ESTADÍSTICA ESPAÑOLA
Vol. 37, Núm. 140, 1995, págs. 511 a 521

Mejora de estimadores de razón y regresión de la varianza poblacional

por
ANTONIO ARCOS CEBRIÁN, MARIA DEL MAR RUEDA GARCÍA
Departamento de Estadística e I. O.
Universidad de Granada

EVA ARTÉS RODRÍGUEZ
Departamento de Estadística e I. O.
Universidad de Almería

RESUMEN

En este trabajo proponemos un método para incorporar información adicional de varias variables auxiliares, en la fase de estimación, del parámetro varianza de una población finita. El estimador propuesto puede verse como una extensión del método multivariante clásico propuesto por *Isaki* (1983).

Palabras clave: estimadores de razón, estimadores de regresión, varianza poblacional.

Clasificación AMS: 62D05.

1. INTRODUCCION

Es bien conocido que el uso de información auxiliar cuando se muestrea en una población finita, puede mejorar notablemente la precisión en la estimación de parámetros poblacionales como la media o el total poblacionales. Son muchas las

situaciones prácticas en las que se dispone de información adicional proporcionada por varias variables y en tales casos es importante saber incorporar en la fase de estimación la información disponible. En este trabajo proponemos un nuevo método que persigue ese fin cuando se trata de estimar la varianza de una población finita.

Supongamos una población finita de tamaño N en cuyas unidades hay definidas dos características: y y x, siendo y la característica de interés y x la característica auxiliar. Por Y_j y X_j, respectivamente, denotamos los valores de y y x en la unidad j. Supongamos una muestra aleatoria simple de tamaño n y que el tamaño de la población es grande como para poder omitir el factor de corrección por finitud. Sean s_y^2 y s_x^2 estimadores insesgados de las varianzas poblacionales S_y^2 y S_x^2, respectivamente,

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_j - \overline{Y}\right)^2$$

donde $\overline{Y}$ es la media poblacional de la variable y.

El método clásico para construir un estimador tipo razón (*Isaki*, 1983) para S_y^2 es mediante el estimador:

$$\hat{S}_I^2 = s_y^2 \frac{S_x^2}{s_x^2},$$

siendo este estimador más preciso que el estimador directo, s_y^2, bajo ciertas condiciones (ver *Isaki*,1983).

En este trabajo extendemos el estimador $\hat{S}_I^2$ de forma que el estimador definido es más preciso que $\hat{S}_I^2$ y también que s_y^2.

2. EL ESTIMADOR PROPUESTO Y SU ERROR CUADRATICO MEDIO

La extensión del estimador $\hat{S}_I^2$ que proponemos, $\hat{S}_E^2$, es la siguiente:

$$\hat{S}_E^2 = w s_y^2 \frac{S_x^2}{s_x^2} + (1-w)s_y^2.$$

Claramente, $\hat{S}_E^2$ es $\hat{S}_I^2$ si $w=1$ y $\hat{S}_E^2$ es s_y^2, si $w=0$. Así, una elección óptima del coeficiente w, producirá un estimador $\hat{S}_E^2$ con menor varianza que las varianzas de los estimadores $\hat{S}_I^2$, y s_y^2.

El estimador $\hat{S}_E^2$, es sesgado. La precisión del estimador se ha de medir por su error cuadrático medio y la constante w ha de elegirse minimizando éste. Denotando

$$t_0 = s_y^2 - S_y^2, \quad t_1 = s_y^2 - s_y^2 \frac{S_x^2}{s_x^2},$$

se obtiene $\hat{S}_E^2 - S_y^2 = t_0 - wt_1$,

$$ECM\left(\hat{S}_E^2\right) = E\left(t_0^2\right) - 2wE(t_0t_1) + w^2E\left(t_1^2\right) = f(w).$$

El punto crítico de la función cuadrática y convexa f coincide con su mínimo, con lo que el coeficiente w que minimiza $ECM\left(\hat{S}_E^2\right)$ es $w_0 = \dfrac{E(t_0t_1)}{E\left(t_1^2\right)}$ y

$$\min ECM\left(\hat{S}_E^2\right) = E\left(t_0^2\right) - \frac{E(t_0t_1)^2}{E\left(t_1^2\right)},$$

mínimo que al ser no negativo puede expresarse como

$$\min ECM\left(\hat{S}_E^2\right) = E\left(t_0^2\right)\left(1-\eta^2\right),$$

donde

$$\eta^2 = \frac{E(t_0t_1)^2}{E\left(t_0^2\right)E\left(t_1^2\right)},$$

con $\eta^2 \leq 1$. Obsérvese que η es como el coeficiente de correlación entre t_0 y t_1 sólo que reemplazando $V(t_1)$ por $E\left(t_1^2\right)$ (ya que s_y^2 es insesgado para S_y^2).

3. APROXIMACION DE PRIMER ORDEN DE ECM $\left(\hat{S}_E^2\right)$

Sean $e_1 = \dfrac{s_x^2 - S_x^2}{S_x^2}$ y $e_0 = \dfrac{s_y^2 - S_y^2}{S_y^2}$. Entonces, $s_y^2 = S_y^2(1+e_0)$ y $t_0 = s_y^2 - S_y^2 = S_y^2 e_0$.

Además, en un primer orden de aproximación,

$$t_1 = s_y^2 - s_y^2 \frac{S_x^2}{s_x^2} = s_y^2\left(1-(1+e_1)^{-1}\right) \cong S_y^2(1+e_0)e_1.$$

Así, $t_1^2 \cong S_y^4 e_1^2$, $t_0 t_1 \cong S_y^4 e_0 e_1$, y

$$E\left(t_1^2\right) \cong S_y^4\left[\frac{V\left(s_x^2\right)}{S_x^4} + O\left(n^{-2}\right)\right],$$

$$E(t_0 t_1) \cong S_y^4\left[\frac{Cov\left(s_x^2, s_y^2\right)}{S_y^2 S_x^2} + O\left(n^{-2}\right)\right].$$

Por tanto, como $t_0^2 = S_y^4 e_0^2$, $E\left(t_0^2\right) = V\left(s_y^2\right)$, y

$$\min ECM\left(\hat{S}_E^2\right) \cong V\left(s_y^2\right) - \frac{Cov^2\left(s_x^2, s_y^2\right)}{V\left(s_x^2\right)} = V\left(s_y^2\right)\left[1-\rho^2\left(s_x^2, s_y^2\right)\right]$$

y ya que en muestreo aleatorio simple (ver *Kendall* y *Stuart* (1977)),

$$V\left(s_x^2\right) = \frac{1}{n} S_x^4(\beta_2(x)-1) + S_x^4 O\left(n^{-2}\right),$$

$$V\left(s_y^2\right) = \frac{1}{n} S_y^4(\beta_2(y)-1) + S_y^4 O\left(n^{-2}\right),$$

$$Cov\left(s_x^2, s_y^2\right) = \frac{1}{n} S_x^2 S_y^2(\theta(y,x)-1) + S_x^2 S_y^2 O\left(n^{-2}\right),$$

donde

$$\beta_2(y) = \frac{\mu_{40}}{\mu_{20}^2}, \beta_2(x) = \frac{\mu_{04}}{\mu_{02}^2}, \theta_2(y,x) = \frac{\mu_{22}}{\mu_{20}\mu_{02}},$$

con

$$\mu_{rs} = \frac{1}{N}\sum\left(Y_j - \overline{Y}\right)^r \sum\left(X_j - \overline{X}\right)^s,$$

obtenemos

$$\min \mathrm{ECM}\left(\hat{S}_E^2\right) \cong \frac{1}{n} S_y^4 \left[\beta_2(y) - 1 - \frac{(\theta(y,x)-1)^2}{(\beta_2(x)-1)}\right]$$

y el *w* óptimo que minimiza $\mathrm{ECM}\left(\hat{S}_E^2\right)$ es

$$w_0 = \frac{(\theta\,(y,x)-1)}{(\beta_2(x)-1)}$$

El método de regresión propuesto por *Isaki* (1983) estima la varianza poblacional mediante el estimador $\hat{S}_{reg}^2 = s_y^2 + b\left(S_x^2 - s_x^2\right)$ donde *b* es una constante, y la varianza mínima del estimador óptimo obtenido coincide con la dada en (2). Así, el estimador propuesto dado en (1), con el coeficiente óptimo (3) y el estimador de regresión óptimo tienen la misma precisión.

En el caso particular de que la distribución conjunta (y,x) tenga los mismos momentos hasta el orden cuatro que una normal bivariante, los coeficientes, *b* y *w*, que dan los estimadores óptimos de regresión, $\hat{S}_{reg}^2$, y del estimador que proponemos, $\hat{S}_E^2$, son

$$b_0 = \frac{S_y^2}{S_x^2}\rho^2 \quad \text{y} \quad w_0 = \rho^2,$$

donde ρ es el coeficiente de correlación entre *y* y *x*. Así, los estimadores óptimos adquieren la forma

$$\hat{S}_{reg^{opt}}^2 = s_y^2 + \frac{S_y^2}{S_x^2}\rho^2\left(S_x^2 - s_x^2\right)$$

y

$$\hat{S}_{E^{opt}}^2 = \rho^2 s_y^2 \frac{S_x^2}{s_x^2} + \left(1-\rho^2\right)s_y^2 \ .$$

De esta forma, para tamaños de muestra grandes, el disponer de información sobre el coeficiente de correlación ρ (información disponible en muchas situaciones prácticas) permite construir un estimador para la varianza poblacional tan preciso como el estimador de regresión y más preciso que los estimadores de razón, $\hat{S}_I^2$, y directo, s_y^2. Además, con sólo esta información no es posible computar el estimador de regresión óptimo pues se requiere conocer también la varianza poblacional de la variable de interés, S_y^2, mientras que el estimador $\hat{S}^2_{E^{opt}}$ sí se puede evaluar.

4. EXTENSION AL CASO MULTIVARIANTE

Supongamos en esta sección que en las *N* unidades de la población finita hay definidas *k*+1 características *y*, $x_1,\ldots, x_k$. Supongamos una muestra aleatoria simple de tamaño *n* y sean $s^2_{x_i}$ estimadores insesgados de las varianzas poblacionales $S^2_{x_i}$, $i=1,\ldots, k$.

Isaki (1983) propuso (de forma similar al método para estimar la media poblacional propuesto por Olkin (1958)) un estimador multivariante tipo razón para S_y^2:

$$\hat{S}^2_{MTR} = \sum_{i=1}^{k} w_i s_y^2 \frac{S^2_{x_i}}{s^2_{x_i}}, \quad \text{con} \quad \sum_{i=1}^{k} w_i = 1.$$

La comparación de éste estimador, en cuanto a precisión, respecto al estimador directo, s_y^2, sólo es posible imponiendo restricciones a los momentos de la distribución conjunta de las variables $y, x_1,\ldots, x_k$.

En esta sección extendemos el estimador $\hat{S}^2_{MTR}$ y el estimador propuesto en la sección 2, $\hat{S}^2_E$, de forma que el estimador obtenido es más preciso que $\hat{S}^2_{MTR}$ y también que s_y^2. Definimos

$$\hat{S}^2_{ME} = \sum_{i=1}^{k} w_i s_y^2 \frac{S^2_{x_i}}{s^2_{x_i}} + w_{k+1} s_y^2, \quad \text{con} \quad \sum_{i=1}^{k+1} w_i = 1.$$

$\hat{S}^2_{ME}$ coincide con $\hat{S}^2_{MTR}$ si $w_{k+1}=0$, $\hat{S}^2_{ME}$ coincide con s_y^2 si $w_i=0$ para $i=1, \ldots k$, y $\hat{S}^2_{ME}$ coincide con $\hat{S}^2_E$ si $k=1$. Por tanto, eligiendo los coeficientes w_i óptimos, se obtendrá un estimador $\hat{S}^2_{ME}$ más preciso que $\hat{S}^2_{MTR}$ y s_y^2.

Sean $t_i = s_y^2 - s_y^2 \frac{S_{x_i}^2}{s_{x_i}^2}, i = 1,\ldots,k$, w el vector columna $(w_1,\ldots,w_k)'$, M la matriz $k \times k$, $M = \left(E(t_i t_j)\right)$, i,j=1,..., k, y m el vector columna $\left(E(t_0 t_1),\ldots,E(t_{0i} t_k)\right)'$

Como $w_{k+1} = 1 - \sum_{i=1}^{k} w_i$, obtenemos $\hat{S}_{ME}^2 - S_y^2 = t_0 - \sum_{i=1}^{k} w_i t_i$, y

$$ECM\left(\hat{S}_{ME}^2\right) = E\left(t_0^2\right) - 2m'w + w'Mw = f(w).$$

Los coeficientes w óptimos que minimizan $ECM\left(\hat{S}_{ME}^2\right)$ son $w_0 = M^{-1}m$ y

$$\min ECM\left(\hat{S}_{ME}^2\right) = E\left(t_0^2\right) - m'M^{-1}m.$$

Además, en un primer orden de aproximación, denotando $e_i = s_{x_i}^2 - \frac{S_{x_i}^2}{s_{x_i}^2}$, obtenemos $t_i \cong S_y^2(1+e_0)e_i$ y $t_i t_l \cong S_y^4 e_i e_l, i,l = 0,\ldots,k$ (para i=l=0, $t_0^2 = S_y^2 e_0^2$). Entonces, usando una notación análoga a la de la sección 2, obtenemos

$$E(t_i t_l) \cong \frac{1}{n} S_y^4 \left[\theta_{x_i x_l} - 1\right], \quad i,l = 0,\ldots,k, \quad (\text{con} \quad x_0 = y),$$

y llamando $A_0 = \left(\theta_{yx_1} - 1,\ldots,\theta_{yx_k} - 1\right)$ $A = A_{(k\times k)} = \left(\theta_{x_i x_l} - 1\right)$, i=l=1,...,k

$$\min ECM\left(\hat{S}_{ME}^2\right) = \frac{1}{n} S_y^4 \left[\beta_2(y) - 1 - A_0' A^{-1} A_0\right]$$

y obtenemos que los coeficientes óptimos son

$$w_0 = M^{-1} m A^{-1} A_0.$$

Así, el estimador $\hat{S}_{ME}^2$ es igual de eficiente que el estimador de regresión multivariante propuesto por (*Isaki*, 1983)

$$\hat{S}_{MR}^2 = s_y^2 + \sum_{i=1}^{k} B_i\left(S_{x_i}^2 - s_{x_i}^2\right),$$

y $\hat{S}^2_{ME}$ tiene la misma forma que un estimador de regresión multivariante salvo que $S^2_{x_i} - s^2_{x_i}$ se ha reemplazado por $s^2_y \frac{S^2_{x_i}}{s^2_{x_i}} - s^2_y$, *i=1,..., k*. Para verificar que el estimador $\hat{S}^2_{ME}$ es tan preciso como el estimador $\hat{S}^2_{MR}$, de (5), tenemos

$$ECM\left(\hat{S}^2_{ME}\right) = E\left(t_0^2\right) - 2m'w + w'Mw \equiv V\left(s^2_y\right) - 2A_0'w + w'Aw =$$

$$= V\left(s^2_y\right) + \sum_{i=1}^{k} w_i^2 V\left(s^2_{x_i}\right)\frac{S^4_y}{S^4_{x_i}} - 2\sum_{i=1}^{k} w_i Cov\left(s^2_{x_i}, s^2_y\right)\frac{S^2_y}{S^2_{x_i}} + \sum_{i\neq j}^{k} w_i w_j Cov\left(s^2_{x_i}, s^2_{x_j}\right)\frac{S^4_y}{S^2_{x_i}S^2_{x_j}}$$

La varianza exacta del estimador de regresión multivariante $\hat{S}^2_{MR}$ es

$$V\left(\hat{S}^2_{MR}\right) = V\left(s^2_y\right) + \sum_{i=1}^{k} B_i^2 V\left(s^2_{x_i}\right) - 2\sum_{i=1}^{k} B_i Cov\left(s^2_{x_i}, s^2_y\right) + \sum_{i\neq j}^{k} B_i B_j Cov\left(s^2_{x_i}, s^2_{x_j}\right).$$

Por tanto, como los coeficientes B_i óptimos, B_i^0 son las soluciones del sistema

$$B_i V\left(s^2_{x_i}\right) - Cov\left(s^2_{x_i}, s^2_y\right) + \sum_{i\neq j}^{k} B_i B_j Cov\left(s^2_{x_i}, s^2_{x_j}\right) = 0, \quad i = 1,\ldots,k,$$

denotando $\chi_i = w_i R_i$, $R_i = S^2_y / S^2_{x_i}$, para *i=1,..., k*, concluimos que los coeficientes χ_i óptimos, χ_i^0, y los coeficientes B_i óptimos son iguales:

$$\chi_i^0 = w_i^0 R_i = B_i^0,$$

y entonces

$$\min ECM\left(\hat{S}^2_{ME}\right) \equiv \min V\left(\hat{S}^2_{MR}\right).$$

Ahora mostramos cómo el estimador propuesto, $\hat{S}^2_{ME}$, presenta ciertas ventajas frente al estimador multivariante de regresión $\hat{S}^2_{MR}$. Si $(y, x_1,..., x_k)$ tiene los mismos momentos que una ley multivariante normal $1\times(k+1)$ hasta el orden ocho y se asume que el coeficiente de correlación poblacional entre todas las variables es el mismo, es decir, $\rho_{yx_i} = \rho_{x_i x_i} = \rho$, $i,j = 1,\ldots,k$ ver también *Isaki* 1983), entonces el estimador multivariante tipo razón, $\hat{S}^2_{MRT}$, es óptimo para

$$w_1^0 = w_2^0 = \cdots = w_k^0 = \frac{1}{k},$$

el estimador multivariante de regresión, $\hat{S}^2_{MR}$, es óptimo para

$$B_i^0 = \frac{\rho^2}{1+(k-1)\rho^2}\frac{S_y^2}{S_{x_i}^2}, \quad i = 1,\ldots,k,$$

y el propuesto, para los coeficientes

$$w_1^0 = w_2^0 = \cdots = w_k^0 = \frac{\rho^2}{1+(k-1)\rho^2}.$$

Así, los estimadores multivariantes óptimos obtenidos por el método de regresión y por el método que proponemos adoptan la forma

$$\hat{S}^2_{MR_{opt}} = s_y^2 + \frac{\rho^2}{1+(k-1)\rho^2}\sum_{i=1}^{k}\frac{S_y^2}{S^2_{x_i\,i}}\left(S_{x_i}^2 - s_{x_i}^2\right)$$

y

$$\hat{S}^2_{ME_{opt}} = \frac{\rho^2}{1+(k-1)\rho^2}\sum_{i=1}^{k} s_y^2\frac{S_{x_i}^2}{s_{x_i}^2} + \left(1 - \frac{\rho^2}{1+(k-1)\rho^2}\right)s_y^2,$$

respectivamente.

De las expresiones anteriores se deduce cómo el método que proponemos mejora ambos estimadores. En primer lugar, de (4) se deduce que $\hat{S}^2_{ME}$ es más preciso que s_y^2 y $\hat{S}^2_{MR}$. En segundo lugar, bajo las restricciones impuestas en el modelo propuesto por *Isaki*, $\hat{S}^2_{ME}$ mejora a $\hat{S}^2_{MR}$ en el sentido de no requerir más que disponer de información acerca del coeficiente de correlación común a todas las variables para ser computado exactamente.

5. APLICACIONES Y COMENTARIOS

Los resultados de este trabajo pueden ser aplicados en cualquier situación en la que un estimador tipo razón pueda ser usado. Una de estas situaciones es la siguiente: Un censo de 1993 muestra que el salario medio de una población es $\overline{X}$. Además, en el mismo censo, como medida de la dispersión de los salarios con respecto a su media (y en cierta forma de la distribución de los salarios), muestra

una varianza poblacional S_x^2. En mitad de una campaña electoral del año 1995 se tiene interés en estimar la varianza de los salarios S_y^2, para observar en qué medida ha variado. Mediante una muestra aleatoria simple es posible estimar esta varianza mediante el método directo, s_y^2. Si además se observa en las unidades de la muestra seleccionada el salario de 1993, es posible estimar la varianza S_y^2 mediante $\hat{S}_I^2 = s_y^2 \frac{S_x^2}{s_x^2}$. El estimador que proponemos es más preciso en la estimación de S_y^2 que ambos.

El estimador que proponemos en este trabajo, $\hat{S}_{ME}^2$, puede verse como un caso especial del estimador de razón de *Isaki*, con sólo introducir una variable auxiliar artificial x_{k+1} que cumpla $s_{x_{k+1}}^2 = S_{x_{k+1}}^2$. Entonces, $\frac{S_{x_{k+1}}^2}{s_{x_{k+1}}^2} = 1$ y el estimador de *Isaki*, $\hat{S}_{MTR}^2$, con k+1 variables auxiliares coincide con el propuesto $\hat{S}_{ME}^2$ para k variables auxiliares.

Rueda y *Arcos* (1996) consideran un estimador de la varianza poblacional obtenido siguiendo el método repetido de sustitución, cuasi-insesgado y asintóticamente igual de eficiente que el estimador de regresión, y con las mismas buenas propiedades que el propuesto en este trabajo en poblaciones normales bivariantes.

Cuando el parámetro de interés es la media poblacional de la variable *y*, *Tankou* y *Dharmadhikari* (1989) consideran, de forma similar, un estimador como extensión del estimador multivariante tipo razón propuesto por Olkin (1958) para estimar la media poblacional .

Finalmente, notar que aunque este trabajo se ha desarrollado suponiendomuestreo aleatorio simple y que el tamaño de la población sea grande como para poder prescindir del factor de corrección por finitud, el método puede aplicarse para construir estimadores tipo razón en diferentes diseños muestrales.

REFERENCIAS

ISAKI, C.T. (1983): «Variance Estimation Using Auxiliary Information» *Journal of the American Statistical Associaton*, 78, 117-123.

KENDALL, M., STUART, A. (1977): «The Advanced Theory os Statistics Vol 1 and 2», *Griffin London*

OLKIN, I. (1958): «Multivariate Ratio Estimator for Finite Population», *Biometrika*, 45, 145-165.

RUEDA, M., ARCOS, A. (1996): «Repeated Substitution Method: The Ratio Estimator for the Population Variance». *Metrika*, 43, 101-105.

TANKOU, V., DHARMADHIKARI, S. (1989): «Improvement of Ratio-Type Estimators». *Biometrical Journal*, 31, 795-802.

IMPROVEMENT OF RATIO AND REGRESSION TYPE ESTIMATORS FOR POPULATION VARIANCE

SUMMARY

This paper proposes a method for using multi-auxiliary information at the estimation stage of sample surveying when the parameter of interest is the variance of a finite population. The proposed estimator is an extension of the classical multivariate method of *Isaki* (1983).

Key Words and Phrases: Ratio estimators, regression estimators, population variance.

AMS Classification: 62D05.