Metodología de Encuestas
Vol 1, Num 1, 1999, 19-28

APORTACIONES AL MUESTREO SUCESIVO

Eva María Artés Rodríguez
Universidad de Almería

Mª del Mar Rueda García
Antonio Arcos Cebrián
Universidad de Granada

RESUMEN

En este trabajo se desarrolla la teoría del muestreo sucesivo utilizando un estimador de razón-producto de doble muestreo para la parte apareada de la muestra. Se obtienen las expresiones para la fracción del apareamiento óptimo y para el estimador combinado junto con su error. Se presentan resultados para algunos casos especiales de aplicación práctica. Para evaluar el estimador propuesto se incluye un ejemplo numérico.

Palabras clave: muestreo sucesivo, fracción de apareamiento, estimador de razón-producto, eficiencia relativa.

Introducción

Un aspecto a destacar en las encuestas continuas es la estructura de la muestra en cada ocasión. Existen varias posibilidades:

1. Extraer una nueva muestra en cada ocasión (muestreo *repetido*)
2. Utilizar la misma muestra en todas las ocasiones (muestreo *panel*)
3. Realizar un reemplazamiento parcial de unidades de una ocasión a otra (muestreo en *ocasiones sucesivas*, o también llamado muestreo *rotativo* cuando los elementos tienen restringido el número de etapas en las que van a formar parte de la muestra, como es el caso de la Encuesta de Población Activa, de periodicidad trimestral, y de la mayoría de las encuestas elaboradas por el Instituto Nacional de Estadística español (INE).

Las circunstancias de la encuesta y las características que se quieran estimar son determinantes para elegir el tipo de diseño muestral más adecuado.

En las encuestas muestrales es frecuente la necesidad de estimar algún parámetro poblacional en intervalos regulares de tiempo. Si existe una relación entre el valor de un elemento de la población en un período de tiempo, y el valor del mismo elemento en el período siguiente, entonces es posible emplear la información contenida en la muestra del período precedente, para mejorar la estimación actual del parámetro poblacional. En este sentido, para que sea posible utilizar la información muestral precedente, se debe obtener la muestra de manera que los elementos muestrales en los dos períodos sucesivos tengan algunos elementos comunes.

Algunos motivos por los que conviene utilizar el reemplazamiento parcial de unidades de la muestra son:

1. Reduce los costes, ya que utilizar una muestra completamente nueva en cada ocasión puede resultar excesivamente costoso.
2. Aumenta la precisión de los estimadores.
3. La permanencia indefinida de las mismas unidades en la muestra puede crear problemas y reducir la eficiencia de los estimadores. Por ejemplo, en las encuestas familiares de tipo *panel* se incrementan los sesgos en las estimaciones debido a la falta de colaboracíon de algunas familias que pertenecen al panel de hogares.

Así, el INE utiliza principalmente encuestas de muestreo *rotativo* debido a que presenta ventajas de las dos encuestas anteriores (*repetidas* y *panel*).

La teoría sobre muestreo *sucesivo* desarrollada hasta el momento va dirigida a obtener el estimador óptimo combinando dos estimadores de las medias: un estimador indirecto de doble muestreo de la parte apareada de la muestra, y un estimador simple de la media de la parte no apareada.

Con frecuencia disponemos de información que ha sido obtenida en la ocasión anterior sobre una variable auxiliar x que se encuentra positivamente correlacionada

con y, la variable objeto de estudio. En este contexto se ha demostrado que el estimador combinado que utiliza un estimador de razón para la parte apareada de la muestra tiene menor varianza que el estimador usual $\bar{y}$ siempre que $\rho > (C_x/2C_y)$ (Sen, Sellers y Smith, 1975).

Otras veces se dispone de la información proporcionada por una variable z que se encuentra negativamente correlacionada con y. En este caso, también se ha demostrado que el estimador óptimo que combina un estimador producto de doble muestreo para la parte apareada de la muestra y una media muestral simple de la parte no apareada, tiene menor varianza que el estimador usual $\bar{y}$ siempre que $\rho < (-C_x/2C_y)$ (Artés, Rueda y Arcos, 1998).

Para cubrir un amplio rango de situaciones prácticas, en este artículo proponemos un estimador que es aplicable cuando x y z están positiva y negativamente correlacionadas con y, respectivamente (donde la variable auxiliar x suele ser el valor de y en la ocasión anterior). Así, generalizamos la teoría en muestreo sucesivo para construir el estimador óptimo de la media en la segunda ocasión utilizando un estimador de razón-producto de doble muestreo para la parte apareada de la muestra, y una media simple basada en una muestra aleatoria de la parte no apareada en la segunda ocasión.

La teoría ha sido aplicada para proporcionar estimaciones más precisas de las variables analizadas en un estudio sobre hábitos de salud y nivel de condición física en escolares llevado a cabo en los colegios de Almería capital.

Teoría

Supongamos que las muestras son de tamaño n en ambas ocasiones, que se utiliza muestreo aleatorio simple y que el tamaño de la población N es suficientemente grande como para poder ignorar el factor de corrección por finitud.

Sea una muestra de tamaño n seleccionada en la primera ocasión de una población de tamaño N. Se dispone de información acerca de dos variables auxiliares x y z, cuyas medias denotamos por $\bar{x}$ y $\bar{z}$. Sea una muestra aleatoria simple de tamaño m submuestreada de las n unidades, que se retiene para la segunda ocasión (muestra apareada), y las restantes $u = n - m$ unidades son reemplazadas por una nueva selección del universo $N - m$ que resulta después de omitir las m unidades. En la segunda ocasión se considera la variable de interés y, que suponemos está correlacionada positivamente con x y negativamente con z.

Notación

m	tamaño muestral de aquellas unidades cuestionadas en ambas ocasiones (muestra apareada)

$u = n\text{-}m$	tamaño muestral de aquellas unidades cuestionadas sólo en la segunda ocasión (muestra no apareada)
$\bar{x}_m, \bar{z}_m\left(\bar{y}_m\right)$	media muestral apareada en la primera (segunda) ocasión estimando $\bar{X}, \bar{Z}\left(\bar{Y}\right)$
$\bar{y}_u$	media muestral no apareada en la segunda ocasión estimando $\bar{Y}$
R_1	$\bar{Y} / \bar{X}$
R_2	$\bar{Y} / \bar{Z}$
Δ_1	$C_x / C_y = \bar{Y} S_x / \bar{X} S_y$
Δ_2	$C_z / C_y = \bar{Y} S_z / \bar{Z} S_y$
ρ	coeficiente de correlación lineal de Pearson
Z_1	$\Delta_1 (2\rho_{xy} - \Delta_1)$
Z_2	$\Delta_2 (2\rho_{yz} + \Delta_2)$
p	m / n , fracción de apareamiento

El método Razón-Producto de estimación

Las partes apareada (m unidades) y no apareada (u unidades) de la muestra en la segunda ocasión proporcionan estimadores independientes ($\bar{y}_m$ y $\bar{y}_u$) de la media poblacional en la segunda ocasión ($\bar{Y}$).

Para la parte apareada podemos obtener un estimador mejor, $\bar{y}'_m$, para la media poblacional, $\bar{Y}$, utilizando el estimador de razón-producto de Singh (1967) bajo la técnica de doble muestreo, dado por

$$\bar{y}'_m = \bar{y}_m \frac{\bar{x}}{\bar{x}_m} \frac{\bar{z}_m}{\bar{z}}$$

Aplicando un razonamiento análogo al que hace Cochran (1977) para el estimador de razón, obtenemos su varianza:

$$V\left(\bar{y}'_m\right) = \frac{S_y^2}{n} + \left(\frac{1}{m} - \frac{1}{n}\right)\left(S_y^2 + S_x^2 R_1^2 + S_z^2 R_2^2 - 2S_{xy} R_1 + 2S_{yz} R_2 - 2S_{xz} R_1 R_2\right) =$$

$$= \frac{S_y^2}{m} + \frac{S_y^2}{n} \frac{u}{n-u}\left(\Delta_1^2 + \Delta_2^2 - 2\rho_{xy}\Delta_1 + 2\rho_{yz}\Delta_2 - 2\rho_{xz}\Delta_1\Delta_2\right) =$$

$$= \frac{S_y^2}{m} + \frac{S_y^2}{n}\frac{u}{n-u}\left(Z_1 - Z_2 + 2\rho_{xz}\Delta_1\Delta_2\right) \qquad (1)$$

Se puede obtener un estimador de la varianza sustituyendo en (1) los parámetros poblacionales S_y^2, Z_1, Z_2, ρ_{xz}, Δ_1 y Δ_2 por sus correspondientes estimadores muestrales.

Puesto que el estimador directo $\bar{y}_m$ basado en las m unidades tiene varianza

$$V(\bar{y}_m) = \frac{S_y^2}{m}$$

deducimos que $\bar{y}'_m$ es más preciso que $\bar{y}_m$ siempre que $Z_1 - Z_2 + 2\rho_{xz}\Delta_1\Delta_2 \geq 0$

Entonces, combinando los estimadores $\bar{y}_m$ y $\bar{y}_u$ (independientes entre sí) con pesos ω y $(1-\omega)$, respectivamente, lo que se obtiene es un estimador $\bar{y}_{2rp}$ de la media de la población en la segunda ocasión $\bar{Y}$ dado por:

$$\bar{y}_{2rp} = \omega\bar{y}'_m + (1-\omega)\bar{y}_u$$

y de aquí

$$V(\bar{y}_{2rp}) = \omega^2 V(\bar{y}'_m) + (1-\omega)^2 V(\bar{y}u)$$

Obtenemos el mejor estimador de $\bar{Y}$ en la segunda ocasión utilizando los valores de ω que minimicen $V(\bar{y}_{2rp})$

$$\omega_{opt} = \frac{V(\bar{y}_u)}{V(\bar{y}_u) + V(\bar{y}'_m)}$$

Como ya se sabe, $V(\bar{y}_u)$ viene dada por

$$V(\bar{y}_u) = \frac{S_y^2}{u}$$

Por tanto, sustituyendo en la expresión de la varianza tenemos que

$$V_{\min}(\bar{y}_{2rp}) = \frac{V(\bar{y}'_m)\,V(\bar{y}_u)}{V(\bar{y}'_m) + V(\bar{y}_u)} = \frac{S_y^2}{n}\frac{1-qA}{1-q^2A}$$

donde $A = Z_1 - Z_2 + 2\rho_{xz}\Delta_1\Delta_2$.

El valor óptimo de u se obtiene minimizando con respecto a la variación en u, y viene dado por

$$\frac{u}{n} = \frac{1-\sqrt{1-A}}{A}$$

o, lo que es lo mismo, la fracción del apareamiento óptimo vale

$$p_{opt} = 1 - \frac{1-\sqrt{1-A}}{A}$$

Sin embargo, si se considera el estimador usual de la media de la población en la segunda ocasión, $\bar{y}$, que es la media muestral basada sólo en las n unidades muestrales de dicha ocasión, su varianza toma la siguiente expresión

$$V(\bar{y}_u) = \frac{S_y^2}{n}$$

Por tanto, la ganancia en precisión, G, de $\bar{y}_{2rp}$ sobre $\bar{y}$ viene dada por

$$G = \frac{V(\bar{y}) - V(\bar{y}_{2rp})}{V(\bar{y}_{2rp})} = \frac{p(1-p)A}{1-(1-p)A}$$

Por definición $p \leq 1$. Si $p = 1$ (apareamiento total) ó $p = 0$ (reemplazamos toda la muestra), la ganancia vale cero. Para cualquier otro valor de p, obtendremos una ganancia positiva siempre que se verifique $A = Z_1 - Z_2 + 2\rho_{xz}\Delta_1\Delta_2 \geq 0$.

Comparación de eficiencias entre estimadores indirectos

Estimador de Razón

Se ha estudiado la precisión del estimador combinado de razón-producto con aquel que utiliza un estimador de razón para la parte apareada de la muestra, a partir de sus varianzas

$$V_{\min}(\bar{y}_{2r}) - V_{\min}(\bar{y}_{2rp}) \geq 0$$

siempre que se verifique

$$Z_2 - 2\rho_{xz}\Delta_1\Delta_2 \leq 0$$

lo que demuestra que el estimador combinado basado en un estimador de razón-producto de la parte apareada de la muestra y una media simple de la parte no apareada, $\bar{y}_{2rp}$, es más preciso que el correspondiente estimador que utiliza un estimador de razón para la muestra apareada, $\bar{y}_{2r}$, sujetos a la anterior condición.

Estimador Producto

Si se considera ahora el estimador combinado que utiliza un estimador producto para la parte apareada de la muestra, $\bar{y}_{2p}$, haciendo uso sólo de la variable auxiliar z correlacionada negativamente con la variable de interés y, y lo comparamos con $\bar{y}_{2rp}$, obtenemos que

$$V_{\min}(\bar{y}_{2p}) - V_{\min}(\bar{y}_{2rp}) \geq 0$$

siempre que se verifique

$$Z_1 + 2\rho_{xz}\,\Delta_1\Delta_2 \geq 0$$

lo que indica que, bajo dicha condición, el estimador que combina un estimador de razón-producto para la parte apareada de la muestra y una media simple de la parte no apareada, $\bar{y}_{2rp}$, es más eficiente que el correspondiente estimador que utiliza un estimador de producto para la parte apareada, $\bar{y}_{2p}$.

Caso especial

Para el caso especial

$$\rho_{xz} = \rho_1 = -\rho_{yz}$$
$$\rho_{xz} = \rho_0$$
$$C_x = C_y = C_z = C$$

tenemos que

$$V(\bar{y}'_m) = = \frac{S_y^2}{m} - \frac{S_y^2}{n}\frac{u}{n-u}(4\rho_1 + 2\rho_0 - 2)$$

y por tanto

$$V_{\min}(\bar{y}_{2rp}) = \frac{S_y^2}{n}\frac{1-qA}{1-q^2A}$$

donde $A = 4\rho_1 + 2\rho_0 - 2$.

El valor óptimo para u, en este caso, viene dado por

$$\frac{u}{n} = \frac{1-\sqrt{3-4\rho_1-2\rho_0}}{4\rho_1+2\rho_0-2}$$

La ganancia en precisión del estimador combinado, $\bar{y}_{2rp}$, sobre el estimador indirecto, $\bar{y}$, se puede obtener mediante la siguiente ecuación

$$G = \frac{p(1-p)(4\rho_1 + 2\rho_0 - 2)}{1-(1-p)(4\rho_1 + 2\rho_0 - 2)}$$

A partir de la expresión anterior deducimos que, para cualquier valor de p ($0<p<1$), el estimador combinado que utiliza un estimador de razón-producto para la parte apareada de la muestra, $\bar{y}_{2rp}$, es más preciso que el estimador usual $\bar{y}$ siempre que

$$\rho_1 + \frac{\rho_0}{2} \geq \frac{1}{2}$$

Por tanto, se puede concluir que la ganancia en precisión de $\bar{y}_{2rp}$ sobre $\bar{y}$ será óptima cuanto mayor sea la dependencia entre las variables auxiliares x e y con la variable objeto de estudio y (ρ_1 crece) y, al mismo tiempo, cuanto más incorrelacionadas estén x y z entre sí (ρ_0 decrece).

Estudio empírico

Para evaluar el buen funcionamiento del método propuesto se han utilizado los datos recogidos en una investigación sobre hábitos saludables y nivel de condición física. Dicho estudio se ha llevado a cabo sobre una población de escolares de 6º de Educación Primaria en los colegios de Almería capital durante los meses de Abril y Junio de 1998.

Se ha pretendido desarrollar un plan de muestreo que proporcione estimadores más precisos de las variables estudiadas. Dicho plan se ha basado en el principio del muestreo *sucesivo* de la misma población, y consistió en dos conjuntos de muestras aleatorias independientes: 1) una muestra de 131 escolares seleccionados, en la primera ocasión (Abril de 1998), entre los 2.211 escolares que formaban la población, y 2) una segunda muestra de 197 escolares seleccionada, en la segunda ocasión (Junio de 1998) entre los 2.080 escolares que no formaron parte de la muestra apareada.

A cada niño de la muestra se le administró un cuestionario sobre hábitos saludables, y se evaluó el nivel de condición física mediante determinados tests y medidas antropométricas.

Para el propósito del presente estudio hemos considerado la estimación del componente *endomorfo* (y, una de las múltiples variables implicadas en la investigación) en la segunda ocasión, tomando como variables auxiliares al índice de masa corporal (x) y al *volumen máximo de oxígeno* (z) de la primera ocasión. El procedimiento de estimación ha consistido en combinar los estimadores de las dos muestras independientes de escolares: $\bar{y}_m$ y $\bar{y}_u$.

Los datos muestrales sobre el número de escolares y parámetros obtenidos en las dos ocasiones han sido los siguientes:

- Primera ocasión (Abril 98): gran muestra de n=328.
- Segunda ocasión (Junio 98): muestra apareada m=131, muestra no apareada u=197.

$\hat{\sigma}_y = 1{,}54$ $\quad \bar{y} = 3{,}67$ $\quad \hat{\rho}_{xy} = 0{,}71$

$\hat{\sigma}_x = 3{,}71$ $\quad \bar{x} = 21{,}37$ $\quad \hat{\rho}_{xz} = -0{,}20$

$\hat{\sigma}_z = 6{,}87$ $\quad \bar{z} = 39{,}4$ $\quad \hat{\rho}_{yz} = -0{,}56$

A partir de los datos obtenemos que

$$\hat{V}_{\min}(\bar{y}_{2rp}) = 0{,}8\,\frac{S_y^2}{n} < \frac{S_y^2}{n} = \hat{V}(\bar{y})$$

lo que supone un 24,31% de ganancia en precisión del estimador propuesto sobre el estimador usual.

Se ha calculado también la fracción de apareamiento óptimo

$$p_{\text{opt}} = 37{,}8\%$$

Además, en la tabla 1 podemos comprobar cómo el estimador combinado basado en un estimador de razón-producto de la parte apareada de la muestra y una media simple de la parte no apareada, $\bar{y}_{2rp}$, es más preciso que el correspondiente estimador que utiliza un estimador de razón para la muestra apareada, $\bar{y}_{2r}$, y aquel que utiliza un estimador producto de la parta apareada, $\bar{y}_{2p}$. En la última columna se muestra la ganancia en eficiencia alcanzada por los distintos estimadores.

Tabla 1: *Comparación de eficiencias entre estimadores*

Estimadores	**Variable auxiliar**	**Varianza**	**Precisión sobre $\bar{y}$**
1. Directo $\bar{y}$	ninguna	S_y^2/n	
2. Producto $\bar{y}_{2p}$	z	0,92 S_y^2/n	8,42%
3. Razón $\bar{y}_{2r}$	x	0,89 S_y^2/n	12,98%
4. Razón-producto $\bar{y}_{2rp}$	x y z	0,8 S_y^2/n	24,31%

Referencias

Artés, E.; Rueda, M. y Arcos, A. (1998) Succesive Sampling using a Product Estimate. *Applied Sciences and the Environment Computational Mechanics Publications*, 85-90.

Biradar, R.S. y Singh, H.P. (1992) A Note on an Almost Unbiased Ratio–cum– Product Estimator. *Metron*, 50, 249-255.

Casimiro, A.J. (1999) Comparación, evolución y relación de hábitos saludables y nivel de condición física–salud en escolares, entre final de Educación Primaria (12 años) y final de Educación Secundaria Obligatoria (16 años), *Tesis Doctoral*, Universidad de Granada.

Cochran, W.G. (1977) *Sampling Techniques*, third edition. John Wiley & Sons, New York.

Rao, P.S.R.S. y Mudholkar, G.S. (1967) Generalized Multivariate Estimator for the Mean of Finite Populations. *Journal of the American Statistical Association*, 62, 1009-1012.

Rao, P.S.R.S. (1988) Ratio and Regression Estimators. En P.S.R.S. Rao y Krishnaiah (Eds.) *Handbook of Statistics,6, Sampling*. 449-468.

Sen, A.R.; Sellers, S. y Smith, G.E.J. (1975) The Use of a Ratio Estimate in Succesive Sampling. *Biometrics*, 31, 673-683.

Singh, M.P. (1967) Ratio–cum–Product Method of Estimation. *Metrika*, 12, 34-42.