ESTADISTICA ESPAÑOLA
Vol. 36, Núm. 135, 1994, págs. 143 a 162

Método de exploración secuencial cuasi equilibrado en muestreo en poblaciones finitas

por
FRANCISCO R. FERNANDEZ GARCIA y JOSE A. MAYOR GALLEGO
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas. Universidad de Sevilla

RESUMEN

En este trabajo se propone un procedimiento de muestreo de tipo secuencial que proporciona probabilidades de inclusión de primer orden proporcionales al tamaño, sobre bloques de la población, siendo las de segundo orden nulas para pares de elementos similares, lo que se traduce en una disminución del error de muestreo.

Para la estimación de la varianza se usa un esquema de replicación que puede ser aplicado en paralelo, bastando una única exploración de la población.

Se realiza un estudio computacional comparativo con procedimientos existentes de tipo ΠPS, tanto de naturaleza secuencial como no secuencial, mediante estimaciones repetidas sobre poblaciones construidas por simulación.

Palabras clave: muestreo con probabilidades variables, muestreo secuencial, estimador de Horvitz-Thompson, modelo de superpoblación, esquema replicado.

Clasificación AMS: 62D05.

1. INTRODUCCION

La obtención de una muestra a partir de una población finita, con objeto de estimar ciertas características poblacionales, está supeditada a las posibilidades de acceso a los diferentes individuos de la población, de manera que, en la selección de los mismos para formar parte de la muestra, resulta de gran importancia el método de acceso a los elementos del marco.

Desde este punto de vista, hemos de considerar métodos de muestreo adaptados a situaciones en las cuales sólo es posible la exploración secuencial de la población, usualmente debido a las limitaciones inherentes al soporte de la misma. Así, por ejemplo, el caso de una población formada por elementos que están siendo producidos en una cadena de fabricación y cuya calidad se quiere estudiar. Otro ejemplo típico se da en auditoría, cuando es necesario estudiar de forma sucesiva las diferentes cuentas de un sistema contable.

Estos métodos realizan un barrido sobre la población, dispuesta en cierto orden, de forma que una vez explorado el elemento final, la muestra está completa.

En su aspecto más simple, no utilizan información auxiliar para seleccionar los elementos, de manera que la existencia de individuos muy diferentes, en relación a ciertas características, no es tenida en cuenta. Así, si realizamos una exploración secuencial en una población

$$U = \{u_1, u_2, u_3, \ldots, u_N\}$$

con N elementos, dispuestos en cualquier orden, y cada elemento de la misma, independientemente de los demás, es seleccionado con probabilidad p, obtendremos como espacio muestral el conjunto de todas las muestras posibles. Si se representa por $n\ (m)$ el tamaño de la muestra m, cada una de éstas tendrá probabilidad $p\ (m) = p^{n\ (m)}\ (1 - p)^{N - n\ (m)}$. El diseño muestral originado se conoce como diseño muestral de Bernouilli (véase Särndal, Swensson y Wretman, 1992).

Como puede observarse, el tamaño muestral es variable pero su esperanza viene dada por $E\ [n\ (m)] = Np$, lo que permite cierto control sobre el mismo.

Otro procedimiento de este tipo, que denominamos *método de inserción* y que proporciona el diseño muestral aleatorio simple, consiste en recorrer la población generando para cada elemento un número aleatorio entre 0 y 1. En todo momento se mantiene una lista de elementos, *ordenada* por la magnitud de los correspondientes números aleatorios, de menor a mayor.

Los n primeros elementos son introducidos directamente en la lista. A partir del elemento n-ésimo, dicha lista es actualizada insertando probabilísticamente el elemento u_i en el lugar correspondiente y eliminando en su caso otro elemento almacenado.

De esta forma, una vez explorada toda la población, se tiene una muestra de n elementos, perteneciente a un diseño muestral aleatorio simple. Hemos de observar que la aplicación de este método no requiere el conocimiento del tamaño de la población, N, siendo posible su aplicación incluso en aquellos casos en los cuales la población no está delimitada de antemano. Consideremos como ejemplo el caso de una línea de producción con un número de unidades a producir no controlado con exactitud.

En la sección 5 de este trabajo se expone una versión algorítmica de este procedimiento, así como una justificación del mismo.

Observemos que ninguno de los métodos descritos tiene en cuenta las diferencias existentes entre los elementos de la población, lo que puede producir estimaciones poco precisas sobre todo en poblaciones muy dispersas. Así, si queremos estimar el total, $T(Y)$, de una variable, Y, definida sobre la población,

$$T(Y) = \sum_{i \in U} Y_i$$

a partir de una muestra m, obtenida de un diseño muestral con tamaño de muestra fijo y probabilidades de inclusión π_i y π_{ij}, la estimación de Horvitz-Thompson tiene por varianza

$$V[\hat{T}(Y)] = -\frac{1}{2} \sum_{i,j \in U} \sum (\pi_{ij} - \pi_i \pi_i) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

cuya disminición puede conseguirse por dos vías:

a) Por una parte, si las probabilidades de inclusión de primer orden son proporcionales a los valores Y_i tendremos varianza nula. Este supuesto no es realista por razones evidentes; sin embargo, sí podemos intentar que dichas probabilidades de inclusión sean proporcionales a los valores de la variable auxiliar, X, relacionada con Y, con lo que conseguimos, si no anular la varianza, al menos reducirla.

b) Por otra parte, también podemos disminuir la varianza intentando que los valores $-\Delta_{ij} = \pi_i \pi_i - \pi_{ij}$ sean pequeños (e incluso nulos) para aquellos pares (i, j) tales que los correspondientes valores Y_i e Y_j sean muy diferentes, pues son éstos los que proporcionan valores $(Y_i / \pi_i - Y_j / \pi_j)^2$ mayores. Nuevamente, habrá que utilizar la variable auxiliar para conseguir este objetivo.

Existen varios procedimientos de exploración secuencial que intentan la reducción de varianza por alguna de las vías mencionadas, dando lugar a diferentes diseños muestrales.

Citemos primeramente el método sistemático de Madow (1949), usado en auditoría, con el cual se consigue que las probabilidades de inclusión de primer orden sean proporcionales a una variable auxiliar, conocida para toda la población. Al ser de tipo sistemático, presenta el inconveniente de la estimación de la varianza, usual en los mismos. Véanse Hájek (1981) y Bellhouse (1988).

Otro método de esta naturaleza es el de Sunter (1977), con el cual se consigue también que las probabilidades de inclusión sean proporcionales al tamaño, aunque no necesariamente para todos los elementos de la población. Este método requiere que la población sea ordenada en orden decreciente, mediante la variable auxiliar.

Finalmente mencionamos el método de exploración secuencial denominado del *estrato móvil,* Tillé (1994), con el cual se consigue que las probabilidades de inclusión de primer orden sean constantes, siendo las de segundo orden tanto mayores cuanto más diferentes son los elementos correspondientes (en relación a una variable auxiliar que se usa para ordenar previamente la población). Este método incide, pues, en la segunda vía de reducción de varianza.

En este trabajo se propone un método de exploración secuencial, que llamamos cuasi equilibrado, y que obtiene la muestra mediante un proceso del tipo anteriormente descrito, consiguiendo probabilidades de inclusión de primer orden proporcionales al tamaño sobre bloques de la población, lo que, como se verá, tiene el mismo efecto de reducción de varianza que los diseños πPS usuales. Por otra parte, las probabilidades de inclusión de segundo orden son nulas para pares de elementos similares, lo que se traduce en una disminución del error de muestreo.

El trabajo se ha estructurado de la siguiente forma:

En la sección dos se presenta el método de exploración secuencial cuasi equilibrado, estudiándose el diseño muestral resultante.

En la sección tres comparamos este método con el mencionado anteriormente de Madow, y con el método de Sampford, de naturaleza no secuencial, tanto cuando la población está dispuesta aleatoriamente como cuando posee cierta estructura.

En la sección cuatro se construye una estimación insesgada del error de muestreo, y se realiza un estudio computacional de la variabilidad de dicha estimación, comparando el método secuencial equilibrado con el método de Sampford.

Finalmente, en la sección cinco se agrupan, en forma algorítmica, los procedimientos propuestos en el trabajo.

2. METODO DE EXPLORACION SECUENCIAL CUASI EQUILIBRADO

Vamos a suponer que sobre la población U existe una variable de estudio, Y, que toma los valores Y_i, $i \in U$, cuyo total, $T(Y) = \sum_U Y_i$, queremos estimar.

Asumimos la existencia de una variable auxiliar, X_i, relacionada con la anterior según el siguiente modelo de superpoblación:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$E_s[\varepsilon_i] = 0$$
$$V_s[\varepsilon_i] = \sigma^2$$
$$E_s[\varepsilon_i \varepsilon_j] = 0 \qquad i \neq j$$

donde $E_s[\cdot]$ y $V_s[\cdot]$ indican, respectivamente, los operadores esperanza y varianza con respecto a dicho modelo de superpoblación.

Supongamos que queremos extraer n elementos de la población U, y denotemos $T(X) = \sum_U X_i$. Supongamos también que la población está ordenada de menor a mayor, por los valores de la variable auxiliar, es decir,

$$X_1 \leq X_2 \leq X_3 \leq \ldots X_{N-1} \leq X_N$$

y llamemos $T_i = \sum_{k=1}^{i} X_k$, siendo $T_0 = 0$.

Los valores de la variable auxiliar han de verificar la condición usual para la aplicabilidad de los métodos ΠPS, es decir,

$$X_i \leq \frac{1}{n} T(X) \qquad \forall\, i \in U$$

de manera que si uno o varios elementos no la cumplen, entran directamente en la muestra, siendo apartados de la población.

Utilizando el rango de variación de la variable X, definimos n subpoblaciones como sigue:

$$U_1 = \{i \in U \mid T_{i-1} < \frac{T(X)}{n} \leq T_i\}$$

$$U_k = \{i \in U \mid T_{i-1} < k\,\frac{T(X)}{n} \leq T_i\} - \bigcup_{j=1}^{k-1} U_j \qquad k = 2, 3, \ldots, n$$

Si suponemos que las unidades de la población tienen tamaños muy pequeños en relación a $T(X) / n$, entonces las cantidades

$$T_X(U_k) = \sum_{i \in U_k} X_i \qquad k = 1, \ldots, n$$

serán muy similares a $T(X) / n$, siendo ésta la razón de que el método se denomine cuasi equilibrado.

Observemos que en una exploración secuencial de la población U se recorren sucesivamente las subpoblaciones $U_1, U_2, \ldots, U_N$.

En particular, en la subpoblación U_k, vamos a considerar el siguiente procedimiento para extraer un elemento, donde con $\mathcal{U}[0,1)$ denotamos una distribución uniforme continua en $[0,1)$.

1. Generar un número aleatorio $r \sim \mathcal{U}[0,1)$.
2. Seleccciónar el elemento $u_i \in U_k$ tal que

$$T_{i-1} \leq r\, T_X(U_k) + \sum_{j=1}^{k-1} T_X(U_j) < T_i$$

Así, después de un barrido completo de la población, habremos seleccionado n elementos de la misma, uno por cada subpoblación.

En el siguiente teorema se calculan los valores de las probabilidades de inclusión y las cantidades $-\Delta_{ij}$, así como la varianza del estimador de Horvitz-Thompson para el total poblacional, $T(Y)$.

Teorema 1

Con el procedimiento previamente descrito se verifica:

(a)

$$\pi_i = \frac{X_i}{T_X(U_k)} \qquad i \in U_k$$

(b)

$$-\Delta_{ij} = \begin{cases} 0 & i \in U_k,\ j \in U_l,\ k \neq l \\ \pi_i \pi_j & i, j \in U_k,\ i \neq j \\ -\pi_i (1 - \pi_i) & i, j \in U_k,\ i = j \end{cases}$$

(c) La varianza del estimador de Horvitz-Thompson para el parámetro $T(Y)$ viene dada por

$$V[\hat{T}(Y)] = \frac{1}{2} \sum_{k=1}^{n} \sum_{i,j \in U_k} \pi_i \pi_i \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

Demostración

(a) Sea $u_i \in U_k$. Se tiene

$$\pi_i = P[i \in m] = P\left[T_{i-1} - \sum_{j=1}^{k-1} T_X(U_j) \leq r\, T_X(U_k) < T_i - \sum_{j=1}^{k-1} T_X(U_j) \right] =$$

$$= P\left[\frac{T_{i-1}}{T_X(U_k)} - \frac{\sum_{j=1}^{k-1} T_X(U_j)}{T_X(U_k)} \leq r < \frac{T_i}{T_X(U_k)} - \frac{\sum_{j=1}^{k-1} T_X(U_j)}{T_X(U_k)} \right] =$$

$$= \frac{T_i - T_{i-1}}{T_X(U_k)} = \frac{X_i}{T_X(U_k)}$$

(b) Es inmediato si se observa que

$$\pi_{ij} = \begin{cases} \pi_i \pi_j & i \in U_k, j \in U_l, k \neq l \\ 0 & i, j \in U_k, i \neq j \\ \pi_i & i, j \in U_k, i = j \end{cases}$$

(c) Basta sustituir los parámetros obtenidos en el apartado **(b)** en la expresión general

$$V[\hat{T}(Y)] = \frac{1}{2} \sum_{i,j \in U} \sum - \Delta_{ij} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

y al anularse los sumandos correspondientes a parejas de elementos de grupos diferentes, queda la expresión propuesta.

Observemos que el método propuesto no se puede considerar ΠPS en sentido estricto ya que el factor de proporcionalidad para las probabilidades de inclusión de primer orden, en relación al tamaño de cada elemento, es diferente en cada una de las subpoblaciones U_k , $k = 1, \ldots, n$. No obstante, ello no afecta a la disminución de varianza debido a la descomposición de la misma obtenida en el apartado **(c)**.

Recalcamos que para elementos de la población distantes, en el sentido de su tamaño con respecto a la variable auxiliar, X, el valor de $- \Delta_{ij}$ se anula, con lo cual el término correspondiente a dicho par de elementos desaparece de la expresión de la varianza.

Mediante un cálculo directo también podemos obtener la siguiente expresión alternativa para la varianza:

$$V [\hat{T} (Y)] = \sum_{i \in U} \frac{Y_i^2}{\pi_i} - \sum_{k=1}^{n} T_Y^2 (U_k)$$

donde denotamos

$$T_Y (U_k) = \sum_{i \in U_k} Y_i \qquad k = 1, \ldots, n$$

Como puede verse, bajo la hipótesis ya mencionada de que los tamaños de los elementos sean pequeños en relación a $T (X) / n$, se tendrá que

$$\pi_i = \frac{X_i}{T_X (U_k)} = n \frac{X_i}{T (X)}$$

con lo cual el primer término de la varianza no resulta muy influenciado por la estructura de ordenación existente sobre la población. Con respecto al segundo término, si denotamos por $N_1, \ldots, N_n$ los tamaños respectivos de $U_1, \ldots, U_n$, tenemos para su esperanza en el modelo de superpoblación

$$E_s \left[\sum_{k=1}^{n} T_Y^2 (U_k) \right] = \sum_{k=1}^{n} V_s [T_Y (U_k)] + E_s^2 [T_Y (U_k)] =$$

$$= \sum_{k=1}^{n} \sigma^2 N_k + (\alpha N_k + \beta T_X (U_k))^2 =$$

$$= \sigma^2 N + \sum_{k=1}^{n} (\alpha N_k + \beta T_X (U_k))^2 -$$

$$- \frac{1}{n} (\alpha N + \beta T (X))^2 + \frac{1}{n} (\alpha N + \beta T (X))^2 =$$

$$= \sigma^2 N + \frac{1}{n} (\alpha N + \beta T (X))^2 + n \sigma^2_{\alpha N_k + \beta T_X (U_k)}$$

donde $\sigma^2_{\alpha N_k + \beta T_X (U_k)}$ representa la varianza de las cantidades $\alpha N_k + \beta T_X (U_k)$, $k = 1, \ldots, n$.

Bajo la hipótesis mencionada, las cantidades $T_X (U_k)$, $k = 1, \ldots, n$ son muy similares, con lo cual, despreciando su variabilidad, se tiene

$$\sigma^2_{\alpha N_k + \beta T_X (U_k)} \simeq \alpha^2 \sigma^2_{N_k}$$

siendo $\sigma^2_{N_k}$ la varianza de las cantidades $N_1, N_2, \ldots, N_n$.

Así pues, la varianza de la estimación será tanto menor cuanto más dispersos estén los valores $N_1, N_2, \ldots, N_n$ y dicha dispersión tenderá a ser elevada cuando la población se ordene por la variable usada para equilibrar las subpoblaciones.

En la sección 5 de este trabajo exponemos una versión algorítmica del método descrito anteriormente.

3. COMPARACION CON OTROS METODOS

En esta sección realizamos una comparación con otros métodos estrictamente ΠPS, exponiendo los resultados obtenidos al aplicar tanto el método secuencial cuasi equilibrado (SCE) como el ya mencionado de Madow, y el de Sampford, de naturaleza no secuencial. Esta aplicación consiste en la estimación de la media poblacional de una variable de estudio, Y, en dos poblaciones generadas por simulación. Ambas poblaciones tienen un tamaño $N = 10000$ y se describen a continuación.

- **Población NOR10000**

Contiene dos variables, Y y X, ligadas por la relación

$$Y_i = 3000 + 8 \times X_i + 10 \times \varepsilon_i, \qquad i = 1, \ldots, 10000$$

Los valores de X_i han sido generados de una distribución normal N (1000, 200) y los de ε_i de una normal N (0, 50) independiente de la anterior.

- **Población EXP10000**

En esta población las variables Y y X están ligadas por la relación

$$Y_i = 1000 + 10 \times X_i + 3 \times \varepsilon_i, \qquad i = 1, \ldots, 10000$$

Los valores de X_i han sido generados de una distribución exponencial de media $\mu = 200$ y los de ε_i se han obtenido restando 50 a los valores generados a partir de una distribución exponencial de media 50.

El tamaño muestral usado es $n = 10$ y para cada estimación se ha calculado el error relativo mediante la expresión

$$\text{ERROR RELATIVO} = 100 \times \left| \frac{\bar{Y} - \hat{\bar{Y}}}{\bar{Y}} \right|$$

promediando dicho error relativo sobre 1000 ejecuciones de cada método. Además, se han diferenciado los resultados según que la población esté ordenada o no por la variable auxiliar X, con el fin de estudiar la sensibilidad de los métodos a la existencia de dicho orden.

- **Población NOR10000 ordenada por la variable auxiliar**

METODO	ERROR RELATIVO MEDIO
SCE	1.2410
Sampford	1.8964
Madow	1.4369

- **Población NOR10000 ordenada aleatoriamente**

METODO	ERROR RELATIVO MEDIO
SCE	1.8547
Sampford	1.9033
Madow	1.8471

- **Población EXP10000 ordenada por la variable auxiliar**

METODO	ERROR RELATIVO MEDIO
SCE	6.6579
Sampford	10.5136
Madow	8.5408

- **Población EXP10000 ordenada aleatoriamente**

METODO	ERROR RELATIVO MEDIO
SCE	10.3961
Sampford	10.3519
Madow	12.6710

Los valores obtenidos indican que el método secuencial cuasi equilibrado proporciona mejores resultados que los métodos de Sampford y Madow, siendo esta mejoría notable con respecto al de Sampford. No obstante, si la población se ordena aleatoriamente, los resultados son similares en los tres métodos.

4. ESTIMACION DEL ERROR DE MUESTREO

Como puede observarse, al no ser mayores que cero todas las probabilidades de inclusión de segundo orden, no es posible aplicar el estimador usual de la varianza.

A continuación veremos cómo el uso de técnicas de replicación puede proporcionarnos estimaciones insesgadas de la varianza. Supongamos que por el procedimiento descrito en la sección anterior hemos obtenido l muestras independientes, por ejemplo reiterando el proceso l veces sobre la población,

$$m_1, m_2, \ldots, m_l$$

y denotemos

$$\hat{T}_j(Y) = \sum_{m_j} \frac{Y_i}{\pi_i} \qquad j = 1, \ldots, l$$

y sea

$$\hat{T}^*(Y) = \frac{1}{l} \sum_{j=1}^{l} \hat{T}_j(Y)$$

Por otra parte, denotemos

$$m_1^*, m_2^*, \ldots, m_n^*$$

a las n muestras formadas cada una, respectivamente, con los l elementos seleccionados en cada grupo U_k. Se verifica entonces el siguiente teorema.

Teorema 2

(a) $\hat{T}^*(Y)$ es un estimador insesgado de $T(Y)$.

(b) Su varianza es

$$V[\hat{T}^*(Y)] = \frac{1}{2l} \sum_{k=1}^{n} \sum_{i,j \in U_k} \pi_i \pi_i \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

(c) Un estimador insesgado de $V[\hat{T}^*(Y)]$ es

$$\hat{V}[\hat{T}^*(Y)] = \frac{1}{2l^2(l-1)} \sum_{k=1}^{n} \sum_{i,j \in m_k^*} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

Demostración

(a) Inmediato por ser insesgado el estimador de Horvitz-Thompson.

(b) Se deduce directamente a partir de la independencia de los estimadores

$$\hat{T}_j(Y) = \sum_{m_j} \frac{Y_i}{\pi_i} \qquad j = 1, \ldots, I$$

(c) Para demostrar que

$$\hat{V}[\hat{T}^*(Y)] = \frac{1}{2I^2(I-1)} \sum_{k=1}^{n} \sum_{i,j \in m_k^*} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

es insesgado expresamos $\hat{T}^*(Y)$ de la siguiente forma:

$$\hat{T}^*(Y) = \frac{1}{I} \sum_{j=1}^{I} \hat{T}_j(Y) = \frac{1}{I} \sum_{k=1}^{n} \sum_{i \in m_k^*} \frac{Y_i}{\pi_i} = \sum_{k=1}^{n} \sum_{i \in m_k^*} \frac{Y_i}{I\pi_i}$$

Observemos que m_k^* es una muestra de I elementos, con posibles repeticiones, extraídos de U_k, con probabilidades de selección que coinciden con las de inclusión. Se tiene, pues, que

$$\sum_{i \in m_k^*} \frac{Y_i}{I\pi_i}$$

es el estimador de Hansen-Hurwitz de $T(U_k) = \sum_{U_k} Y_i$. Luego, por la independencia,

$$\hat{V}[\hat{T}^*(Y)] = \sum_{k=1}^{n} V \left[\sum_{i \in m_k^*} \frac{Y_i}{I\pi_i} \right]$$

y aplicando la estimación insesgada usual para la varianza de dicho estimador, obtenemos

$$\hat{V}[\hat{T}^*(Y)] = \sum_{k=1}^{n} \hat{V}\left[\sum_{i \in m_k^*} \frac{Y_i}{I\pi_i}\right] = \sum_{k=1}^{n} \frac{1}{I} s_z^2(m_k^*) =$$

$$= \frac{1}{2I^2(I-1)} \sum_{k=1}^{n} \sum_{i,j \in m_k^*} \sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

donde con $s_z^2(m_k^*)$ hemos denotado la cuasivarianza, sobre los elementos de la muestra m_k^*, de los valores de la variable $Z_i = Y_i / \pi_i$, habiéndose aplicado la siguiente expresión alternativa de la cuasivarianza:

$$s_z^2(m_k^*) = \frac{1}{I-1} \sum_{i \in m_k^*} \left(\frac{Y_i}{\pi_i} - \frac{1}{I} \sum_{i \in m_k^*} \frac{Y_i}{\pi_i}\right)^2 =$$

$$= \frac{1}{2I(I-1)} \sum_{i,j \in m_k^*} \sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

Observemos que para estimar $\hat{V}[\hat{T}^*]$ también es posible usar el estimador genérico en las técnicas de agrupaciones aleatorias, es decir,

$$\hat{V}^*[\hat{T}^*] = \frac{1}{I(I-1)} \sum_{j=1}^{I} (\hat{T}_j(Y) - \hat{T}^*(Y))^2$$

que, en nuestro caso, también es insesgado. En efecto, puede demostrarse [véase Fernández y Mayor (1994)] la siguiente relación:

$$E[\hat{V}^*[\hat{T}^*(Y)]] = E\left[\frac{1}{I(I-1)} \sum_{j=1}^{I} (\hat{T}_j(Y) - \hat{T}^*(Y))^2\right] =$$

$$= V[\hat{T}^*(Y)] - \frac{1}{I(I-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{I} \sum \text{Cov}[\hat{T}_i(Y), \hat{T}_j(Y)] +$$

$$+ \frac{1}{I(I-1)} \sum_{j=1}^{I} (E[\hat{T}_j(Y)] - E[\hat{T}^*(Y)])^2$$

Por la independencia de $\hat{T}_j\ (Y)$, $j = 1, \ldots, l$ se verifica que las covarianzas entre ellos son nulas, es decir,

$$\text{Cov}\ [\hat{T}_i\ (Y), \hat{T}_j\ (Y)] = 0 \qquad \forall\ i \neq j$$

Y, por otra parte, al ser insesgado el estimador de Horvitz-Thompson, se tiene

$$E\ [\hat{T}_j\ (Y)] = T\ (Y) \qquad \forall\ j$$

de donde se deduce

$$E\ [\hat{V}^*\ [\hat{T}^*\ (Y)]] = V\ [\hat{T}^*\ (Y)]$$

En la práctica, para obtener las muestras replicadas, $m_1, \ldots, m_l$, no es necesario explorar la población l veces, basta realizar una única exploración secuencial, extrayendo las l muestras de forma paralela. En la sección 5 se indica, en forma algorítmica, la realización de dicho proceso en paralelo.

Dicho algoritmo ha sido aplicado a la estimación de la media poblacional de la variable Y en la población NOR10000 ordenada por la variable X, usando los estimadores propuestos en el teorema anterior, para lo cual se han extraído de la población 30 elementos, tomando $l = 3$ y $n = 10$. Este proceso se ha realizado 1000 veces, para las cuales se ha calculado el error relativo medio y el coeficiente de variación de las varianzas estimadas.

Los resultados se exponen en la siguiente tabla, comparándolos con los obtenidos para el método de Sampford con un tamaño muestral $n = 30$. Para cada método se indica, además, el valor de la varianza real de la estimación a efectos comparativos.

Para el método de Sampford, tanto las probabilidades de inclusión de segundo orden como la varianza real de la estimación han sido calculadas usando las correspondientes expresiones asintóticas dadas por Asok y Sukhatme (1976) [véase también Sukhatme *et al.* (1984)].

	SCE en paralelo ($l = 3$, $n = 10$)	Sampford ($n = 30$)
ERROR RELATIVO MEDIO	0.726471	1.130521
COEFICIENTE DE VARIACION DE LAS VARIANZAS ESTIMADAS	0.444047	0.436414
VARIANZA DE LA ESTIMACION	10080.497	22043.954

Como puede verse, el error relativo medio es notablemente más pequeño para el método secuencial replicado que para el de Sampford, siendo resaltable que el método de Sampford tenga una varianza superior al doble de la del método secuencial equilibrado. Finalmente, podemos también afirmar que la estabilidad de la estimación de la varianza es similar en ambos métodos, como puede deducirse de los coeficientes de variación de las varianzas estimadas.

5. ESQUEMAS ALGORITMICOS

En esta sección presentamos, en forma algorítmica, los diferentes métodos propuestos en este trabajo. En todos ellos, la notación $r \sim \mathcal{U}[0, 1)$ indica que r es una magnitud aleatoria generada a partir de una distribución uniforme continua en [0, 1).

5.1. Método de inserción

En este algoritmo, Θ representa una lista, que siempre se mantiene ordenada, de elementos de la población, usando como criterio de ordenación el indicado más adelante.

ALGORITMO 1. METODO DE INSERCION

$\Theta := (), i := 1$

REPETIR

$\alpha_i \sim \mathcal{U}[0, 1)$

SI $(i \leq n)$

Introducir u_i en la lista Θ de forma que los elementos de dicha lista aparezcan ordenados por la magnitud del número aleatorio asociado.

SI NO

SI $(\alpha_i \leq \max_{j \in \Theta} \{\alpha_j\} = \alpha_l)$

Eliminar u_l de Θ e insertar en dicha lista el elemento u_i.

FIN SI

FIN SI

$i := i + 1$

HASTA $(u_i \notin U)$

Una vez finalizado el algoritmo, la lista Θ contiene la muestra deseada. Este procedimiento asegura la obtención de una muestra perteneciente a un diseño muestral aleatorio, es decir, todas las muestras de tamaño fijo n con distribución de probabilidad uniforme sobre las mismas, a partir de una exploración secuencial de la población, sin requerir el conocimiento previo del tamaño de la población.

Para probarlo, basta observar que si ordenamos la población completa usando como criterio la magnitud de los números aleatorios, cualquiera de las $N!$ ordenaciones posibles tiene probabilidad $1 / N!$, por lo cual la probabilidad de obtener la muestra m será

$$p(m) = \frac{n!\,(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$$

5.2. Método secuencial cuasi equilibrado

Seguidamente exponemos una versión algorítmica del método de exploración secuencial cuasi equilibrado. En dicho algoritmo, m representa la muestra a extraer.

ALGORITMO 2. METODO DE EXPLORACION SECUENCIAL CUASI EQUILIBRADO

$m := \{\}$, $i := 0$, $k := 1$, $T_g := 0$
REPETIR
  $r \sim \mathcal{U}[0, 1)$
  $j := 0$
  REPETIR
    $i := i + 1$
  SI $((r\ T_X(U_k) + T_g < T_i)$ Y $(j = 0))$
    $m := m \cup \{i\}$
    $j := 1$
  FIN SI
  HASTA $\left(T_i \geq k \dfrac{T(X)}{n}\right)$
  $T_g := T_g + T_X(U_k)$
  $k := k + 1$
HASTA $(k > n)$

5.3. Método de exploración secuencial cuasi equilibrado en paralelo

El algoritmo siguiente obtiene l muestras independientes, cada una de n elementos, con un solo barrido de la población.

ALGORITMO 3. METODO DE EXPLORACION SECUENCIAL CUASI EQUILIBRADO EN PARALELO

$i := 0,\ k := 1,\ T_g = 0$

$m_1 := \{\}, \ldots, m_l := \{\}$

REPETIR

$r_1, \ldots, r_l \sim \mathcal{U}[0, 1)$

$j_1 := 0, \ldots, j_l := 0$

REPETIR

$i := i + 1$

$p := 1$

REPETIR

SI $((r_p\ T_X(U_k) + T_g < T_i)$ Y $(j_p = 0))$

$m_p := m_p \cup \{i\}$

$j_p := 1$

FIN SI

$p := p + 1$

HASTA $p > l$

HASTA $\left(T_i \geq k \dfrac{T(X)}{n}\right)$

$T_g := T_g + T_X(U_k)$

$k := k + 1$

HASTA $(k > n)$

6. CONCLUSIONES

Como se ha visto, el método secuencial cuasi equilibrado, tanto en su versión básica como replicada, se presenta como una alternativa muy adecuada en aquellos casos en los cuales las particularidades de la población sólo permiten una exploración secuencial de la misma, sacando partido de la ordenación existente en relación a una variable auxiliar. Es importante observar que tal ordenación no tiene por qué ser estricta, de manera que la existencia de una ordenación aproximada también producirá una disminución del error de muestreo.

Por otra parte, dicho método, en comparación con otros de tipo secuencial de similares características como el de Madow, presenta la ventaja de disponer de un estimador de la varianza adaptado a sus particularidades.

REFERENCIAS

ASOK, C., y SUKHATME, B. V. (1976): «On Sampford's procedures of unequal probability sampling without replacement», *J. Amer. Satist. Assoc,* 71, 912-918.

BELLHOUSE, D. R. (1988): «Systematic Sampling», *Handbook of Statistics 6. Sampling,* P. R. Krishnaiah y C. R. Rao (eds.), Amsterdam: North-Holland.

FERNÁNDEZ, F. R., y MAYOR, J. A. (1994): *Muestreo en poblaciones finitas: Curso básico,* Barcelona: P.P.U.

HÁJEK, J. (1981): *Sampling from a Finite Population,* New York: Marcel Dekker.

MADOW, W. G. (1949): «On the theory of systematic sampling II», *Ann. Math. Statist.,* 20, 333-354.

SAMPFORD, M. R. (1967): «On sampling without replacement with unequal probabilities of selection», *Biometrika,* 54, 499-513.

SÄRNDAL, C.; SWENSSON, B., y WRETMAN, J. (1992): *Model Assisted Survey Sampling,* New York: Springer-Verlag.

SUKHATME, P. V., *et al.* (1984): *Sampling Theory of Surveys Applications,* Iowa State University Press.

SUNTER, A. B. (1977): «List Sequential Sampling with Equal or Unequal Probabilities without Replacement», *Appl. Statist.,* 26, N. 3, 261-268.

TILLÉ, I. (1994): «L'algorithme de la strate mobile», XXI Congreso Nacional de Estadística, Investigación Operativa e Informática, S.E.I.O., Calella, Barcelona.

A SAMPLING SCHEME WITH SEQUENTIAL QUASI BALANCED EXPLORATION

SUMMARY

In this paper we propose a sampling sequential procedure with first-order inclusion probabilities wich are proportional to size over subpopulations from the entire population, and second-order inclusion probabilities equal to zero for pairs of similar units. We show that this method reduces the sampling error with respect to similar methodologies.

A replication method is used to estimate the variance. Its main advantage is that all the sampling replications are obtained with only one sequential exploration of the population.

The relative performance of this method, in relation with another unequal probability sampling schemes, either sequential-type or not, is studied. For that, using simulated populations, the relative errors are calculated under different sampling schemes.

Key Words: sampling with unequal probabilities, sequential sampling, Horvitz-Thompson estimator, superpopulation model, replicated scheme.

AMS Classification: 62D05.