del Teorema 3:

Caso 1: $N-1$ par y eliminando las muestras sistemáticas duplicadas.

Sea ns el número de muestras posibles bajo un diseño $mmas(l, m)$.

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

Obsérvese que para cada selección aleatoria entre 1 y N , digamos k , se tienen $N(N-1)/m$ muestras sisemáticas y todos los elementos de la población U , a excepción del k -ésimo número aleatorio, aparecen una vez (por simplicidad, estas $N(N-1)/m$ posibles muestras se denominarán como el k -ésimo bloque). Después de efectuar algunas manipulaciones algebraicas, el k -ésimo bloque tiene la siguiente forma:

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

La suma de los k -ésimos bloques de 1 a N es igual a:

$$\frac{N-1}{m} m(y_1^2 + \dots + y_N^2) + (N-1)(y_1^2 + \dots + y_N^2) - 4 \sum_{l=1}^{N-1} \sum_{j>l}^N y_l y_j$$

Sustituimos este valor en la esperanza de la varianza muestral entre elementos:

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

Usando la identidad, $(\sum_{k=1}^N y_k)^2 = \sum_{k=1}^N y_k^2 + 2\sum_{i=1}^{N-1} \sum_{j>i}^N y_i y_j$, la última expression resulta ser:

$$E(\hat{s}_{r,s}^2) = \frac{(N-1)\sum_{k=1}^N y_k^2 - (\sum_{k=1}^N y_k)^2 + \sum_{k=1}^N y_k^2}{N(N-1)} = \frac{\sum_{k=1}^N y_k^2 - N\bar{y}_U^2}{N-1}, \text{ lo cual complete la}$$

prueba.

Case 2: $N-1$ impar.

Observe que para cada número aleatorio entre 1 y N , digamos k , hay $(N-1)$ muestras sistemáticas y todos los elementos de la población U , a excepción del número aleatorio k , aparecen m veces (por brevedad, estas $(N-1)$ posibles muestras se denominarán como el k -ésimo bloque). Después de efectuar algunas manipulaciones algebraicas, el k -ésimo bloque tiene la siguiente forma:

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

La suma de los k -ésimos bloques de 1 a N es igual a:

$$(N-1)m(y_1^2 + \dots + y_N^2) + (N-1)m(y_1^2 + \dots + y_N^2) - 4m \sum_{l=1}^{N-1} \sum_{j>l}^N y_l y_j$$

Usando la misma identidad para el cuadrado de una suma, como en el caso previo, y reemplazando este valor en la esperanza de la varianza entre elementos, se obtiene el resultado deseado.

Caso 3: $N-1$ par y sin eliminación de las muestras sistemáticas duplicadas.

Misma prueba que en el caso 2.

Bibliografía

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.