Metodología de Encuestas
Vol 2, Núm 2, 2000, 261-272

ESTIMADORES DE RAZÓN CON DATOS FALTANTES

María del Mar Rueda García
Universidad de Granada

Silvia González Aguilera
Universidad de Jaén

RESUMEN

Los estimadores de razón se presentan como una estrategia muy aceptable cuando se cuenta con la información de una variable auxiliar. No obstante, en las investigaciones con encuestas es muy frecuente encontrar ausencias de medida para determinados puntos de la matriz resultante de datos. Estas ausencias pueden afectar tanto a la variable objetivo como a la auxiliar.

En este trabajo se proponen tres clases de estimadores de razón que utilizan la información que proporcionan las unidades de la muestra en la que existe pérdida de datos en cualquiera de las dos variables de interés, comparando su comportamiento con los estimadores de razón basados sólo en las unidades muestrales con la información completa de todas las variables. Los resultados indican que la estimación que considera los registros parciales además de la información completa, mejora la que se realiza únicamente con esta última.

Palabras clave: estimador de razón, datos faltantes.

Introducción

El muestreo de una población finita está dirigido a obtener información acerca de alguna o algunas características de ésta a partir, no de la población entera, sino de una parte de ella llamada muestra, debido a la imposibilidad de obtener información de todas las unidades poblacionales, ya sea por falta de recursos económicos, porque la población es excesivamente grande o simplemente por conveniencia. Apoyándonos en las unidades muestrales debemos proponer estimadores sobre dichas características de forma tal que proporcionen la información más eficiente.

La eficiencia de los métodos se suele medir en términos de la precisión, siendo el objetivo encontrar estimaciones cuyos errores de muestreo sean pequeños.

La mayoría de los métodos de muestreo más sencillos y utilizados, consideran estimadores que utilizan sólo los valores observados de la característica objeto de estudio. Sin embargo es frecuente que la variable objeto de estudio, y, esté altamente relacionada con una característica auxiliar, x, cuyos datos están disponibles o son muy fáciles de obtener para todos los elementos de la población. En esta situación, esta información auxiliar se puede utilizar para aumentar la precisión de las estimaciones mediante dos caminos: bien, modificando el diseño muestral, o bien, proponiendo estimadores más complejos, pero con menor error de muestreo, que reciben el nombre de estimadores indirectos.

Entre estos métodos se encuentra el método de estimación de razón que permite obtener estimadores de la media o el total poblacional más precisos, basándose en el comportamiento de la razón de las medias en la muestra. Además, hay ocasiones en las cuales el propio parámetro que se está interesado en estimar es, en sí mismo, un cociente de medias o totales poblacionales.

En la mayor parte de los desarrollos concernientes al uso de información auxiliar en la estimación de parámetros, se asume que todas las observaciones de las unidades seleccionadas en la muestra están disponibles. En la práctica esto no siempre es así, de hecho es muy común que en los estudios de mercado, encuestas de opinión, investigaciones socioeconómicas y otros experimentos científicos, se produzca una pérdida de observaciones. En estas circunstancias, los procedimientos tradicionales para deducir inferencias no pueden ser aplicados fácilmente.

La razón de medias poblacionales se estima de forma convencional como la razón entre las correspondientes medias muestrales. Este procedimiento de estimación no puede aplicarse cuando algunas de las observaciones no están disponibles. La alternativa es despreciar las unidades para las cuales hay pérdida de datos en alguna de las dos variables, restringiéndose así el tamaño de la muestra considerablemente y generándose, posiblemente, sesgo en las estimaciones (Morales, 2000).

El uso de las unidades en las que hay pérdida de los datos ya fue utilizada por Sing y Joarder (1998) para estimar la varianza a través de estimadores de razón.

Estos autores consideran que la información, tanto de la variable principal como de la variable auxiliar, no se puede obtener para ciertas unidades de la muestra.

En este trabajo vamos a considerar la situación en la que hay algunas observaciones para las cuales no están disponibles los valores en alguna de las dos variables, de forma que la pérdida de datos ocurre para ambas características por separado, pero no simultáneamente.

Teniendo en cuenta esto, definiremos tres clases de estimadores alternativos: la razón entre las medias muestrales calculadas utilizando sólo las observaciones completas. La primera clase de estimadores usará las observaciones incompletas en las que el valor de x está disponible, además de las observaciones completas. La segunda clase utiliza las observaciones incompletas en las que y está disponible, además de las observaciones completas. Por último, la tercera clase está basada en todas las observaciones, tanto completas como incompletas, generalizando a las clases anteriores, y pudiéndose aplicar bajo cualquier diseño muestral.

Al final del trabajo se realizará una comparación de estos estimadores en diversas poblaciones, mediante procedimientos de simulación.

Estimadores de razón

Consideremos una población U de N unidades en las cuales hay definido un diseño muestral $d = (S_d, p_d)$ mediante el cual se obtiene una muestra $s \in S_d$. Para dicha muestra, que supondremos de tamaño fijo n, se observan los valores de dos variables: $(y_i, x_i), i = 1, \ldots, n$.

Sea $R = \dfrac{Y}{X}$

el parámetro que se desea estimar. Asumiremos que sólo están disponibles en la muestra un conjunto de n-p-q observaciones completas:

$(y_1, x_1), \ldots, (y_{n-p-q}, x_{n-p-q})$

También están disponibles los valores de x para p unidades de la muestra, para las cuales las correspondientes observaciones de la variable y han desaparecido. Además se tiene un conjunto de q observaciones de la variable y cuyos correspondientes valores de x no están disponibles. Supondremos p y q números enteros verificando $0 < p, q < \dfrac{n}{2}$.

La estructura de los datos se puede visualizar en la tabla 1.

Para mayor simplicidad vamos a separar las unidades de la muestra s en tres conjuntos disjuntos:

$s_1 = \{i \in s / x_i \text{ e } y_i \text{ están disponibles}\}$

$s_2 = \{i \in s / x_i \text{ está disponible pero } y_i \text{ no lo está}\}$

$s_3 = \{i \in s / y_i \text{ está disponible pero } x_i \text{ no lo está}\}$

Tabla 1: *Estructura de datos con información ausente.*

Datos Muestrales	
Variable y	Variable x
y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_{n-p-q}	x_{n-p-q}
Faltante	$x_{n-p-q+1}$
Faltante	$x_{n-p-q+2}$
$\vdots$	$\vdots$
Faltante	x_{n-q}
y_{n-q+1}	Faltante
y_{n-q+2}	Faltante
$\vdots$	$\vdots$
y_n	Faltante

El estimador usual para R es el cociente entre los estimadores de Horvitz-Thompson basados en las n-p-q unidades de la muestra que están completas:

$$r_1 = \frac{\hat{y}'_{HT}}{\hat{x}'_{HT}} = \frac{\sum_{i \in s_1} \frac{y_i}{\pi_i}}{\sum_{i \in s_1} \frac{x_i}{\pi_i}}$$

Siendo π_i la probabilidad que tiene la unidad i de pertenecer a la muestra. Si los valores de p y q son grandes, el estimador r_1 estará basado en una muestra de tamaño pequeño y es de esperar por ello que disminuya considerablemente su precisión.

Ahora bien, si existen valores disponibles para la variable x, es lógico utilizar también estos datos para la estimación del total X.

Proponemos pues la siguiente clase de estimadores:

$$r_2 = \frac{\hat{y}'_{HT}}{a\hat{x}^2_{HT} + (1-a)\hat{x}'_{HT}} = \frac{\sum_{i \in s_1} \frac{y_i}{\pi_i}}{a\sum_{j \in s_2} \frac{x_j}{\pi_j} + (1-a)\sum_{i \in s_1} \frac{x_i}{\pi_i}}$$

basados en n - q unidades muestrales.

Análogamente, podemos utilizar los datos de s_2 para estimar mejor Y, y definir así los estimadores:

$$r_3 = \frac{b\hat{y}^3_{HT} + (1-b)\hat{y}'_{HT}}{\hat{x}'_{HT}}$$

siendo $\hat{y}^3_{HT} = \sum_{k \in s_3} \frac{y_k}{\pi_k}$, $0 \le b \le 1$, basados en n - p unidades muestrales.

Por último si consideramos todos los datos disponibles, podemos construir la siguiente clase de estimadores:

$$r_4 = \frac{b\hat{y}^3_{HT} + (1-b)\hat{y}'_{HT}}{a\hat{x}^2_{HT} + (1-a)x'_{HT}}$$

que engloba como casos particulares a r_1, r_2 y r_3. La idea de utilizar los datos de s_2 y s_3 para construir estimadores de R alternativos al usual ya fue considerada por Toutenburg y Srivastava (1998) en muestreo aleatorio simple, quienes proponen tres estimadores basados en las medias muestrales ponderadas por los tamaños relativos, y comparan su comportamiento en cuanto a eficiencia.

Estimadores óptimos

El paso siguiente es buscar en cada clase de estimadores r_2, r_3 y r_4, aquel estimador que tenga un mejor comportamiento. Como usualmente se hace en poblaciones finitas, la elección del *mejor estimador* se realiza en función de que minimice el error de estimación, es decir minimice el error cuadrático medio (evidentemente los estimadores no tienen por qué ser insesgados).

Utilizando el método de aproximación de Taylor, se llega, después de varios cálculos un poco tediosos, a las aproximaciones lineales de los errores cuadráticos medios, cuya minimización es simple, y se pueden obtener así los estimadores óptimos en cada clase:

Para r_2 se obtiene varianza mínima para

$$a = \frac{-(-2R^2 V(\hat{x}'_{HT}) + 2R^2 \operatorname{cov}(\hat{x}^2_{HT}, \hat{x}'_{HT}) - 2R \operatorname{cov}(\hat{y}'_{HT}, \hat{x}^2_{HT}) + 2R \operatorname{cov}(\hat{y}'_{HT}, \hat{x}'_{HT}))}{2R^2 (V(\hat{x}^2_{HT}) + V(\hat{x}'_{HT}) - 2\operatorname{cov}(\hat{x}^2_{HT}, \hat{x}'_{HT}))}$$

para r_3 se obtiene varianza mínima con

$$b = \frac{-(-2V(\hat{y}'_{HT}) + 2\operatorname{cov}(\hat{y}^3_{HT}, \hat{y}'_{HT}) - 2R \operatorname{cov}(\hat{y}^3_{HT}, \hat{x}'_{HT}) + 2R \operatorname{cov}(\hat{y}'_{HT}, \hat{x}'_{HT}))}{2(V(\hat{y}^3_{HT}) + V(\hat{y}'_{HT}) - 2\operatorname{cov}(\hat{y}^3_{HT}, \hat{y}'_{HT}))}$$

y se obtiene la varianza mínima de r_4 para

$$a = \frac{-C + (C'B - \frac{C}{A}A'B)/(B' - BA'/A)}{A} \quad \text{y} \quad b = \frac{-C' + \frac{C}{A}A'}{B' - BA'/A},$$

siendo

$$A = 2R^2V(\hat{x}^2_{HT}) + 2R^2V(\hat{x}'_{HT}) - 4R^2\,\mathrm{cov}(\hat{x}^2_{HT}, \hat{x}'_{HT})$$

$$B = -2R\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) + 2R\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{x}'_{HT}) + 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}^2_{HT}) - 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}'_{HT})$$

$$C = -2R^2V(\hat{x}'_{HT}) + 2R^2\,\mathrm{cov}(\hat{x}^2_{HT}, \hat{x}'_{HT}) - 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}^2_{HT}) + 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}'_{HT})$$

$$A' = -2R\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) + 2R\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{x}'_{HT}) + 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}^2_{HT}) - 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}'_{HT})$$

$$B' = 2V(\hat{y}^3_{HT}) + 2V(\hat{y}'_{HT}) - 4\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{y}'_{HT})$$

$$C' = -2V(\hat{y}'_{HT}) + 2\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{y}'_{HT}) - 2R\,\mathrm{cov}(\hat{y}^3_{HT}, \hat{x}'_{HT}) + 2R\,\mathrm{cov}(\hat{y}'_{HT}, \hat{x}'_{HT})$$

Desgraciadamente, estos valores óptimos dependen de las varianzas y covarianzas teóricas entre los estimadores de Horvitz-Thompson, que son desconocidas en general, pero pueden estimarse a partir de sus estimadores correspondientes una vez obtenida la muestra, o bien se pueden dar aproximaciones basadas en técnicas de replicación como bootstrap, semimuestras, grupos aleatorios,.... (véase p.e. Wolter, 1985).

Estas aproximaciones permiten obtener valores aproximados ($\tilde{a}$ de a, $\tilde{b}$ de b, $\tilde{a}'$ de a' y $\tilde{b}'$ de b', construyendo los estimadores:

$$\tilde{r}_2 = \frac{\hat{y}^1_{HT}}{\tilde{a}\hat{x}^2_{HT} + (1-\tilde{a})\hat{x}^1_{HT}}, \quad \tilde{r}_3 = \frac{\tilde{b}\hat{y}^3_{HT} + (1-\tilde{b})\hat{y}'_{HT}}{\hat{x}'_{HT}} \quad \text{y} \quad \tilde{r}_4 = \frac{\tilde{b}\hat{y}^3_{HT} + (1-\tilde{b})\hat{y}'_{HT}}{\tilde{a}\hat{x}^2_{HT} + (1-\tilde{a})x^1_{HT}}$$

cuyos errores de muestreo es de esperar que estén próximos a los obtenidos a partir de los mínimos teóricos.

Comparación de estimadores

Por su proceso de construcción, la clase de estimadores r_4 engloba el resto de estimadores r_1, r_2 y r_3, por tanto el estimador obtenido, minimizando el error en la clase r_4, es el mejor en el sentido de mínimo error, de entre todos los estimadores

considerados. Ahora bien, para cualesquiera valores particulares y distintos de a y b, la comparación no es simple para cualquier diseño muestral. Además, los estimadores no tienen por qué mejorar obligatoriamente a r_1.

En efecto, si consideramos el diseño muestral más simple, el muestreo aleatorio simple y suponemos

$$\frac{n-p}{2} > q\,,\quad \frac{n-q}{2} > p \text{ y } p > q$$

se puede calcular la precisión relativa entre los estimadores propuestos, frente al estimador que no usa los datos parcialmente faltantes, cuya expresión viene dada por:

$$\frac{ECM(r_4)}{ECM(r_1)} = \frac{(-b^2+2b+((1-f_q)/q)(1-b^2)/((1-f_{p+q})/p+q))C_y^2}{C_y^2+C_x^2-2C_{yx}} +$$

$$+\frac{(-a^2+2a+((1-f_p)/p)(1-a^2)/(\frac{1}{n-p-q}-\frac{1}{N}))C_x^2}{C_y^2+C_x^2-2C_{yx}} +$$

$$-\frac{2(1-a+ab+(\frac{1}{p}-\frac{1}{N})(1-b)(1-a)/(\frac{1}{n-p-q}-\frac{1}{N}))C_{yx}}{C_y^2+C_x^2-2C_{yx}}$$

Sin más que hacer b=0 y a=0, se obtienen las expresiones correspondientes a r_2 y r_3.

Como se observa, la ganancia o pérdida de precisión respecto al estimador de razón usual depende de los coeficientes a y b, del número de unidades faltantes p y q, así como de las características de la población C_x^2, C_y^2 y C_{xy}.

Así pues, no se pueden dar reglas generales sobre el comportamiento de estos estimadores que sirvan para cualquier diseño muestral y para cualquier población, teniendo que hacer la comparación para cada situación concreta.

A continuación, y a título ilustrativo, se presenta un estudio realizado para algunos estimadores particulares mediante una simulación en una población real usada inicialmente por Chambers y Dunstan (1993). La población está constituida por 430 granjas con 50 o más cabezas de ganado, que se encuestaron en 1988 en un estudio del Australian Bureau of Agricultural and Resource Economics. Las variables que se consideran son los ingresos obtenidos del ganado, y el número de cabezas de ganado en cada granja.

Vamos a considerar el caso particular de que la muestra sea elegida mediante muestreo aleatorio simple, y vamos a calcular los estimadores propuestos por Toutenburg y Srivastava (1998).

$$r_2^T = \frac{(n-q)\sum_{i\in s_1} \frac{y_i}{n-p-q_i}}{p\sum_{j\in s_2} \frac{x_j}{p} + (n-p-q)\sum_{i\in s_1} \frac{x_i}{n-p-q}}$$

$$r_3^T = \frac{q\bar{y}^3 + (n-p-q)\bar{y}^1}{(n-.p)\bar{x}^1}$$

$$r_4^T = \left(\frac{n-p}{n-q}\right)\frac{q\bar{y}^3 + (n-p-q)\bar{y}^1}{p\bar{x}^2 + (n-p-q)\bar{x}^1}$$

siendo

$\bar{y}^1$ la media de la variable y para los valores de la muestra en que x e y están disponibles.

$\bar{y}^3$ la media de la variable y para los valores de la muestra en que y está disponible pero no x.

$\bar{x}^1$ la media de la variable x para los valores de la muestra en que x e y están disponibles.

$\bar{x}^2$ la media de la variable x para los valores de la muestra en que x está disponible, pero y no.

Que corresponden a las clases r_2, r_3 y r_4 para los valores:

$$a = \frac{p}{n-q} \qquad b = \frac{q}{n-p}$$

La metodología seguida para la simulación es la siguiente:

En primer lugar, tomamos dos números, P y Q, que representan el número de observaciones faltantes en la variable x y el número de observaciones faltantes en la variable y, respectivamente. Una vez seleccionados estos dos números, generamos P

números aleatorios de una distribución uniforme entre 1 y el tamaño de la población dividido por dos y tomaremos las unidades correspondientes de la población como faltantes en la variable x.

Para la variable y repetimos el mismo proceso, siendo el número de observaciones faltantes Q.

Una vez elegidos los datos faltantes, tomamos mil muestras aleatorias simples de tamaño n. Para cada una de estas muestras hemos calculado los estimadores propuestos por Toutenburg y Srivastava .

A partir de estas mil muestras calculamos la esperanza y el error de cada uno de los estimadores

Este proceso se repite para distintos tamaños muestrales y diferentes valores de P y Q.

Los resultados correspondientes a las simulaciones para los valores de P y Q, y los tamaños muestrales seleccionados se muestran a continuación, en las tablas 2 (n=25), 3 (n=50) y 4 (n=75).

Tabla 2: *Simulación para un tamaño muestral 25.*

P \ Q	50	75	100	125
50		1 1.1267904 0.7075668 0.7600087	1 0.9931740 0.6318240 0.5655190	1 0.9997311 1.2289105 0.9371253
75	1 1.2814836 0.7643606 0.9852199		1 1.1732596 0.7234598 0-7734959	
100	1 1.0571857 0.7349423 0.8186630	1 1.5186040 0.6716093 0.8314007		
125	1 1.5288649 0.7710113 1.1791655			

Los valores de r_i se expresan en función de r_1, con el objetivo de facilitar la comparación entre los resultados obtenidos en cada caso. Así, el contenido de cada casilla o celda de las tablas 2 a 4 se refiere a:

$$\frac{r_1}{r_1} = 1, \quad \frac{r_2}{r_1}, \quad \frac{r_3}{r_1}, \quad \frac{r_4}{r_1}$$

Tabla 3: *Simulación para un tamaño muestral de 50.*

P \ Q	50	75	100	150
50		1 1.0954380 0.6713275 0.7280987	1 0.9799891 0.6838822 0.6167044	1 0.9888860 0.7042624 0.5516985
75	1 1.2212249 0.7572224 0.9551143		1 1.2527764 0.7193210 0.7776260	
100	1 1.4739508 0.7319812 1.1366769	1 1.3277785 0.7337735 0.9411156		
150	1 1.7792880 0.7203138 1.4261163			

Tabla 4: *Simulación para un tamaño muestral de 75.*

P \ Q	50	75	100	175
50	1 1.0775461 0.9311378 0.9332164		1 1.0747715 0.6504965 0.6652470	1 1.0402962 0.4779632 0.3746396
75	1 1.2831378 0.7974720 1.0527250		1 1.2645905 0.5656187 0.6634295	
100	1 1.5018593 0.7658340 1.1526000	1 1.4155833 0.7567496 1.0063108		
175	1 2.4689325 0.7056509 1.8666235			

Como podemos observar, r_3 y r_4 mejoran al estimador r_1 en la mayor parte de las simulaciones (r_3 en el 96% de los casos y r_4 en el 71% de los casos). El comportamiento de r_2, sin embargo, no es tan bueno. Ello, creemos, se debe por una parte a que la correlación entre las variables no es muy elevada, y por otra a que la variabilidad de la variable x es bastante mayor que la de la variable y. Esto explica que el mejor funcionamiento lo tenga el estimador r_3 pues el estimador de la media de y basado en los datos parcialmente faltantes tiene un buen comportamiento con lo cual la estimación de la media poblacional de y se realiza con más precisión cuando se utilizan también los datos parcialmente faltantes que cuando sólo se usan los datos completos.

En cuanto al número de datos faltantes en x y en y (que es aleatorio), no hemos podido establecer cuál es su relación exacta con el comportamiento de los estimadores. Sí se detecta que el comportamiento de r_4 es mejor para tamaños de Q grandes, es decir, cuando faltan bastantes datos en la población de la variable principal (incluso mejora a r_3), mientras que si Q es pequeño y P grande, no mejora al estimador r_1.

Por último, observar que estos estimadores no son los óptimos en la clase, y pese a ello han producido, en la mayor parte de los casos, una disminución en los errores de muestreo.

Conclusiones

Los estimadores de las clases r_2 , r_3 y r_4, que utilizan los datos parcialmente faltantes, pueden utilizarse como alternativa al estimador usual de la razón, el cual presenta problemas si el número de datos parcialmente faltantes es elevado. Si se puede determinar el óptimo de la clase r_4 para el diseño muestral que se esté usando en la población que se estudia, podremos construir un estimador que garantice que tiene mayor precisión que el estimador usual. Si no se puede determinar dicho óptimo, se pueden probar con distintos estimadores de las clases, comprobando cuáles tienen un buen comportamiento para la población dada, pudiendo obtener así reducciones considerables en el error de estimación.

La ganancia en precisión se asegura conforme los estimadores se aproximan más a los óptimos determinados en las clases. Por tanto, un camino importante a seguir será la determinación de buenas aproximaciones de estos óptimos para los distintos tipos de diseños muestrales.

Estas consideraciones son igualmente válidas en el caso de querer estimar otros parámetros, como medias y totales a través de los estimadores de razón correspondientes. En este caso, se podría extender el estudio al caso de que se utilice más de una variable auxiliar.

Referencias

Morales, L. (2000). El efecto de la no respuesta parcial en el análisis de datos de encuesta: una comparación entre la eliminación de observaciones y la imputación múltiple. *Metodología de Encuestas*, 2(2), 217-238.

Särndal, C. E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer –Verlag, New York.

Singh, S., Joarder, A.M. (1998). Estimation of finite population variance using non-response in survey sampling. *Metrika* 47, 241-249.

Sukhatme, P. V., Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*. Second edition, Asia Publishing House, Bombay, India.

Toutenburg, H., Srivastava, V., K. (1998). Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika* 48, 177-187.

Wolter, K.M. (1985). *Introduction to variance estimation*. Springer–Verlag. New York.