Marcos de Muestreo Imperfectos

Autor: Lic. Fernando O. Rivero Suguiura

1. Planteamiento de los problemas de Marcos Muestrales

En la teoría de poblaciones finitas se supone que el marco de lista de la que se selecciona la muestra, coincide con las unidades que son parte de la población objeto de estudio. En la práctica no siempre esto es cierto, ya que las listas presentan con cierta frecuencia algunos defectos que pueden dar lugar a la aparición de sesgos y a la alteración de las varianzas de los estimadores. Los problemas más frecuentes en los marcos muestrales de lista, son:

- *La sobrecobertura*
- *La subcobertura*
- *La duplicación*

Las dos últimas limitaciones de marco muestral no son tratados en esta investigación, pero vale la pena nombrarlos:

La subcobertura se presenta cuando existen unidades que están en la población objetivo y no en el marco, luego el marco muestral se convierte en un subconjunto de la población. Los siguientes ejemplos describen esta situación:

- Suponer que existen nuevas viviendas ocupadas en zonas marginales de una ciudad que son parte de la población, pero por falta de actualización del marco muestral estas no están presentes en dicho marco, dándoles probabilidad igual a cero de ser seleccionadas en la muestra. En este caso se convierten en unidades faltantes.

- Se muestrea de una lista de establecimientos económicos manufactureros del censo 2000 sin considerar establecimientos de nueva creación que son parte de la población. También son unidades faltantes.

La duplicación de unidades muestrales es cuando el marco muestral contiene elementos repetidos.

Siguiendo los dos ejemplos anteriores:

- Suponer que una vivienda se registra en el marco muestral como ocupada con tres personas, tal cual ocurre con la población objetivo en el periodo de tiempo **1.** Suponer que las tres personas dejan dicha vivienda y son ocupadas por otras cinco personas distintas a las anteriores. Esta vivienda es nuevamente registrada en el marco como ocupada en el periodo de tiempo **2.** Si no se elimina del marco la vivienda del periodo de tiempo 1, se presenta el problema de unidad duplicada.

- El establecimiento económico de nombre "El Buen Productor" está

registrado en el marco como parte de la población objetivo. Suponer el fallecimiento del dueño del establecimiento económico y el cierre temporal. Después de un periodo de tiempo, se abre nuevamente con el nombre "Carlos Andrade" y éste es registrado en el marco muestral como nuevo establecimiento económico. Es un caso de duplicación.

La existencia de *unidades extrañas* en el marco muestral es tema de esta investigación. Por unidad extraña se entiende una unidad de muestreo que, incluida en el marco, no pertenece a la población que se desea estudiar, o que no es una unidad del colectivo que se desea seleccionar para ser parte de la muestra.

Como ejemplos consideremos los siguientes:

- Para una encuesta se desea seleccionar una muestra aleatoria de viviendas ocupadas del marco muestral disponible. Si muchas de las especificadas en el marco están como ocupadas y en la población objetivo son desocupadas, se pueden muestrear unidades extrañas.

- Si para estimar alguna característica de producción en empresas manufactureras del país, se muestrea de una lista de establecimientos económicos manufactureros que en la actualidad cerraron sus puertas por falta de utilidad. Se puede seleccionar unidades extrañas.

Los ejemplos anteriores ponen de manifiesto que la presencia de unidades extrañas en el marco ocurre principalmente en dos tipos de situaciones prácticas:

a) Por no tener actualizado el marco muestral. La lista incluye unidades que han dejado de pertenecer a la población objetivo.

b) La población que se desea muestrear se convierte en una subpoblación del marco muestral generándose el problema de la *sobrecobertura.*

2. El problema de las unidades extrañas en el marco

Ante la situación de sobrecobertura de marcos muestrales, pueden adoptarse diferentes reglas de solución, no todas igualmente adecuadas, cuya puesta en práctica depende de los recursos disponibles.

a) Depuración del marco de muestreo

Eliminar del marco las unidades extrañas sabiendo el número total de inclusión. Con este proceso queda solucionado el problema de sobrecobertura puesto que luego se procede a seleccionar la muestra del marco depurado. En muchos casos esto no será posible con los recursos dados, bien sea porque el marco disponible no contiene información acerca de cuáles unidades son extrañas (siendo necesario un trabajo de campo de carácter exhaustivo o la construcción de un nuevo marco muestral) o por limitaciones de tiempo, presupuesto, personal no disponible, etc.

b) Reemplazo de las unidades extrañas en la muestra

Otra solución consiste en seleccionar una muestra aleatoria del marco disponible no depurado y sustituir las unidades que resulten ser extrañas por otras aleatoriamente seleccionadas de entre las restantes del marco, hasta completar el

tamaño de la muestra planificada con unidades todas no extrañas. Lamentablemente esta solución da lugar a estimaciones sesgadas, como se verá luego.

De acuerdo a los dos puntos anteriores, se considera algunos procedimientos para la estimación del total.

3. Relación de párametros poblacionales en marcos depurados y no depurados

Suponer que de las N unidades de muestreo incluidas en el marco, N^* son unidades no extrañas, y por lo tanto $N - N^*$ son unidades extrañas. Entonces se puede enumerar de 1 a N^* las unidades que no son extrañas y de $N^* + 1$ a N las unidades extrañas del marco. Es decir que el marco de muestreo disponible no depurado se constituye en el conjunto:

$$\Omega = \{U_1, U_2, \ldots, U_N\} \qquad (1)$$

y el conjunto

$$\Omega^* = \{U_1, U_2, \ldots, U_{N^*}\} \qquad (2)$$

es el marco de muestreo depurado.

La proporción de unidades extrañas en Ω es:

$$W = \frac{N - N^*}{N} = 1 - \frac{N^*}{N} \qquad (3)$$

Sea Y_i el valor de una variable que se desea investigar de la i-ésima unidad U_i, se atribuye el valor de Y_i igual a cero cuando U_i es una unidad extraña. Por tratarse de una unidad extraña, el valor de Y_i puede ser realmente cero o no estar definido. Ya que la unidad muestral no pertenece a la población objeto de estudio cuyo total se desea estimar, su contribución a dicho total es nula, justificando así el valor de Y_i igual a cero.

Los valores totales de ambos marcos son entonces coincidentes, es decir:

$$T = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N^*} Y_i \qquad (4)$$

al igual que:

$$\sum_{i,j}^{N} Y_i Y_j = \sum_{i,j}^{N^*} Y_i Y_j$$

$$\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N^*} Y_i^2 \qquad (5)$$

sin embargo las medias en uno u otro marco no son iguales:

$$\overline{Y} = (1 - W)\overline{Y}^* \qquad (6)$$

puesto que

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{N^*}{N^*} \frac{\sum_{i=1}^{N^*} Y_i}{N} =$$

$$= \frac{N^*}{N} \frac{\sum_{i=1}^{N^*} Y_i}{N^*} = (1 - W)\overline{Y}^*$$

y la varianza queda como:

$$\sigma^2 = (1 - W)\sigma^{*2} + \frac{W}{1 - W}\overline{Y}^2 \qquad (7)$$

La relación (7) entre varianzas en el marco no depurado σ^2 y en el marco

depurado σ^{*2} se obtiene de la siguiente manera:

$$N\sigma^2 = \sum_{i=1}^{N} Y_i^2 - N\left(\frac{\sum_{i=1}^{N} Y}{N}\right)^2 =$$

$$= \sum_{i=1}^{N} Y_i^2 - \frac{T^{*2}}{N^*} + \frac{T^{*2}}{N^*} - \frac{T^2}{N}$$

$$N\sigma^2 = N^*\sigma^{*2} + T^2\left(\frac{1}{N^*} - \frac{1}{N}\right) =$$

$$= N^*\sigma^{*2} + T^2\left(\frac{N - N^*}{N^* N}\right) \qquad (8)$$

Despejando σ^2 y sustituyendo $\frac{N - N^*}{N}$ por W y $\frac{N^*}{N}$ por $1 - W$, resulta la relación propuesta (7).

Si se admite la aproximación de $N - 1 \cong N$ y $N^* - 1 \cong N^*$, la relación de cuasivarianzas de ambos marcos, es:

$$S^2 = (1 - W)S^{*2} + \frac{W}{1 - W}\bar{Y}^2 \qquad (9)$$

4. Estimador del total $\hat{T}^*$ cuando se muestrea con el marco depurado

Si se considera un muestreo aleatorio simple sin reemplazo con probabilidades iguales de selección, se tiene que el estimador del total $\hat{T}^*$, es:

$$\hat{T}^* = \frac{N^*}{n} t \qquad (10)$$

con n tamaño de muestra y $t = \sum_{i=1}^{n} Y_i$ total muestral. Es un estimador insesgado de T, cuya varianza es:

$$V(\hat{T}^*) = N^{*2}(1 - f^*)\frac{S^{*2}}{n} \qquad (11)$$

donde $f^* = \frac{n}{N^*}$ es la fracción de muestreo y S^{*2} denota la cuasivarianza en el marco depurado.

Estos dos resultados forman parte de la teoría elemental de muestreo de poblaciones finitas, por ello, se aceptan sin demostración.

5. Estimador del total $\hat{T}$ cuando se muestrea con el marco no depurado

Se señalan dos aspectos en la estimación, estos son:

5.1. No se sustituyen las unidades extrañas que aparecen en la muestra

En estas condiciones, si se considera un muestreo aleatorio simple sin reemplazo con probabilidades iguales de selección, se tiene que el estimador del total $\hat{T}_1$, es:

$$\hat{T}_1 = \frac{N}{n} t \qquad (12)$$

la expresión (12) es un estimador insesgado y su varianza está dada por:

$$V(\hat{T}_1) = N^2(1 - f)\frac{S^2}{n} \qquad (13)$$

donde $f = \frac{n}{N}$ y S^2 es la cuasivarianza en el marco no depurado.

La varianza del total estimado con marco muestral no depurado de la expresión (13), es mayor a la varianza del total estimado con marco muestral depurado de la expresión (11). Para probar que $V(\hat{T}_1) > V(\hat{T}^*)$ se determina la diferencia entre ambas, es decir:

$$d = N^2(1-f)\frac{S^2}{n} - N^{*2}(1-f^*)\frac{S^{*2}}{n} \qquad (14)$$

Para verificar que $d > 0$, se debe tener en cuenta que:

$$N^2(1-f) > N^{*2}(1-f^*) \qquad (15)$$

puesto que $N > N^*$ y que

$$f^* = \frac{n}{N^*} > f = \frac{n}{N}$$

Luego, analizar diferencias entre cuasivarianzas de ambos marcos muestrales, es decir:

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - \frac{T^2}{N}\right) =$$

$$= \frac{1}{N-1}\sum_{i=1}^{N} Y_i^2 - \frac{T^2}{N(N-1)}$$

$$S^{*2} = \frac{1}{N^*-1}\left(\sum_{i=1}^{N^*} Y_i^2 - \frac{T^2}{N^*}\right) =$$

$$= \frac{1}{N^*-1}\sum_{i=1}^{N^*} Y_i^2 - \frac{T^2}{N^*(N^*-1)} \qquad (16)$$

de acuerdo con (11) y (13) se obtiene:

$$d = \frac{1}{n}\left\{\sum_{i=1}^{N} Y_i^2\left[\frac{N(N-1)}{N-1} - \frac{N^*(N^*-1)}{N^*-1}\right] - T^2\left(\frac{N-n}{N-1} - \frac{N^*-n}{N^*-1}\right)\right\} \qquad (17)$$

Puesto que $T^2 = \sum_{i=1}^{N} Y_i^2 + \sum_{i\neq j}^{N} Y_i Y_j$,

reemplazando en (17) y simplificando, se tiene:

$$d = \frac{N-N^*}{(N-1)(N^*-1)n}\left[(N-1)(N^*-1)\left(\sum_{i=1}^{N} Y_i^2\right) - (n-1)\left(\sum_{i\neq j}^{N} Y_i Y_j\right)\right] \qquad (18)$$

el primer factor de d es positivo, se debe probar que también lo es el segundo factor. En efecto, por ser:

$$N\sigma^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 > 0 \qquad (19)$$

se cumple que

$$(N-1)\sum_{i=1}^{N} Y_i^2 > \sum_{i=1}^{N} Y_i Y_j \qquad (20)$$

Por tanto, si se multiplica el primer miembro de la relación (20) por N^*-1 y el segundo por $n-1$, dado que $N^*-1 > n-1$ y ambos positivos, también es:

$$(N-1)(N^*-1)\sum_{i=1}^{N} Y_i^2 > (n-1)\sum_{i=1}^{N} Y_i Y_j \qquad (21)$$

lo cual completa la demostración:

$$V(\hat{T}_1) > V(\hat{T}^*)$$

La comparación de las varianzas de $\hat{T}_1$ y $\hat{T}^*$ también se puede realizar mediante el cociente de forma muy simple con el grado de aproximación que permita la sustitución de $N \cong N-1$, $N^* \cong N^*-1$. Bajo el criterio anterior, las varianzas son iguales a las cuasivarianzas, y de acuerdo a (6) y (9) se obtiene:

$$\frac{V(\hat{T}_1)}{V(\hat{T}^*)} = \frac{N-n}{N^*-n}\left(1+W\frac{\bar{Y}^{*2}}{S^{*2}}\right) \qquad (22)$$

Relación que también pone de manifiesto la mayor varianza de $\hat{T}_1$, por ser $N > N^*$ cuando existen unidades extrañas.

5.2 Se sustituyen aleatoriamente las unidades extrañas que aparecen en la muestra

Así, si se selecciona del marco no depurado sin reemplazo y con probabilidades iguales, el estimador del total $\hat{T}_2$, es:

$$\hat{T}_2 = \frac{N}{n}t \qquad (23)$$

La expresión (23) es un estimador sesgado. Para determinar el sesgo y la varianza de $\hat{T}_2$ basta tener en cuenta que la sustitución aleatoria de las unidades extrañas de la muestra hasta obtener n no extrañas, equivale a efectuar el muestreo en el marco depurado; así resulta que para las variables indicadoras $e_i = 1$, si U_i pertenece a la muestra; y $e_i = 0$, si U_i no pertenece a la muestra, se cumple:

$$E(e_i) = 0 \qquad \text{si } U_i \text{ es extraña}$$

$$E(e_i) = \frac{n}{N^*} \qquad \text{si } U_i \text{ es no extraña} \qquad (24)$$

por tanto:

$$E(\hat{T}_2) = \frac{N}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{N}{n}E\left(\sum_{i=1}^{N} Y_i e_i\right) =$$

$$= \frac{N}{n}\sum_{i=1}^{N} Y_i E(e_i) = \frac{N}{n}\sum_{i=1}^{N^*} Y_i \frac{n}{N^*}$$

$$E(\hat{T}_2) = \frac{N}{N^*}T = \frac{1}{1-W}T \neq T \qquad (25)$$

en consecuencia el sesgo, es:

$$B(\hat{T}_2) = T\left(\frac{1}{1-W}-1\right) = \frac{W}{1-W}T \qquad (26)$$

luego la varianza sería:

$$V(\hat{T}_2) = N^2(1-f^*)\frac{S^{*2}}{n} \qquad (27)$$

donde $f^* = \frac{n}{N^*}$ y S^{*2} es la cuasivarianza similar al del marco depurado.

Teniendo en cuenta el resultado de (11):

$$V(\hat{T}_2) = \frac{1}{(1-W)^2}V(\hat{T}^*) \qquad (28)$$

La relación anterior (28), prueba que la varianza de $\hat{T}_2$ es mayor que la de $\hat{T}^*$, puesto que cuando existen unidades extrañas se comprueba que $0 < (1-W)^2 < 1$.

De (27) y (28) se obtiene el error cuadrático medio de $\hat{T}_2$, que es igual a:

$$ECM\left(\hat{T}_2\right)=\frac{1}{(1-W)^2}\left[V\left(\hat{T}^*\right)+W^2T^{*2}\right]$$

6. Conclusiones y recomendaciones

- Si no se conoce el número de unidades extrañas incluidas en el marco, que es lo más probable, los estimadores son insesgados pero con varianza mayor al de un marco muestral con unidades depuradas. Si se identifican unidades extrañas seleccionadas en la muestra y se sustituyen por otras unidades no extrañas de un marco no depurado, entonces los estimadores son sesgados con varianza mayor al de un marco muestral depurado.

- Lo ideal es realizar actualizaciones de marcos de muestreo en periodos cortos de tiempo (año) antes de levantar una encuesta eliminando unidades extrañas. El costo y los recursos, impiden dicha labor, especialmente en los países subdesarrollados, tal es el caso nuestro. Una alternativa de solución al problema, es estimar la proporción de unidades extrañas en el marco ($\hat{W}$), mediante los listados de actualización de una muestra de unidades primarias de muestreo que arrojen un resultado estimativo y hagan que la varianza de los estimadores a obtenerse sea menor y así aumentar la precisión y disminuir el error de muestreo de los estimadores que se obtengan.

7. Bibliografía

[1] *Pandurang V. Sukhatme,* Teoría de Encuestas por Muestreo con Aplicaciones

[2] *Sharon L. Lohr,* Muestreo – Diseño y Análisis

[3] *Carl – Erik Sarndal,* Model Assisted Survey Sampling

[4] *Azorin Poch,* Curso de Muestreo y Aplicaciones

[5] *Des Raj,* La Estructura de las Encuestas por Muestreo

[6] *Leslie Kish,* Muestreo de Encuestas

"El tranquilo ha cambiado nuestro mundo, no tanto descubriendo nuevos hechos o desarrollos técnicos, sino cambiando los modos de razonar, de experimentar y de formar nuestras opiniones acerca de él."

Hacking