MODELOS DE SUPERPOBLACION EN EL MUESTREO POR CUOTAS: ESTIMACION DE LA MEDIA POBLACIONAL

Marta Guijarro Garví

RESUMEN.— Se utiliza un modelo de análisis de la varianza como modelo de superpoblación en la estimación de la media poblacional mediante muestreo por cuotas, comprobándose que la expresión matricial del estimador basado en el modelo es la misma para los casos de uno y dos factores. En cada uno de estos contextos se estudia la insesgadez del estimador y se obtiene la expresión de su varianza.

1. INTRODUCCION

Básicamente, el muestreo por cuotas consiste en clasificar los N elementos de la población finita objeto de estudio según las distintas modalidades de uno o varios factores que representaremos por $i, j, \ldots, h$ $(i = 1, \ldots, I; j = 1, \ldots, J; \ldots; h = 1, \ldots, H)$. Así, $N_{ij\ldots h}$ expresa el número de unidades poblacionales pertenecientes a la casilla $(i, j, \ldots, h)$ de una tabla de contingencia múltiple, siendo $n_{ij\ldots h}$ el número de ellas que forman parte de la muestra de tamaño n.

En este trabajo supondremos que el muestreo es por cuotas marginales, es decir, tomaremos una única fracción de muestreo, f, para cada conjunto de cuotas:

$$n_{i\cdots} = fN_{i\cdots}; \; n_{\cdot j\cdots} = fN_{\cdot j\cdots}; \; \ldots; \; n_{\cdots h} = fN_{\cdots h}$$

indicando por la suma en todas las modalidades de la característica correspondiente.

La falta de aleatoriedad en la selección de la muestra y la consecuente imposibilidad de realizar inferencias nos llevan a la necesidad de utilizar enfoques basados en modelos, entre los que se encuentra el de los modelos de superpoblaciones que adoptaremos en este trabajo.

Admitiremos, por tanto, que la población finita ha sido generada como muestra aleatoria de una superpoblación infinita, esto es, que el valor de la variable de interés en cada individuo de la población es una realización de una variable aleatoria, formulando la distribución conjunta del total de variables mediante un modelo de análisis de la varianza.

En la sección 2 obtendremos, tanto la expresión matricial del estimador de la media poblacional basado en un modelo con un único factor, como la de su varianza. En la sección 3 realizaremos idéntico análisis para el caso de dos factores[1].

2. MODELOS DE ANALISIS DE LA VARIANZA CON UN FACTOR

La población finita formada por N unidades está dividida en I cuotas, determinadas por las correspondientes modalidades de un único factor o categoría. Denotaremos por N_i, el número de elementos de i-ésima cuota $(i = 1, \ldots, I)$.

El valor de la característica de interés en la unidad k-ésima de la i-ésima cuota, vendrá dado por y_{ik} con $i = 1, \ldots, I$, $k = 1, \ldots, N_i$.

Siguiendo el enfoque proporcionado por los modelos de superpoblación, el vector de valores poblacionales, $y = (y_{11}, \ldots, y_{1N_1}, \ldots, y_{I1}, \ldots, y_{IN_I})'$ es una realización del vector de variables aleatorias $Y = (Y_{11}, \ldots, Y_{1N_1}, \ldots, Y_{I1}, \ldots, Y_{IN_I})'$, cuya distribución conjunta viene formulada a través de un modelo de análisis de la varianza, ξ, dado por:

$$Y = A\beta + \varepsilon \qquad [1]$$

donde A es una matriz de dimensión $N \times I$,

$$A = \begin{pmatrix} 1_{N1} & 0 & \ldots & 0 \\ 0 & 1_{N2} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1_{NI} \end{pmatrix}$$

con $1_{Ni} = (1, \ldots, 1)'$, N_i-vector unidad $(i = 1, \ldots, I)$, $\beta = (\beta_1, \ldots, \beta_I)'$, vector paramétrico de dimensión $I \times 1$ y $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1N_1}, \ldots, \varepsilon_{I1}, \ldots, \varepsilon_{IN_I})'$ vector

1 Deville (1991) considera modelos de análisis de la varianza aunque con hipótesis ligeramente distintas a las formuladas en este trabajo. Sin embargo, dicho autor no utiliza un enfoque matricial, más adecuado al contexto del muestreo en superpoblaciones.

aleatorio de dimensión $N \times 1$ verificando:

$$E_\xi(\varepsilon) = 0$$
$$E_\xi(\varepsilon\varepsilon') = \Sigma$$

donde $E_\xi(\cdot)$ denota la esperanza respecto al modelo y Σ es una matriz de dimensión $N \times N$ con la siguiente estructura:

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \Sigma_I \end{pmatrix}$$

siendo $\Sigma_i = diag\,(o_i^2, \dots, o_i^2)$ $N_i \times N_i$-*matriz diagonal* $(i = 1, \dots, I)$.

Sin pérdida de generalidad, listaremos las unidades de la muestra en primer lugar, teniendo en cuenta que ésta incluye n_i individuos de la categoría i $(i = 1, \dots, I)$. Tendremos, así, las siguientes particiones de Y, A, ε y 1, vector de unos de dimensión $N \times 1$:

$$Y = (Y'_s, Y'_r)'$$
$$A = \begin{pmatrix} A_s \\ A_r \end{pmatrix}$$
$$\varepsilon = (\varepsilon'_s, \varepsilon'_r)$$
$$1 = (1'_s, 1'_r)$$

donde

$$Y_s = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{In_I})'$$
$$Y_r = (Y_{11}, \dots, Y_{1N_1-n_1}, \dots, Y_{I1}, \dots, Y_{IN_I-n_I})'$$

$$A_s = \begin{pmatrix} 1_{n_1} & 0 & \dots & 0 \\ 0 & 1_{n_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1_{n_I} \end{pmatrix}$$

$$A_r = \begin{pmatrix} 1_{N_1-n_1} & 0 & \dots & 0 \\ 0 & 1_{N_2-n_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1_{N_I-n_I} \end{pmatrix}$$

$$\varepsilon_s = (\varepsilon_{11}, \dots, \varepsilon_{1n_1}, \dots, \varepsilon_{I1}, \dots, \varepsilon_{In_I})'$$

$$1_s = (1'_{n_1}, \dots, 1'_{n_I})'$$

Siguiendo a Royall (1970), el estimador de la media poblacional basado en el modelo, resulta de dividir por N la suma de los valores observados y la predicción de los no observados. En este caso:

$$e_R = N^{-1}\,(1'_s Y_s + 1'_r \hat{Y}_r) = N^{-1}\,(1'_s Y_s + 1'_r A_r \hat{\beta})$$

Dado que el estimador de Gauss-Markov de β_i es la media de los valores observados en la categoría i-ésima, es decir,

$$\hat{\beta} = \bar{y}_i = \left(\sum_{k=1}^{n_i} y_{ik} \right) \Big/ n_i \quad i = 1, \ldots, \ldots I$$

es estimador del vector paramétrico es $\hat{\beta} = (\bar{y}_1, \ldots, \bar{y}_I)'$.

Esto nos lleva a

$$1'_s Y_s = 1'_s A_s \hat{\beta}$$

con lo que, en este caso, el estimador predictivo de la media poblacional es

$$e_R = N^{-1}(1'_s A_s \hat{\beta} + 1'_r A_r \hat{\beta}) = N^{-1}(1'_s A_s + 1'_r A_r)\,\hat{\beta} = N^{-1} 1' A \hat{\beta}$$

2.1. Esperanza y varianza del estimulador predictivo

Veamos que el sesgo del estimador es cero. En efecto:

$$N^{-1}E_\xi(1'A\hat{\beta} - 1'Y) = N^{-1}\,E_\xi(1'A\hat{\beta} - 1'A\hat{\beta} - 1'\varepsilon) = N^{-1}[1'AE_\xi\,(\hat{\beta} - \beta) - 1'E_\xi\,(\varepsilon)] = 0$$

puesto que

$$E_\xi(\hat{\beta}_i - \beta_i) = E_\xi\,(\bar{\varepsilon}_i) = 0 \quad i = 1, \ldots, N$$

con

$$\bar{\varepsilon}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} e_{ik}$$

es decir, $\hat{\beta}$ es un estimador insesgado de β.

Por tanto, y considerando que

$$Var_\xi \left(\frac{1}{N}\, 1' A \beta \right) = \frac{1}{N^2}\, 1' AV\, ar_\xi(\beta)\, A' 1$$

bastará calcular $Var_\xi(\hat{\beta})$ para obtener la varianza del estimador:

$$Var_\xi(\hat{\beta}) = E_\xi[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \begin{pmatrix} \sigma_1^2/n_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \sigma_I^2/n_I \end{pmatrix}$$

ya que

$$E_\xi(\hat{\beta}_i - \beta_i)^2 \; E_\xi(\bar{\varepsilon}_i)^2 = \sigma_i^2/n_i \quad i = 1, \dots, I$$

Si denotamos por

$$V = \begin{pmatrix} \sigma_1^2/n_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \sigma_I^2/n_I \end{pmatrix}$$

tendremos:

$$Var_\xi(e_R) = N^{-2} 1' AV A' 1$$

Esta varianza puede ser estimada sustituyendo σ_I^2 por

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (Y_{ik} - \bar{y}_i)$$

3. MODELO DE ANALISIS DE LA VARIANZA CON DOS FACTORES

Supongamos, ahora, la influencia de dos factores sobre los valores poblacionales de la variable de interés. Denotaremos por $N_{ij}(i = 1, \dots, I; j = 1, \dots, J)$, el número de individuos de la población que poseen las modalidades i y j, es decir, que pertenecen a la casilla, (i, j) de la correspondiente tabla de contingencia múltiple. Así, el valor de la característica objeto de estudio, en la unidad k-ésima de la casilla (i, j) se representará por $y_{ijk}(k = 1, \dots, N_{ij})$.

El vector de valores poblacionales $y = (y_{111}, \dots, y_{11N_{11}}, \dots, y_{IJ_1}, \dots, y_{IJN_{IJ}})'$, es, en este caso, una realización del vector de variables aleatorias $Y = (Y_{111}, \dots, Y_{11N_{11}}, \dots, Y_{IJ_1}, \dots, Y_{IJN_{IJ}})'$ siendo el modelo de superpoblación el dado por la expresión [1], donde A es ahora de dimensión $N \times (I + J + 1)$:

$$A = \begin{pmatrix} 1_{N_{11}} & 1_{N_{11}} & \dots & 0 & 1_{N_{11}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{1J}} & 1_{N_{1J}} & \dots & 0 & 0 & \dots & 1_{N_{1J}} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{I1}} & 0 & \dots & 1_{N_{I1}} & 1_{N_{I1}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{IJ}} & 0 & \dots & 1_{N_{IJ}} & 0 & \dots & 1_{N_{IJ}} \end{pmatrix}$$

con $1N_{ij} = (1, \dots, 1)'$, N_{ij}-vector unidad ($i = 1, \dots, I$; $j = 1, \dots, J$), $\beta = (\mu, \alpha_1, \dots, \alpha_I, \gamma_1, \dots, \gamma_J)$, vector paramétrico de dimensión $(I + J + 1)\ x\ 1$ y $\varepsilon = (\varepsilon_{111}, \dots, \varepsilon_{11N_{11}}, \dots, \varepsilon_{1J_1}, \dots, \varepsilon_{1JN_{1J}})'$, vector aleatorio de dimensión $N\ x\ 1$, cuyas componentes son centradas, independientes y, además,

$$Var_{\xi}(\varepsilon_{ijk}) = \tau_i^2 + w_j^2 \qquad i = 1, \dots, I, \quad j = 1, \dots, J, \quad k = 1, \dots, N_{ij}$$

Consideraremos las mismas particiones de Y, A, ε y 1 empleadas en el caso de un único factor. En esta nueva situación tendremos:

$$Y_s = (Y_{111}, \dots, Y_{11n_{11}}, \dots, Y_{1J_1}, \dots, Y_{1Jn_{1J}})'$$
$$Y_r = (Y_{111}, \dots, Y_{11N_{11}-n_{11}}, \dots, Y_{1J_1}, \dots, Y_{1JN_{11}-n_{1J}})'$$

$$A_s = \begin{pmatrix} 1_{n_{11}} & 1_{n_{11}} & \dots & 0 & 1_{n_{11}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{n_{1J}} & 1_{n_{1J}} & \dots & 0 & 0 & \dots & 1_{n_{1J}} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{n_{I1}} & 0 & \dots & 1_{n_{I1}} & 1_{n_{I1}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{n_{IJ}} & 0 & \dots & 1_{n_{IJ}} & 0 & \dots & 1_{n_{IJ}} \end{pmatrix}$$

$$A_r = \begin{pmatrix} 1_{N_{11}-n_{11}} & 1_{N_{11}-n_{11}} & \dots & 0 & 1_{N_{11}-n_{11}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{1J}-n_{1J}} & 1_{N_{1J}-n_{1J}} & \dots & 0 & 0 & \dots & 1_{N_{1J}-n_{1J}} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{I1}-n_{I1}} & 0 & \dots & 1_{N_{I1}-n_{I1}} & 1_{N_{I1}-n_{I1}} & \dots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1_{N_{IJ}-n_{IJ}} & 0 & \dots & 1_{N_{IJ}-n_{IJ}} & 0 & \dots & 1_{N_{IJ}-n_{IJ}} \end{pmatrix}$$

$$\varepsilon_s = (\varepsilon_{111}, \dots, \varepsilon_{11n_{11}}, \dots, \varepsilon_{1J_1}, \dots, \varepsilon_{1Jn_{1J}})'$$
$$1_s = (1'_{n_{11}}, \dots, 1'_{n_{1J}})'$$

con $1_{n_{ij}}$ y $1_{N_{ij}-n_{ij}}$ vectores unidad de dimensiones n_{ij} y $N_{ij} - n_{ij}$, respectivamente.

Admitiremos que se cumplen las siguientes condiciones de identificabilidad del modelo:

$$\sum_j n_{ij}\,\alpha_i = 0 \text{ y } \sum_j n_{ij}\,\beta_j = 0$$

Utilizaremos el muestreo por cuotas marginales, con muestras de tamaño fijo $n = fN$, incluyendo $n_{i\cdot} = fN_{i\cdot}$ individuos por cada modalidad i del primer factor y $n_{\cdot j} = fN_{\cdot j}$ por cada j del segundo.

Al igual que en la situación presentada en la sección anterior, nuestro objetivo es calcular el estimador de Royall, basándonos en la predicción de los valores observados. Es decir,

$$e_R = \frac{1}{N}(1'_s Y_s + 1'_r \hat{Y}_r) = \frac{1}{N}(1'_s Y_s + 1'_r A_r \hat{\beta})$$

La estimación del vector paramétrico por mínimos cuadrados ordinarios nos lleva a:

$$\hat{\beta} = (\bar{y}, \bar{y}_{1\cdot} - \bar{y}, \ldots, \bar{y}_{I\cdot} - \bar{y}, \bar{y}_{\cdot 1} - \bar{y}, \ldots, \bar{y}_{\cdot J} - \bar{y})'$$

donde $\bar{y}$ es la media de los valores muestrales; $\bar{y}_{i\cdot}$ e $\bar{y}_{\cdot j}$ son las medias de las observaciones muestrales de las modalidades i-ésima y j-ésima del primer y segundo factor, respectivamente $(i = 1, \ldots, I; j = 1, \ldots, J)$.

Podemos obtener una expresión más sencilla del estimador predictivo teniendo en cuenta que,

$$1'_s A_s \hat{\beta} = n\bar{y} + \sum_i \sum_j n_{ij}(\bar{y}_{i\cdot} - \bar{y}) + \sum_i \sum_j (\bar{y}_{\cdot j} - \bar{y}) = n\bar{y} = 1'_s Y_s$$

Por tanto,

$$e_R = \frac{1}{N}(1'_s A_s \hat{\beta} + 1'_r A_r \hat{\beta}) = \frac{1}{N} 1' A \hat{\beta} \quad [2]$$

2 Operando obtenemos

$$1'_r A_r \hat{\beta} = (N - n)\,\bar{y}$$

con lo que el estimador de Royall coincide, en este caso, con $\bar{y}$.

3.1. ESPERANZA Y VARIANZA DEL ESTIMADOR PREDICTIVO

Siguiendo el mismo razonamiento que en el apartado 2.1, el estimador e_R es insesgado de la media poblacional, puesto que la independencia entre ambos factores nos lleva a que $\hat{\beta}$ es un estimador insesgado de β.

Con el fin de hallar la varianza del estimador, denotemos por *m* el vector de dimensión *N x 1* de componentes $m_{ijk} = E_\xi (Y_{ijk})$. A partir de él obtenemos, de manera habitual, el correspondiente vector m_s para los valores de la muestra, pudiendo, así, escribir:

$$\frac{1}{N^2} E_\xi (1'A\hat{\beta} - 1'Y)^2 = \frac{1}{N^2} E_\xi \left[\frac{N}{n} 1'_s (Y_s - m_s) - 1'(Y - m) \right]^2 =$$

$$\frac{1}{N^2} E_\xi \left(\frac{N}{n} 1'_s \varepsilon_s - 1'\varepsilon \right)^2 = \frac{1}{N^2} \left[E_\xi \left(\frac{N}{n} 1'_s \varepsilon_s \right)^2 + E_\xi (1'\varepsilon)^2 - 2E_{xi} \left(\frac{N}{n} 1'_s \varepsilon_s 1'\varepsilon \right) \right] =$$

$$\frac{1}{n^2} \sum_i \sum_j n_{ij} (\tau_i^2 + \omega_j^2) + \frac{1}{N^2} \sum_i \sum_j N_{ij} (\tau_i^2 + \omega_j^2) - \frac{2}{Nn} \sum_i \sum_j n_{ij} (\tau_i^2 + \omega_j^2)$$

Ahora bien, aplicando las condiciones del muestreo por cuotas marginales, deducimos que:

$$\sum_i \sum_j n_{ij} (\tau_i^2 + \omega_j^2) = \sum_i n_{i\cdot} \tau_j^2 + \sum_j n_{\cdot j} \omega_j^2 = \frac{n}{N} \left(\sum_i N_{i\cdot} \tau_j^2 + \sum_j N_{\cdot j} \omega_j^2 \right)$$

con lo que la varianza del estimador predictivo puede expresarse como:

$$\frac{1}{N^2} E_\xi (1'A\hat{\beta} - 1'Y)^2 = \frac{1}{Nn} (1 - f) \left(\sum_i N_{i\cdot} \tau_j^2 + \sum_j N_{\cdot j} \omega_j^2 \right)$$

Obsérvese que esta varianza depende sólo de las cuotas y no de la muestra; este hecho justifica, en cierta medida, el empleo de cuotas marginales.

BIBLIOGRAFIA

Deville, J. (1991): A theory of quota surveys. *Surbey Methodology* 17 n. 2, 163-181.

Gourieroux, C. (1981): Theory des sondages. *Economica.*

Herson, J. y Royall, R. M. (1973): Robust estimation in finite populations. *Journal of American Statistical Association* 68, 880-893.

Kendall, M. G. y Stuard, A. (1967): *The advanced theory of statistics.* New York: Hafner.

Rohatgi, V. K. (1976): *An introduction to probability theory and mathematical statistics*. New York: John Wiley.

Royall, R. M. (1970): On finite population sampling theory under certain linear regression models. *Bioetrika* 57, 377-387.

Tam, S. M. (1988b): Some results son robust estimation in finite population sampling. *Journal of American Statistical Association* 83, 242-248.

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