Una metodología para el control por muestreo de cuentas económicas

MARIANO RUIZ ESPEJO

Universidad Nacional de Educación a Distancia

RESUMEN

Presentamos una metodología estadística para el control por muestreo de cuentas económicas. Asumimos el modelo de población finita fija para la estimación insesgada en auditorías, contabilidad y estimación de macromagnitudes como la renta nacional, el total defraudado o la proporción de unidades defraudadoras. El control infiere sobre tales parámetros con no respuesta, errores de medida y en algunos casos tamaños desconocidos de los estratos de poblaciones oblicuas de tipo económico.

Palabras clave: Control por muestreo, Cuentas económicas.

Clasificación A.M.S.: 62D05.
Clasificación UNESCO: 5303, 1209.

ABSTRACT

We present a statistical methodology for the control by sampling of economic accounts. We assume the fixed population model for the unbiased estimation in auditing, accounting and estimation of macromagnitudes as the gross national product, the disappointed total or the proportion of units with fraud. The control infers on such parameters with nonresponse, measure errors and in some cases strata unknown sizes of skewed populations of economic kind.

Key words: Control by sampling, Economic accounts.

A.M.S. classification: 62D05.
UNESCO classification: 5303, 1209.

Dad al César lo que es del César...

1. Introducción

En este trabajo se presenta una metodología basada en técnicas de "muestreo probabilístico de poblaciones finitas", mppf, que garantizan imparcialidad dentro de cada estrato en la selección de un subconjunto de unidades a observar y, en su caso inspeccionar algunas unidades también por mppf, de un conjunto de unidades contables, datos económicos que contribuyen como un sumando más a uno o varios parámetros medidos en "unidades monetarias", u.m., u otras que interesan ser conocidos por las autoridades y expertos sin necesidad de un estudio y reconocimiento exhaustivo de cada una de tales unidades.

Este método muestral representa una sustancial ganancia en tiempo (para disponer de información precisa), ahorro de recursos (en medios de comunicación y económicos) y de trabajo humano, necesarios para una recogida de datos y recuento censal o exhaustivo de los mismos, aunque con la incorporación de medios tecnológicos informáticos para su procesamiento.

Algunos trabajos recientes en este tipo de investigaciones son los de Armstrong *et al.* (1993), Fuller (1991), Lavallée *et al.* (1988), Pechman (1990) y Press (1995), que usan técnicas de muestreo en economía, auditoría, hacienda, etc. En este contexto se sitúa la contribución metodológica que presentamos mediante muestreo bifásico con estratificación en la segunda fase y asumiendo poblaciones finitas oblicuas (*skewed*), cuando se involucran errores de declaración, no respuesta o datos faltantes, e incluso desconociendo en algunos casos los tamaños de los estratos. En estas condiciones se aportan estimadores insesgados del total económico declarado, del total económico verdadero (y de la renta per cápita), del total defraudado, estimando insesgadamente la varianza de los estimadores empleados, incluyendo estimaciones por intervalo de la proporción de unidades que defraudan con un nivel de confianza acotado inferiormente por $1 - \alpha$.

2. Control por muestreo de cuentas económicas

Desde el punto de vista estadístico existen diversas maneras de abordar el estudio de cuentas económicas en las que el "total de unidades monetarias", es el resultado de una suma de un gran número de términos correspondientes a partidas contables, personas físicas o incluso personas jurídicas, que contribuyen con sumandos fijos en las mismas u.m. al total que interesa conocer.

El planteamiento que sugerimos en este trabajo consiste en conocer una población finita o "conjunto de unidades identificadas" que denotamos por U (universo),

$$U = \{1, 2, ..., i, ..., N\},$$

siendo N el tamaño de la población finita, o número de unidades en U.

El identificador i lleva asociado una unidad contable: una factura, una persona física accesible, una empresa, un capital, una tierra, etc. que contribuyen con una variable económica verdadera x_i' al *total económico verdadero* en u.m.

$$X' = \sum_{i=1}^{N} x_i' = N\bar{X}'.$$

Sin embargo en la práctica la variable económica declarada o reflejada en la información inicial, por la unidad i, es el valor x_i. El *total declarado* en u.m. será entonces

$$X = \sum_{i=1}^{N} x_i = N\bar{X}.$$

El *fraude total, F*, será entonces el *total verdadero* X' menos el *total declarado* X:

$$F = X' - X = \sum_{i=1}^{N} (x_i' - x_i) = \sum_{i=1}^{N} t_i = N\bar{F}.$$

El *fraude para la unidad* i será $t_i = x_i' - x_i$, cuando esta diferencia sea *no nula y vulnere* la normativa legal correspondiente a cada caso. En tal caso, la unidad i defrauda. La *proporción de defraudadores* de la población U será:

$$\bar{D} = \frac{1}{N} \sum_{i=1}^{N} d_i,$$

en donde

$$d_i = \begin{cases} 1 & \text{si la unidad } i \text{ sí defrauda} \\ 0 & \text{si la unidad } i \text{ no defrauda} \end{cases}$$

y el *total de unidades defraudadoras* en la población U es D:

$$D = \sum_{i=1}^{N} d_i = N\bar{D}.$$

El problema económico-estadístico a resolver consiste en que conociendo el tamaño de la población, N:

1. Estimar sin sesgo F, X' y D, ya que no se dispone de recursos para obtener todos los datos x_i' y x_i $(i = 1, 2, ..., N)$ al ser N un número extraordinariamente grande (como en la renta nacional) o muy grande (como en auditorías).
2. Estimar sin sesgo las varianzas de los estimadores usados en el punto 1.

Otro problema es:

3. Estimar por intervalo la proporción de unidades defraudantes, $\bar{D}$.

Cuando para toda unidad $i \in U$, $t_i = x_i' - x_i = 0$ (sin fraude en todas las unidades i de la población U), entonces $F = D = 0$ y $X' = X$. En tales hipótesis de trabajo, puede usarse la técnica de muestreo estratificado especial introducido por Dalenius (1957) en poblaciones infinitas, y que puede utilizarse también para poblaciones finitas. Esta estratificación especial consiste en que la población se divide en dos estratos de tamaños conocidos N_1 y N_2, con $N = N_1 + N_2$. El estrato 1 contiene las unidades pequeñas, que se muestrean en un número n_1 $(2 \leq n_1 < N)$; mientras que el estrato 2 recoge a las unidades grandes y que contribuyen usualmente más a los totales X', X y F, realizando un censo en este estrato $(n_2 = N_2)$ coincidiendo el tamaño muestral en el segundo estrato, n_2, con el propio tamaño del estrato, N_2. El estimador a usar para estimar el total X cuando existen unidades grandes, es la generalización del citado en el problema de "¡Oh! Calcuta" por Rao (1994, pp. 106-107) adaptado cuando $N_2 > 1$:

$$\hat{X} = N_1\bar{x}_1 + X_2,$$

donde $\bar{x}_1$ es la media muestral en el estrato primero con diseño mas-n_1 (muestreo aleatorio simple sin reemplazamiento de tamaño n_1) y X_2 es el total del estrato

segundo. Su varianza es

$$V\left(\hat{X}\right) = N_1^2 V\left(\bar{x}_1\right),$$

y su estimador insesgado es

$$\hat{V}\left(\hat{X}\right) = N_1^2 \hat{V}\left(\bar{x}_1\right),$$

donde $\hat{V}\left(\bar{x}_1\right)$ es un estimador insesgado del parámetro $V\left(\bar{x}_1\right)$. De hecho, este método requiere conocer N y N_1 ($N_2 = N - N_1$), y supone la inexistencia de fraudes y de no respuesta (o datos faltantes o *missing data*), algo que puede ser una hipótesis de trabajo aceptable en ciencias exactas, físicas, naturales, etc. pero difícilmente cierto en ciencias económicas, empresariales y sociales.

Ya el profesor Glasser (1962) utilizó el estimador correspondiente de la media poblacional, $\bar{X}$, en poblaciones finitas del tipo económico que aquí estudiamos

$$\widehat{\bar{X}} = \frac{\hat{X}}{N} = W_1\bar{x}_1 + W_2\bar{X}_2,$$

donde $W_i = N_i/N$, $\bar{X}_2 = X_2/N_2$ y además investigó condiciones para la estratificación óptima, supuesta la ordenación de los datos de menor a mayor. En Ruiz Espejo (1996) se presentan algunas mejoras en su análisis.

Admitiendo que usamos muestreo estratificado especial como diseño muestral de selección de unidades, y que pueden existir "errores o fraudes" y "no respuesta o datos faltantes", un estimador insesgado de la media poblacional verdadera, $\bar{X}'$, es

$$\bar{x}'_{st} = W_1\bar{x}'_1 + W_2\bar{x}'_2,$$

donde $\bar{x}'_1$ y $\bar{x}'_2$ son los estimadores insesgados de los parámetros $\bar{X}'_1$ y $\bar{X}'_2$ estudiados en Ruiz Espejo (1988, p. 77; 1994), quedando resuelto el problema de la estimación insesgada de la varianza de $\bar{x}'_{st}$. El estimador $\bar{x}'_1$ (y $\bar{x}'_2$) utilizan a su vez muestreo bifásico. En primera fase se selecciona una muestra en el estrato 1 (y 2) de tamaño $\bar{n}_1$ (y $\bar{n}_2 = N_2$); $\bar{n}_1 = \bar{n}_{11} + \bar{n}_{12}$ ($\bar{n}_2 = \bar{n}_{12} + \bar{n}_{22}$) siendo $\bar{n}_{11}$ (y $\bar{n}_{21}$) los que responden y $\bar{n}_{12}$ (y $\bar{n}_{22}$) los que no responden. El total declarado u observado de estas respuestas no necesariamente verdaderas, es x_{1s1} (y x_{2s1}). Seguidamente en la segunda fase, de los $\bar{n}_{11}$ (y $\bar{n}_{21}$) que responden se obtiene una submuestra de tamaño $\bar{n}_{111} < \bar{n}_{11}$ (y $\bar{n}_{211} < \bar{n}_{21}$) donde se estima e_{111} (y e_{211}) o total muestral del error de observación respecto de los valores verdaderos obtenidos por inspección.

De las $\bar{n}_{12}$ (y $\bar{n}_{22}$) unidades que no responden, de las respectivas muestras, se obtiene una submuestra de tamaño $\bar{n}_{121}$ (y $\bar{n}_{221}$) empleando inspección y obteniendo los valores verdaderos y cuyo total muestral será x'_{121} (y x'_{221}). Con este diseño muestral, el estimador insesgado $\bar{x}'_1$ (y $\bar{x}'_2$) es:

$$\bar{x}'_1 = \frac{1}{\bar{n}_1}\left(x_{1s1} - \frac{\bar{n}_{11}}{\bar{n}_{111}}e_{111} + \frac{\bar{n}_{12}}{\bar{n}_{121}}x'_{121}\right)$$

$$\left[\text{y} \quad \bar{x}'_2 = \frac{1}{\bar{n}_2}\left(x_{2s1} - \frac{\bar{n}_{21}}{\bar{n}_{211}}e_{211} + \frac{\bar{n}_{22}}{\bar{n}_{221}}x'_{221}\right)\right]$$

El muestreo estratificado especial, inicialmente considerado para 2 estratos puede ser generalizado al caso de un número de estratos $L \geq 2$, donde en los $L-1$ primeros se realiza un muestreo estratificado y en el L-ésimo se realiza el censo a las unidades grandes.

Basándonos en estos estudios conocidos, en lo que sigue vamos a dar cobertura metodológica para las cuestiones 1 y 2 planteadas en el control en *auditorías*, impuestos de *hacienda* y de cómo integrar ésta última en el estudio y control de la *renta nacional*. En la recaudación de hacienda, también podemos resolver la cuestión 3.

3. Control por muestreo en auditorías

En el problema de estimar en auditorías el total verdadero X', pueden detectarse un número desconocido o no N_2 de unidades grandes, permaneciendo desconocido o no el tamaño N_1 del primer estrato. La estimación estratificada especial se obtiene dividiendo la población U en dos estratos según las unidades presenten una característica y_k menor o mayor que el punto de estratificación y, para la variable auxiliar que puede observarse en una muestra piloto, s_p, de tamaño m. Si el diseño muestral piloto es de masr-m (muestreo aleatorio simple con reemplazamiento de tamaño m) sobre la población U, denotamos por m_1 al número de selecciones de la muestra piloto s_p cuya variable auxiliar sea menor o igual a y (que es una variable aleatoria binomial previamente a la selección muestral),

$$m_1 = card\{i \in U:\ i \in s_p \text{ e } y_i \leq y\} \rightarrow B(m, W_1)$$

con

$$m_2 = m - m_1 \rightarrow B(m, W_2),$$

denotando para $i = 1, 2$,

$$w_i = m_i/m.$$

En la etapa de observación, posterior a la etapa piloto, el estimador recomendado para $\bar{X}'$ es

$$\bar{x}''_{st} = w_1\bar{x}'_1 + w_2\bar{X}'_2,$$

en el que $\bar{x}'_1$ es el estimador propuesto en Ruiz Espejo (1988, 1994) que admite, como ya hemos indicado, un estimador insesgado de la varianza $\hat{V}_2(\bar{x}'_1)$ al referirse el subíndice de $\hat{V}_2$ a la segunda etapa (posterior a la etapa piloto o primera etapa).

Los tamaños relativos de los estratos W_1 y W_2 son desconocidos. Denotando por $\mathbf{m} = (m_1, m_2)$ a la variable aleatoria multinomial, la esperanza matemática del estimador $\bar{x}''_{st}$ es:

$$E(\bar{x}''_{st}) = E_2E_1\left(w_1\bar{x}'_1 + w_2\bar{X}'_2|\mathbf{m}\right) =$$

$$= E_2\left(W_1\bar{x}'_1 + W_2\bar{X}'_2\right) = W_1\bar{X}'_1 + W_2\bar{X}'_2 = \bar{X}',$$

y la varianza de $\bar{x}''_{st}$ es:

$$V(\bar{x}''_{st}) = V_2E_1(\bar{x}''_{st}|\mathbf{m}) + E_2V_1(\bar{x}''_{st}|\mathbf{m}) =$$

$$= V_2\left(W_1\bar{x}'_1 + W_2\bar{X}'_2\right) +$$

$$+E_2\left[\bar{x}'^2_1V_1(w_1|\mathbf{m}) + \bar{X}'^2_2V_1(w_2|\mathbf{m}) + 2\bar{x}'_1\bar{X}'_2Cov_1(w_1, w_2|\mathbf{m})\right] = (*)$$

y como

$$V_1(w_1|\mathbf{m}) = V_1(w_2|\mathbf{m}) = -Cov_1(w_1, w_2|\mathbf{m}) = \frac{W_1W_2}{m},$$

$$(*) = W_1^2V_2(\bar{x}'_1) + \frac{W_1W_2}{m}\left\{\left[\bar{X}'^2_1 + V_2(\bar{x}'_1)\right] + \bar{X}'^2_2 - 2\bar{X}'_1\bar{X}'_2\right\} =$$

$$= W_1\left(W_1 + \frac{W_2}{m}\right)V_2(\bar{x}'_1) + \frac{W_1W_2}{m}\left(\bar{X}'_1 - \bar{X}'_2\right)^2.$$

Aunque estos resultados son ciertos, es más eficiente (o preciso, en el sentido de "menor varianza") usar un diseño muestral piloto tipo mas-m (muestreo aleatorio simple sin reemplazamiento de tamaño m), en cuyo caso $\mathbf{m} = (m_1, m_2)$ es una

variable aleatoria hipergeométrica de parámetros (m, N, W_i), véase Casas Sánchez y Santos Peñas (1996), con

$$E_1(w_1|\mathbf{m}) = W_1, \quad E_1(w_2|\mathbf{m}) = W_2,$$

y

$$V_1(w_1|\mathbf{m}) = V_1(w_2|\mathbf{m}) = -Cov_1(w_1, w_2|\mathbf{m}) = \frac{N-m}{(N-1)m} W_1 W_2,$$

y de aquí,

$$E(\bar{x}''_{st}) = \bar{X}'$$

y

$$V(\bar{x}''_{st}) = W_1^2 V_2(\bar{x}'_1) +$$

$$+\frac{N-m}{(N-1)m} W_1 W_2 \left\{ \left[\bar{X}_1'^2 + V_2(\bar{x}'_1) \right] + \bar{X}_2'^2 - 2\bar{X}'_1 \bar{X}'_2 \right\} =$$

$$= W_1 \left[W_1 + \frac{N-m}{(N-1)m} W_2 \right] V_2(\bar{x}'_1) + \frac{N-m}{(N-1)m} W_1 W_2 \left(\bar{X}'_1 - X'_2 \right)^2.$$

Denotando ahora con signo superior circunflejo a los estimadores insesgados del parámetro que subindican,

$$\hat{V}_2(\bar{x}'_1) = \frac{N_1 - \bar{n}_1}{N_1 \bar{n}_1} s^2_{x'_{\bar{n}_{111}+\bar{n}_{121}}} +$$

$$+\frac{1}{\bar{n}_1^2} \left[\left(\bar{n}_1^2 \frac{N_{11}^2}{N_1^2} + \frac{N_1 - \bar{n}_1}{N_1 - 1} \bar{n}_1 \frac{N_{11} N_{12}}{N_1^2} \right) (k-1) s^2_{e_{n_{111}}} + \right.$$

$$\left. + \left(\bar{n}_1^2 \frac{N_{12}^2}{N_1^2} + \frac{N_1 - \bar{n}_1}{N_1 - 1} \bar{n}_1 \frac{N_{11} N_{12}}{N_1^2} \right) (k-1) s^2_{x'_{n_{121}}} \right]$$

que está estudiado en Ruiz Espejo (1994), donde $s^2_{x'_{\bar{n}_{111}+\bar{n}_{121}}}$ es la cuasivarianza muestral de las $\bar{n}_{111} + \bar{n}_{121}$ primeras unidades de la muestra, N_{11} (resp. N_{12}) es el tamaño del subestrato de respuesta (resp. de no respuesta) en el primer estrato de unidades pequeñas, $s^2_{e_{\bar{n}_{111}}}$ es la cuasivarianza muestral de errores de las $\bar{n}_{111}$ primeras unidades de respuesta en la primera fase de selección y que se inspeccionan en la segunda fase, $s^2_{x'_{\bar{n}_{121}}}$ es la cuasivarianza muestral de las $\bar{n}_{121}$ primeras unidades de la muestra en la segunda fase de la variable verdadera x' entre las $\bar{n}_{12}$ que no respondieron en la primera fase, y $k = \bar{n}_{11}/\bar{n}_{111} = \bar{n}_{12}/\bar{n}_{121} =$

$\bar{n}_{21}/\bar{n}_{211} = \bar{n}_{22}/\bar{n}_{221}$ es la constante común de submuestreo de una fase a otra ($k = k_{11} = k_{21}$ de Ruiz Espejo, 1988, pp. 79-80); en la última fórmula se consideran conocidas las constantes que intervienen y, en otro caso, las constantes pueden sustituirse por sus estimadores insesgados respectivos dando lugar a estimadores $\hat{V}_2(\bar{x}'_1)$ aproximadamente insesgados. También,

$$\left[\widehat{\left(\bar{X}'_1 - \bar{X}'_2\right)^2}\right] = \left(\bar{X}'^2_1 - \widehat{2\bar{X}'_1\bar{X}'_2}\right) + \bar{X}'^2_2 =$$

$$= \bar{x}'^2_1 - \hat{V}_2(\bar{x}'_1) - 2\bar{x}'_1\bar{X}'_2 + \bar{X}'^2_2 =$$

$$= \left(\bar{x}'_1 - \bar{X}'_2\right)^2 - \hat{V}_2(\bar{x}'_1),$$

$$\widehat{W_1^2} = w_1^2 - \hat{V}_1(w_1|\mathbf{m}) = w_1^2 - \frac{N-m}{(N-1)m}\widehat{[W_1(1-W_1)]} =$$

$$= w_1^2 - \frac{N-m}{N(m-1)}w_1(1-w_1)$$

y

$$\widehat{(W_1W_2)} = \frac{(N-1)m}{N(m-1)}w_1(1-w_1),$$

por lo que sustituyendo estos estimadores insesgados, podemos obtener un estimador insesgado de la varianza $V(\bar{x}''_{st})$ como

$$\hat{V}(\bar{x}''_{st}) = \left[\widehat{(W_1^2)} + \frac{N-m}{(N-1)m}\widehat{(W_1W_2)}\right]\hat{V}_2(\bar{x}'_1) +$$

$$+\frac{N-m}{(N-1)m}\widehat{(W_1W_2)}\left[\widehat{\left(\bar{X}'_1 - \bar{X}'_2\right)^2}\right],$$

admitiendo conocido el tamaño poblacional N. De todo lo anterior, es directo estimar el total verdadero X', con el estimador $N\bar{x}''_{st}$, siendo un estimador insesgado de su varianza

$$\hat{V}(N\bar{x}''_{st}) = N^2\hat{V}(\bar{x}''_{st}).$$

Si todas las unidades de la muestra declaran (y supuesto que el "valor registrado de libro", X, es desconocido), el estimador insesgado del total declarado, X, es $N\bar{x}^*_{st}$, donde

$$\bar{x}^*_{st} = w_1\bar{x}_1 + w_2\bar{X}_2,$$

siendo $\bar{x}_1$ la media muestral declarada en el estrato de unidades pequeñas (estrato 1). Por ello, en el caso de ausencia de "datos faltantes o no respuesta" en la muestra, un estimador insesgado del total defraudado, es

$$\hat{F} = N\left(\bar{x}''_{st} - \bar{x}^{*}_{st}\right) = N\left[w_1\left(\bar{x}'_1 - \bar{x}_1\right) + w_2\left(\bar{X}'_2 - \bar{X}_2\right)\right],$$

y su varianza resulta ser

$$V\left(\hat{F}\right) = N_1^2 V_2\left(\bar{x}'_1 - \bar{x}_1\right) +$$

$$+N^2 E_2\left[\left(\bar{x}'_1 - \bar{x}_1\right)^2 V_1\left(w_1|\mathbf{m}\right) + \left(\bar{X}'_2 - \bar{X}_2\right)^2 V_1\left(w_2|\mathbf{m}\right) +\right.$$

$$\left. + 2\left(\bar{x}'_1 - \bar{x}_1\right)\left(\bar{X}'_2 - \bar{X}_2\right) Cov_1\left(w_1, w_2|\mathbf{m}\right)\right]$$

que puede estimarse insesgadamente por

$$\hat{V}\left(\hat{F}\right) = \widehat{N_1^2}\hat{V}_2\left(\bar{x}'_1 - \bar{x}_1\right) +$$

$$+N^2\left[\left(\bar{x}'_1 - \bar{x}_1\right)^2 \hat{V}_1\left(w_1|\mathbf{m}\right) + \left(\bar{X}'_2 - \bar{X}_2\right)^2 \hat{V}_1\left(w_2|\mathbf{m}\right) +\right.$$

$$\left. + 2\left(\bar{x}'_1 - \bar{x}_1\right)\left(\bar{X}'_2 - \bar{X}_2\right) \widehat{Cov}_1\left(w_1, w_2|\mathbf{m}\right)\right]$$

en donde

$$\widehat{N_1^2} = \left(Nw_1\right)^2 - \hat{V}_1\left(Nw_1|\mathbf{m}\right) = N^2\left[w_1^2 - \hat{V}_1\left(w_1|\mathbf{m}\right)\right],$$

$$\hat{V}_1\left(w_1|\mathbf{m}\right) = \frac{N-m}{(N-1)\,m}\left[W_1\left(\widehat{1-W_1}\right)\right] = \frac{N-m}{N\left(m-1\right)} w_1\left(1-w_1\right) =$$

$$= \hat{V}_1\left(w_2|\mathbf{m}\right) = -\widehat{Cov}_1\left(w_1, w_2|\mathbf{m}\right),$$

y al no haber datos faltantes, para $i = 1$,

$$V_2\left(\bar{x}'_i - \bar{x}_i\right) = V_2\left(\bar{x}'_i\right) + V_2\left(\bar{x}_i\right) - 2Cov_2\left(\bar{x}'_i, \bar{x}_i\right),$$

por lo que

$$\hat{V}_2\left(\bar{x}'_i - \bar{x}_i\right) = \hat{V}_2\left(\bar{x}'_i\right) + \hat{V}_2\left(\bar{x}_i\right) - 2\widehat{Cov}_2\left(\bar{x}'_i, \bar{x}_i\right),$$

donde $\hat{V}_2\left(\bar{x}'_i\right)$ está ya estudiado y es conocido (Ruiz Espejo, 1994)

$$\hat{V}_2\left(\bar{x}_i\right) = \frac{N_i - n_i}{N_i}\frac{\widehat{S^2_{x_i}}}{n_i} = \frac{N_i - n_i}{N_i}\frac{s^2_{x_i}}{n_i}$$

donde

$$S_{x_i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{ij} - \bar{X}_i \right)^2,$$

(x_{ij} es el dato declarado de la unidad j en el estrato i), y éste es estimable insesgadamente por

$$\widehat{S_{x_i}^2} = s_{x_i}^2 = \frac{1}{n_i - 1} \sum_{j \in s_i} (x_{ij} - \bar{x}_i)^2$$

(donde ahora x_{ij} es el dato muestral de la j-ésima extracción en la muestra s_i con diseño mas-n_i en el estrato i). Finalmente,

$$Cov_2 \left(\bar{x}_i', \bar{x}_i \right) = \frac{N_i - n_i}{N_i} \frac{S_{x_i' x_i}}{n_i}$$

y por ello su estimador insesgado será

$$\widehat{Cov}_2 \left(\bar{x}_i', \bar{x}_i \right) = \frac{N_i - n_i}{N_i} \frac{\widehat{S_{x_i' x_i}}}{n_i},$$

con

$$S_{x_i' x_i} = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \left(x_{ij}' - \bar{X}_i' \right) \left(x_{ij} - \bar{X}_i \right),$$

que es estimable insesgadamente (en el caso de respuesta en las n_i unidades de la muestra, y de submuestrear en s_{i1} el error de declaración en n_{i1} unidades con diseño mas-n_{i1} de entre la muestra o subpoblación de n_i unidades de la muestra inicial) por el estimador

$$\widehat{S_{x_i' x_i}} = s_{x_i' x_i} = \frac{1}{n_{i1} - 1} \sum_{j \in s_{i1}} \left[x_{ij}' - \bar{x}' \left(s_{i1} \right) \right] \left[x_{ij} - \bar{x} \left(s_{i1} \right) \right]$$

(donde $\bar{x}' \left(s_{i1} \right)$ y $\bar{x} \left(s_{i1} \right)$ son las medias muestrales de x_{ij}' y x_{ij} en la muestra s_{i1}) y éste, $s_{x_i' x_i}$, a su vez es estimable insesgadamente (en el caso de existir o aparecer no respuesta) por el estimador

$$\widehat{s_{x_i' x_i}} = \frac{1}{n_{i11} - 1} \sum_{j \in s_{i11}} \left[x_{ij}' - \bar{x}' \left(s_{i11} \right) \right] \left[x_{ij} - \bar{x} \left(s_{i11} \right) \right],$$

siendo n_{i11} el tamaño muestral de la submuestra s_{i11} ($s_{i11} \subset s_{i1}$) de respondientes inspeccionados de entre las submuestras de no respuesta y de respuesta errada

en el estrato i; las medias muestrales $\bar{x}'(s_{i11})$ y $\bar{x}(s_{i11})$ son las correspondientes de s_{i11} para x'_{ij} y x_{ij} (que sólo en la situación de ausencia de "no respuesta" y de "respuesta errada" en el estrato i, será $n_{i11} = n_{i1}$, deduciéndose entonces que $\widehat{s_{x'_i x_i}} = s_{x'_i x_i}$).

4. Control por muestreo en hacienda pública

Básicamente, el planteamiento del control por muestreo del total recaudado de los contribuyentes como macroparámetro a estimar, tiene fundamentos análogos al del control por muestreo en auditorías que hemos visto en la sección 3 anterior.

Hay que tener en cuenta que la teoría de muestras de poblaciones finitas se apoya en la delimitación e identificación de las unidades que contribuyen con una característica cuantitativa como es su renta o gravamen fiscal.

La "renta nacional" es la suma o el agregado de rentas parciales que se acumulan en el macroparámetro "total de rentas parciales". Entendemos por "renta parcial" a la renta obtenida por uno o varios (siempre en número finito) contribuyentes con su renta personal (originada por su trabajo, capital, propiedades como tierra, construcciones, etc.).

Si además contribuyen con un gravamen personal, el total gravado es la suma de tales gravámenes personales. En este sentido, admitir que varias unidades se unan como un único contribuyente a hacienda, haría su gravamen conjunto mayor que la suma de sus contribuciones parciales o personales, si la escala impositiva es progresiva y creciente (a mayor renta, mayor porcentaje de gravamen). Recíprocamente, si un conjunto de contribuyentes formando una familia, sociedad o grupo, desagrega su renta conjunta en distintas rentas individuales o personales, ante una escala impositiva progresiva y creciente, el gravamen de la sociedad sería mayor que la suma de los gravámenes de los contribuyentes de la familia, sociedad o grupo, desagregados individualmente.

Ante una escala impositiva creciente, la desintegración de la renta en un número grande de contribuyentes hará que el fisco recaude menores tasas.

La metodología estadística para la estimación del total recaudado, del total defraudado, etc. puede realizarse en líneas similares a las tratadas en auditorías. En este caso se supone que no existe la no respuesta, admitiendo que todas las unidades implicadas declaran.

¿Cómo estimar la proporción de defraudadores?

Llamando al "número de unidades de la población, seleccionadas por muestreo aleatorio simple sin reemplazamiento de tamaño n_1^* y en el estrato 1, defraudadoras" como $(d,1)$, y si conocemos los tamaños relativos de los estratos (concretamente W_1), la proporción de defraudadores de tal estrato 1, $p(d|1)$, es (aplicando la definición de probabilidad condicionada de un suceso sobre otro):

$$p(d|1) = \frac{p(d,1)}{p(1)} = \frac{p(d,1)}{W_1} = \frac{(d,1)/N}{W_1} = \frac{(d,1)}{N_1},$$

y un estimador insesgado de $p(d|1)$ es:

$$\hat{p}(d|1) = \frac{\hat{p}(d,1)}{W_1},$$

siendo $\hat{p}(d,1)$ la proporción muestral de unidades defraudadoras en el estrato 1, al inspeccionar las unidades de la submuestra donde se obtiene el dato verdadero. Así, la distribución que sigue el estimador $\hat{p}(d,1)$ es hipergeométrica dividida por n_1^*

$$\hat{p}(d,1) \to \frac{H(n_1^*, N_1, p(d,1))}{n_1^*},$$

y por ello, $\hat{p}(d|1)$ es un estimador insesgado de $p(d|1)$ pues

$$E[\hat{p}(d|1)] = \frac{E[\hat{p}(d,1)]}{W_1} = \frac{p(d,1)}{p(1)} = p(d|1).$$

También la varianza se obtiene así:

$$V[\hat{p}(d|1)] = \frac{1}{W_1^2} V[\hat{p}(d,1)] = \frac{1}{W_1^2} \frac{N_1 - n_1^*}{N_1 - 1} \frac{1}{n_1^*} p(d,1)[1 - p(d,1)] \leq$$

$$\leq \frac{1}{4} \frac{1}{W_1^2} \frac{N_1 - n_1^*}{N_1 - 1} \frac{1}{n_1^*}.$$

Intervalos de confianza para el parámetro $p(d|1)$ pueden obtenerse mediante la desigualdad de Chebychev:

$$p\left[|\hat{p}(d|1) - p(d|1)| < k\right] \geq 1 - \frac{V\left[\hat{p}(d|1)\right]}{k^2} \geq 1 - \frac{N_1 - n_1^*}{W_1^2 (N_1 - 1) n_1^* 4k^2} = 1 - \alpha$$

siendo $1-\alpha$ una cota inferior del nivel de confianza del intervalo para el parámetro $p(d|1)$:

$$\hat{p}(d|1) - k < p(d|1) < \hat{p}(d|1) + k$$

donde, si α está prefijada como cota superior del nivel de significación del intervalo anterior,

$$k = \frac{1}{2W_1}\sqrt{\frac{N_1 - n_1^*}{(N_1 - 1) n_1^* \alpha}}.$$

Estos intervalos de confianza aseguran tal nivel de significación acotado superiormente por α. Otros posibles intervalos aproximados podrían ajustarse más al nivel de significación α, pero no garantizan que el nivel de significación esté con *seguridad* a lo sumo acotado por α. En este sentido puede decirse que estos intervalos propuestos anteriormente son más amplios que los teóricamente exactos, pero éstos son impracticables pues requieren conocer el parámetro $p(d,1)$ y tener tabulada la distribución hipergeométrica de parámetros $H(n_1^*, N_1, p(d,1))$.

Si hubiéramos considerado fijados k y α, el tamaño muestral para tales especificaciones se obtiene de la relación anterior, y operando:

$$N_1 - n_1^* = 4W_1^2 k^2 (N_1 - 1) n_1^* \alpha$$

o bien

$$n_1^* \left[4W_1^2 k^2 (N_1 - 1)\alpha + 1\right] = N_1,$$

es decir, despejando n_1^* de la ecuación anterior,

$$n_1^* = \frac{N_1}{4W_1^2 k^2 (N_1 - 1)\alpha + 1}.$$

El tamaño muestral con el diseño mas-n_1^* no puede ser mayor que el tamaño poblacional N_1.

De modo diferente, $p(d,2)$, la proporción de defraudadores en el estrato de unidades grandes, se puede calcular directamente, ya que se realiza un censo en el estrato 2 citado. Entonces, la proporción de unidades defraudadoras de la población global es $p(d) = p(d,1) + p(d,2)$, estimable insesgadamente por

$$\hat{p}(d) = \hat{p}(d,1) + p(d,2) \rightarrow \frac{H\left(n_1^*, N_1, p\left(d,1\right)\right)}{n_1^*} + p\left(d,2\right),$$

y su varianza es

$$V\left[\hat{p}(d)\right] = V\left[\frac{H\left(n_1^*, N_1, p\left(d,1\right)\right)}{n_1^*}\right] =$$

$$= \frac{N_1 - n_1^*}{N_1 - 1}\frac{1}{n_1^*}p\left(d,1\right)\left[1 - p\left(d,1\right)\right] \leq \frac{N_1 - n_1^*}{\left(N_1 - 1\right)n_1^* 4},$$

permitiéndonos construir intervalos de confianza para el parámetro "proporción de defraudadores", $p(d)$. Por la desigualdad de Chebychev,

$$p\left[\left|\hat{p}(d) - p(d)\right| < k\right] \geq 1 - \frac{V\left[\hat{p}\left(d\right)\right]}{k^2} \geq$$

$$\geq 1 - \frac{N_1 - n_1^*}{\left(N_1 - 1\right)n_1^* 4k^2} = 1 - \alpha.$$

Así, fijados n_1^* y α:

$$k = \frac{1}{2}\sqrt{\frac{N_1 - n_1^*}{\left(N_1 - 1\right)n_1^*\alpha}}.$$

Si fijamos n_1^* y k:

$$\alpha = \frac{N_1 - n_1^*}{\left(N_1 - 1\right)n_1^* 4k^2},$$

con la restricción $\alpha \leq 1$.

Si fijamos k y α:

$$n_1^*\left[\left(N_1 - 1\right)4k^2\alpha + 1\right] = N_1,$$

de donde

$$n_1^* = \frac{N_1}{\left(N_1 - 1\right)4k^2\alpha + 1}$$

que sería el tamaño muestral a observar para la variable verdadera x_i' en el estrato 1, para k y α. En todos los casos anteriores el intervalo de confianza es

$$\hat{p}(d) - k < p(d) < \hat{p}(d) + k.$$

Los comentarios realizados al desarrollar la construcción del intervalo para estimar $p(d|1)$, son trasladables a este caso con $p(d)$.

5. Control por muestreo de la renta nacional

Ahora nos encontramos con una población dividida en tres estratos:

Estrato 1: Las unidades que no tienen el deber legal de declarar a hacienda.
Estrato 2: Compuesto por el estrato primero utilizado en el control por muestreo de hacienda.
Estrato 3: Estrato de "inclusión segura en la muestra" de hacienda o que se examinará completamente (estrato de unidades grandes).

Ahora el tamaño poblacional N es mayor que en la sección 4 anterior. Un estimador insesgado que integra la información de hacienda (sección 4) al estudio y control de la renta nacional, es:

$$\bar{x}_{st}''' = W_1\bar{x}_1' + W_2\bar{x}_2' + W_3\bar{X}_3' = W_1\bar{x}_1' + (1 - W_1)\,\bar{x}_{st}'$$

puesto que

$$W_2\bar{x}_2' + W_3\bar{X}_3' = \frac{N_2}{N}\bar{x}_2' + \frac{N_3}{N}\bar{X}_3' = \frac{N_2 + N_3}{N}\bar{x}_{st}' = \frac{N - N_1}{N}\bar{x}_{st}'.$$

El estimador $\bar{x}_{st}'''$ estima insesgadamente la renta media verdadera (media no per cápita, sino por unidad poblacional). Efectivamente,

$$E\,(\bar{x}_{st}''') = \bar{X}'.$$

La estimación insesgada de la renta nacional per cápita, se obtiene mediante la corrección

$$\frac{N}{M}\bar{x}_{st}''',$$

siendo N el número de unidades de la población, y M el número de habitantes del país.

La varianza del estimador es

$$V\left(\bar{x}_{st}'''\right) = W_1^2 V\left(\bar{x}_1'\right) + \left(1 - W_1\right)^2 V\left(\bar{x}_{st}'\right),$$

que es estimable insesgadamente, pues $\hat{V}\left(\bar{x}_1'\right)$ está proporcionada en Ruiz Espejo (1994) y $\hat{V}\left(\bar{x}_{st}'\right)$ está en las secciones anteriores, transformando adecuadamente las notaciones y su significado.

Un estimador insesgado del total poblacional defraudado, $T_d = X' - X$, respecto del parámetro verdadero $\bar{X}'$, es

$$\hat{T}_d = N\left(\bar{x}_{st}''' - \bar{x}_{st}\right) = N\left[W_1\left(\bar{x}_1' - \bar{x}_1\right) + W_2\left(\bar{x}_2' - \bar{x}_2\right)\right],$$

siendo

$$\bar{x}_{st} = \sum_{i=1}^{3} W_i \bar{x}_i,$$

con $\bar{x}_3 = \bar{X}_3'$. En efecto,

$$E\left(\hat{T}_d\right) = N_1\left(\bar{X}_1' - \bar{X}_1\right) + N_2\left(\bar{X}_2' - \bar{X}_2\right) =$$

$$= X_1' + X_2' - \left(X_1 + X_2\right) =$$

$$= \sum_{i=1}^{3} X_i' - \sum_{i=1}^{3} X_i = X' - X = T_d.$$

Y la varianza es

$$V\left(\hat{T}_d\right) = N^2 \sum_{i=1}^{2} W_i^2 V_2\left(\bar{x}_i' - \bar{x}_i\right),$$

cuyo estimador insesgado es

$$\hat{V}\left(\hat{T}_d\right) = N^2 \sum_{i=1}^{2} \widehat{W_i^2} \hat{V}_2\left(\bar{x}_i' - \bar{x}_i\right)$$

siendo

$$\widehat{W_i^2} = w_i^2 - \hat{V}_1\left(w_i|\mathbf{m}\right) = w_i^2 - \frac{N-m}{(N-1)\,m} \frac{N-1}{N} \frac{m}{m-1} w_i\left(1 - w_i\right) =$$

$$= w_i^2 - \frac{N-m}{N(m-1)} w_i (1-w_i),$$

cuando W_i fuera desconocido. En el caso en el que W_i fuera conocido, $\widehat{W_i^2} = W_i^2$ directamente (sin estudio muestral piloto). Además,

$$V_2(\bar{x}_i' - \bar{x}_i) = E_2\left[(\bar{x}_i' - \bar{x}_i)^2\right] - \left(\bar{X}_i' - \bar{X}_i\right)^2,$$

y así, un estimador insesgado de esta varianza, es

$$\hat{V}_2(\bar{x}_i' - \bar{x}_i) = (\bar{x}_i' - \bar{x}_i)^2 - \left[\widehat{\left(\bar{X}_i' - \bar{X}_i\right)^2}\right]$$

con

$$\left[\widehat{\left(\bar{X}_i' - \bar{X}_i\right)^2}\right] = \left(\bar{X}_i'^2 + \widehat{\bar{X}_i^2} - 2\bar{X}_i\widehat{\bar{X}_i'}\right) =$$

$$= \widehat{\bar{X}_i'^2} + \widehat{\bar{X}_i^2} - 2\left(\widehat{\bar{X}_i\bar{X}_i'}\right) =$$

$$= \bar{x}_i'^2 - \hat{V}_2(\bar{x}_i') + \bar{x}_i^2 - \hat{V}_2(\bar{x}_i) - 2\left[\bar{x}_i\bar{x}_i' - \widehat{Cov}_2(\bar{x}_i, \bar{x}_i')\right],$$

donde

$\hat{V}_2(\bar{x}_i')$ está en el trabajo de Ruiz Espejo (1994),

$$\hat{V}_2(\bar{x}_i) = \frac{N_i - n_i}{N_i}\frac{\widehat{S_{x_i}^2}}{n_i} = \frac{N_i - n_i}{N_i}\frac{s_{x_i}^2}{n_i},$$

y

$$\widehat{Cov}_2(\bar{x}_i, \bar{x}_i') = \frac{N_i - n_i}{N_i}\frac{\widehat{S_{x_i x_i'}}}{n_i} = \frac{N_i - n_i}{N_i}\frac{s_{x_i x_i'}}{n_i},$$

siendo

$$S_{x_i x_i'} = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(x_{ij}' - \bar{X}_i'\right)\left(x_{ij} - \bar{X}_i\right)$$

que es estimable insesgadamente por

$$s_{x_i x_i'} = \frac{1}{n_i - 1}\sum_{j \in s_i(n_i)}\left(x_{ij}' - \bar{x}_i'\right)(x_{ij} - \bar{x}_i),$$

en el que hemos denotado por $s_i(n_i)$ a la muestra de tamaño n_i, en el estrato i. A su vez, si se submuestrea en $s_i(n_i)$, la nueva cuasicovarianza submuestral es un estimador insesgado de $s_{x_i x_i'}$.

La estimación de la renta nacional en la situación en la que los pesos relativos de los estratos, W_i, sean desconocidos y se estimen por una muestra piloto a través de $w_i = m_i/m$ $(i = 1, 2, 3)$, puede basarse en el estimador

$$\bar{x}''''_{st} = w_1\bar{x}'_1 + w_2\bar{x}'_2 + w_3\bar{X}'_3,$$

que es insesgado para $\bar{X}'$, y su varianza es

$$V\left(\bar{x}''''_{st}\right) = V_2\left(W_1\bar{x}'_1 + W_2\bar{x}'_2 + W_3\bar{X}'_3\right) +$$

$$+E_2\left\{\frac{1}{m^2}\frac{N-m}{N-1}m\left[W_1\left(1-W_1\right)\bar{x}'^2_1 + W_2\left(1-W_2\right)\bar{x}'^2_2 + W_3\left(1-W_3\right)\bar{X}'^2_3\right] +\right.$$

$$\left. + 2\bar{x}'_1\bar{x}'_2 Cov\left(w_1, w_2|\mathbf{m}\right) + 2\bar{x}'_1\bar{X}'_3 Cov\left(w_1, w_3|\mathbf{m}\right) + 2\bar{x}'_2\bar{X}'_3 Cov\left(w_2, w_3|\mathbf{m}\right)\right\} =$$

$$= \sum_{i=1}^{2} W_i^2 V_2\left(\bar{x}'_i\right) + \frac{N-m}{(N-1)\,m}\sum_{i=1}^{3} W_i\left(1-W_i\right)\left[\bar{X}'^2_i + V_2\left(\bar{x}'_i\right)\right] -$$

$$-2\frac{N-m}{(N-1)\,m}\sum_{i=1}^{2}\sum_{j=i+1}^{3} W_i W_j \bar{X}'_i\bar{X}'_j,$$

donde $\bar{x}'_3 = \bar{X}'_3$.

Esta varianza, $V\left(\bar{x}''''_{st}\right)$, puede estimarse insesgadamente por

$$\hat{V}\left(\bar{x}''''_{st}\right) = \sum_{i=1}^{2}\widehat{W_i^2}\hat{V}_2\left(\bar{x}'_i\right) +$$

$$+\frac{N-m}{(N-1)\,m}\sum_{i=1}^{3}\left[W_i\widehat{\left(1-W_i\right)}\right]\left[\widehat{\bar{X}'^2_i} + \hat{V}_2\left(\bar{x}'_i\right)\right] -$$

$$-2\frac{N-m}{(N-1)\,m}\sum_{i=1}^{2}\sum_{j=i+1}^{3}\left(\widehat{W_i W_j}\right)\widehat{\bar{X}'_i}\widehat{\bar{X}'_j},$$

donde

$$\widehat{W_i^2} = w_i^2 - \frac{N-m}{(N-1)\,m}w_i\left(1-w_i\right),$$

como ya conocíamos; también,

$$\left[W_i\widehat{\left(1-W_i\right)}\right] = \widehat{W_i} - \widehat{W_i^2} = w_i - w_i^2 + \frac{N-m}{(N-1)\,m}w_i\left(1-w_i\right) =$$

$$= \left[1 - \frac{N-m}{(N-1)\,m}\right][w_i\,(1-w_i)] = \frac{N\,(m-1)}{(N-1)\,m} w_i\,(1-w_i)\,,$$

y

$$\widehat{(W_i W_j)} = w_i w_j - \widehat{Cov}_1\,(w_i, w_j|\mathbf{m}) =$$

que por Ruiz Espejo (n.d.),

$$= w_i w_j - \frac{N-m}{N}\frac{\widehat{S_{I_i(\cdot),I_j(\cdot)}}}{m} = (**)$$

donde hemos denotado

$$I_i(k) = \begin{cases} 1 & \text{si la unidad } k \text{ sí pertenece al estrato } i \\ 0 & \text{si la unidad } k \text{ no pertenece al estrato } i \end{cases}$$

por lo que

$$I_i(k) I_j(k) = 0,$$

para todo k cuando los estratos i y j son distintos ($i \neq j$); y por Ruiz Espejo (1995),

$$(**) = w_i w_j - \frac{N-m}{Nm}\left\{\frac{1}{N-1}\sum_{k=1}^{N}\widehat{[I_i(k) - W_i]\,[I_j(k) - W_j]}\right\} =$$

$$= w_i w_j - \frac{N-m}{Nm}\frac{1}{m-1}\sum_{k=1}^{m}[I_i(k) - w_i]\,[I_j(k) - w_j] =$$

$$= w_i w_j + \frac{N-m}{N\,(m-1)} w_i w_j = \left[\frac{N-m}{N\,(m-1)} + 1\right] w_i w_j =$$

$$= \frac{(N-1)\,m}{N\,(m-1)} w_i w_j.$$

Además $\hat{V}_2\,(\bar{x}'_i)$ es conocido (Ruiz Espejo, 1994),

$$\widehat{\bar{X}'^2_i} = \bar{x}'^2_i - \hat{V}_2\,(\bar{x}'_i)\,,$$

y $\widehat{\bar{X}'_i} = \bar{x}'_i$.

El estimador insesgado del total defraudado en la renta nacional será

$$\hat{T}_d = N\left(\bar{x}'''_{st} - \sum_{i=1}^{3} w_i \bar{x}_i\right) = N\sum_{i=1}^{2} w_i\left(\bar{x}'_i - \bar{x}_i\right)$$

donde hemos supuesto que $\bar{x}_3 = \bar{X}'_3$ (la generalización cuando no se admite este supuesto, es inmediata). Bajo esta hipótesis,

$$V\left(\hat{T}_d\right) = \sum_{i=1}^{2} W_i^2 V_2\left(\bar{x}'_i - \bar{x}_i\right) +$$

$$+E_2\left\{\left[\sum_{i=1}^{2}\left(\bar{x}'_i - \bar{x}_i\right)^2 V_1\left(w_i|\mathbf{m}\right)\right] + 2\left(\bar{x}'_1 - \bar{x}_1\right)\left(\bar{x}'_2 - \bar{x}_2\right) Cov_1\left(w_1, w_2|\mathbf{m}\right)\right\}$$

que es estimable insesgadamente, ya que W_i^2,

$$V_1\left(w_i|\mathbf{m}\right) = \frac{N-m}{(N-1)\,m} W_i\left(1 - W_i\right),$$

y

$$Cov_1\left(w_1, w_2|\mathbf{m}\right) = -\frac{N-m}{(N-1)\,m} W_1 W_2$$

lo son; así como en el caso de inexistencia de no respuesta,

$$V_2\left(\bar{x}'_i - \bar{x}_i\right) = V_2\left(\bar{x}'_i\right) + V_2\left(\bar{x}_i\right) - 2Cov_2\left(\bar{x}'_i, \bar{x}_i\right),$$

vistas en Ruiz Espejo (1994) para $V_2\left(\bar{x}'_i\right)$, y sus estimadores insesgados $\hat{V}_2\left(\bar{x}'_i\right)$,

$$\hat{V}_2\left(\bar{x}_i\right) = \frac{N_i - n_i}{N_i}\frac{s_{x_i}^2}{n_i},$$

y

$$\widehat{Cov}_2\left(\bar{x}'_i, \bar{x}_i\right) = \frac{N_i - n_i}{N_i}\frac{\widehat{S^*_{x'_i x_i}}}{n_i},$$

siendo

$$\widehat{S^*_{x'_i x_i}} = \widehat{s_{x'_i x_i}} = \frac{1}{n_{i1} - 1}\sum_{j \in s_i^*}\left[x'_{ij} - \bar{x}'_i\left(s_i^*\right)\right]\left[x_{ij} - \bar{x}_i\left(s_i^*\right)\right],$$

donde hemos denotado por s_i^* a la submuestra de tamaño n_{i1} de entre las n_i unidades de la submuestra s_i del estrato i de tamaño N_i; también

$$\bar{x}'_i\left(s_i^*\right) = \frac{1}{n_{i1}}\sum_{j \in s_i^*} x'_j$$

con $x'_j = x_j - e_j$ ($-e_j$ es el fraude, error de observación, de medida o de declaración

en la unidad j de s_i^*) y

$$\bar{x}_i\left(s_i^*\right) = \frac{1}{n_{i1}} \sum_{j \in s_i^*} x_j.$$

Bibliografía

ARMSTRONG, J., BLOCK, C. y SRINATH, K.P. (1993): Two-phase sampling of tax records for business surveys. *J. Bus. Econ. Statist.* **11**, 407-416.

CASAS SÁNCHEZ, J.M. y SANTOS PEÑAS, J. (1996): *Introducción a la Estadística para Economía y Administración de Empresas.* Madrid, Centro de Estudios Ramón Areces.

DALENIUS, T. (1957): *Sampling in Sweden.* Estocolmo, Almqvist and Wicksell.

FULLER, W.A. (1991): Simple estimators for the mean of skewed populations. *Statistica Sinica* **1**, 137-158.

GLASSER, G.J. (1962): On the complete coverage of large units in a statistical study. *Rev. Internat. Statist. Inst.* **30**, 28-32.

LAVALLÉE, P. e HIDIROGLOU, M.A. (1988): On the stratification of skewed populations. *Survey Methodology* **14**, 33-43.

PECHMAN, J.A. (1990): The future of the income tax. *Amer. Econ. Rev.* **80**, 1-20.

PRESS, S.J. (1995): Sample-audit tax assessment for businesses: what's fair? *J. Bus. Econ. Statist.* **13**, 357-359.

RAO, C.R. (1994): *Estadística y Verdad.* Barcelona, PPU.

RUIZ ESPEJO, M. (1988): Estimación insesgada con observaciones erradas y no respuesta. *Trabajos de Estadística* **3**, 71-80.

--------- (1994): Problema 52. *Qüestiió* **18**, 281, 467-468.

--------- (1995): Una relación entre cuasicovarianzas muestral y poblacional. *Rev. Acad. Cienc. Zaragoza* **50**, 51-53.

--------- (1996): Problema 58. *Qüestiió* **20**, 109, 373-374.

--------- (n.d.): Covariance of sample moments for finite populations. *Proc. Roy. Irish Acad. Sect. A*, en prensa.