TRABAJOS DE ESTADISTICA
Vol. 5, núm. 2, 1990, pp. 73 a 81

DISEÑO MUESTRAL EN DOS FASES SOBRE UNA POBLACION FINITA: UNA PERSPECTIVA BAYESIANA

J. BASULTO SANTOS
Univ. de Sevilla
S. MURGUI IZQUIERDO
Univ. de Valencia
(Facultad de Ciencias Económicas y Empresariales)
Avda. Blasco Ibáñez, 30. 46010 Valencia.

RESUMEN

Se plantea la estimación de la media de una población finita bajo el supuesto de un modelo de superpoblación. Se considera la existencia de una variable auxiliar cuyo coste de observación es inferior al de interés y un diseño en dos fases. En la primera fase se observa únicamente la variable auxiliar, mientras que en la segunda se hace lo propio con la de interés. En el trabajo se determina el estimador óptimo, así como el tamaño y la muestra a seleccionar en cada fase.

Palabras claves: Modelo de superpoblación, Perspectiva bayesiana, Diseño muestral en dos fases.

ABSTRACT

The estimation of the mean in a finite population is considered with a superpopulation model. We use an auxiliary variable which has an observation cost inferior to the corresponding cost of the variable we want to study. The design is made in two phases. The auxiliary variable is observed first. In this essay the optimal estimator is determined as well as the samples to be selected in each phase.

Recibido: Diciembre 1989.
Revisado: Julio 1990.

1. INTRODUCCION

El objetivo propuesto en este trabajo es la construcción de un estimador para la media $\bar{y} = \sum_{i=1}^{N} y_i/N$, de los valores que toma una característica Y sobre un colectivo finito, cuyas unidades se encuentran identificadas por los enteros $1, 2, 3, ..., N$. Se supone para ello que las y_i son valores de variables aleatorias Y_i que verifican el modelo $Y_i = \alpha x_i + e_i x_i^{1/2}$, para $i = 1, ..., N$; siendo las e_i variables independientes e idénticamente distribuidas con media cero y varianza σ^2, α un parámetro desconocido y las x_i valores de variables positivas X_i independientes e idénticamente distribuidas con media μ y varianza σ_x^2.

En lo que sigue, se admitirá que el coste de observar cada valor x_i es una cantidad fija c_1 y que ésta es menor que el correspondiente coste c_2 de observar cada valor y_i. Este supuesto conduce a seleccionar la muestra en dos fases (Cochran, 1978). En la primera se elige una muestra s_1 de tamaño n y sobre sus unidades se observan únicamente los valores x_i. En la segunda se elige una muestra s_2 de tamaño m ($m \leqslant n$) de entre las unidades ya seleccionadas en s_1, observando en este caso los correspondientes valores y_i.

Bajo una perspectiva bayesiana, la incertidumbre asociada con las cantidades desconocidas del modelo debe ser descrita mediante distribuciones de probabilidad adecuadas. En la sección 2 se especifican éstas y en la sección 3 se resuelve el problema de estimación.

La determinación del diseño muestral exige conocer tanto los tamaños n y m de las muestras como la forma en que deben ser seleccionadas las unidades. Tales cuestiones son abordadas en las secciones 4 y 5.

2. ESPECIFICACION DEL MODELO DE SUPERPOBLACION

Reparametrizando el modelo en $h = 1/\sigma^2$ y $\delta = 1/\sigma_x^2$, supondremos que α, h, μ y δ son variables aleatorias, tales que los vectores (α, h) y (μ, δ) siguen distribuciones Normales-Gammas (DeGroot, 1970), independientes y con parámetros conocidos (α', h', a', b') y (μ', δ', u', v'), respectivamente.

Admitiendo ahora una distribución Normal tanto para las variables

e_i como X_i, las anteriores hipótesis suponen la intercambiabilidad de los vectores (X_i, Y_i) para $i = 1, ..., N$ (Sugden, 1978).

La especificación del modelo considerado permite deducir como distribución de la variable $\bar{Y} = \sum_{i=1}^{N} Y_i/N$, condicionada a las observaciones x_i, una t de Student con media $\alpha\bar{x}$, precisión

$$\frac{Nh'}{h'\bar{x} + N\bar{x}^2}\,\frac{a'}{b'}$$

y grados de libertad a', siendo $\bar{x} = \sum_{i=1}^{N} x_i/N$.

La función de densidad de tal distribución la representaremos por $f(\bar{y}/\vec{x})$.

3. ESTIMACION DE LA MEDIA POBLACIONAL

Seleccionada una muestra s_1 de n unidades y observados los datos $\vec{x}_{s_1} \equiv (x_i; i \in s_1)$, la media $\bar{x}$ puede descomponerse en la forma

$$\bar{x} = [n\bar{x}_s + (N - n)\bar{x}_{s'_1}]/N, \quad \text{donde } \bar{x}_{s_1} = \sum_{i \in s_1} x_i/n$$

es un factor conocido y $\bar{x}_{s'_1} = \sum_{i \notin s_1} x_i/(N-n)$ el valor de una variable $\bar{X}_{s'_1}$.

Se comprueba que la distribución de $\bar{X}_{s'_1}$ condicionada a los datos $\vec{x}_{s_1}$ es una t de Student de media

$$\mu'' = \frac{\delta'\mu' + n\bar{x}_{s_1}}{\delta' + n},$$

precisión

$$\frac{(N - n)\delta''}{N - n + \delta''}\,\frac{u''}{v''}$$

y grados de libertad u''. Siendo $\delta'' = \delta' + n$; $u'' = n$;

$$v'' = v' + nS^2_{s_1} + \frac{n\delta'(\bar{X}_{s_1} - \mu')^2}{\delta' + n}$$

y $S^2_{s'_1}$ la varianza de los datos $\vec{x}_{s_1}$.

La distribución de la media $\bar{Y}$ condicionada a los datos $\vec{x}_{s_1}$ puede expresarse a través de su función de densidad

$$f(\bar{y}/\vec{x}_{s_1}) = \int_R f(\bar{y}/\vec{x}) \cdot f(\bar{x}_{s'_1}/\vec{x}_{s_1})\, d\bar{x}_{s'_1} \tag{1}$$

por lo que se trata de la integral del producto de dos densidades t de Student.

Supongamos ahora que de entre las unidades que constituyen la muestra s_1 se ha seleccionado una muestra s_2 de tamaño m ($m \leqslant n$) y que sobre la misma se han observado los datos $\vec{y}_{s_2} \equiv (y_i\,; i \in S_2)$. La función de densidad de la variable

$$\bar{Y}_{s'_2} = \frac{\sum\limits_{i \notin s_2} Y_i}{(N-m)}$$

condicionada a los datos $\vec{x}_{s_1}$ y $\vec{y}_{s_2}$ puede expresarse mediante la integral

$$f(\bar{y}_{s'_2}/\vec{x}_{s_1}, \vec{y}_{s_2}) = \int_R f(\vec{y}_{s'_2}/\bar{x},\ \vec{y}_{s_2}) \cdot f(\bar{x}_{s'_1}/\vec{x}_{a_1})\, d\bar{x}_{s'_1} \tag{2}$$

donde $f(\bar{y}_{s'_2}/\bar{x},\ \vec{y}_{s_2})$ corresponde a una distribución t de Student con media

$$\alpha''[(n-m)\bar{x}_{s_1 - s_2} + (N-n)\bar{x}_{s'_1}],$$

precisión

$$\frac{(N-n)h''}{h''\bar{x}_{s'_2} + (N-m)\bar{x}^2_{s'_2}} \cdot \frac{a''}{b''}$$

y grados de libertad a''. Siendo

$$\alpha'' = \frac{h'' + m\bar{y}_{s_2}}{h' + m\bar{x}_{s_2}} \quad , \quad h'' = h' + m\bar{x}_{s_2} \quad , \quad a'' = a' + n \quad \text{y}$$

$$b'' = b' + \frac{1}{h' + m\bar{x}_{s_2}} (h'(\sum_{i \in s_2} (y_i - \alpha' x_i)^2/x_i) + (\sum_{i \in s_2} y_i^2/x_i) m\bar{x}_{s_2} - m^2 \bar{y}^2_{s_2}),$$

donde

$$\bar{x}_{s_2} = \sum_{i \in s_2} x_i/m,\ \bar{y}_{s_2} = \sum_{i \in s_2} y_i/m;\quad \bar{x}_{s_1 - s_2} = \sum_{i \in s_1 \cap s_2} x_i/(n-m)$$

y

$$\bar{X}_{s_2'} = (n - m)\bar{x}_{s_1 - s_2} + (N - n)\bar{x}_{s_1'})/(N - m).$$

La distribución de $\bar{y}$ condicionada al conjunto de todos los datos $\vec{x}_{s_1}$ y $\vec{y}_{s_2}$ se obtiene fácilmente teniendo en cuenta la descomposición $\bar{y} = (m\bar{y}_{s_2} + (N - m)\bar{y}_{s_2'})/N$ y la relación (2).

Como estimador de $\bar{y}$ puede adoptarse la media de la distribución $f(\bar{y}/\vec{x}_{s_1}, \vec{y}_{s_2})$ cuya expresión es

$$(m\bar{y}_{s_2} + \alpha''((n - m)\bar{x}_{s_1 - s_2} + (N - n)\mu''))/N.$$

Si se considera como distribución inicial para los parámetros una no-informativa como $f(\mu, \delta, \alpha, h) \propto \delta^{-1} \cdot h^{-1}$, la expresión anterior se reduce a $\dfrac{\bar{y}_{s_2}}{\bar{x}_{s_2}} \bar{x}_{s_1}$, que es el estimador de razón propuesto por Cochran (1978).

El error de la estimación viene expresado por la varianza de la distribución $f(\bar{y}/x_{s_1}, y_{s_2})$. Su valor se obtiene operando por condicionalidad, siendo el resultado

$$\frac{1}{N^2}\left(\frac{b''}{h''(a'' - 2)}(h''(n - m)\bar{x}_{s_1 - s_2} + h''(N - n)\mu'' + ((N - n)\mu'' + \right.$$

$$\left. + (n - m)\bar{x}_{s_1 - s_2})^2) + (N - n)^2\left(\alpha''^2 + \frac{b''}{h''(a'' - 2)}\right) \cdot \mathrm{Var}\,(\bar{X}_{s_1'}/\vec{x}_{s_1})\right),$$

siendo

$$\mathrm{Var}\,(\bar{X}_{s_1'}/\vec{x}_{s_1}) = \frac{v''}{u'' - 2} \cdot \frac{N - n + \delta''}{(N - n)\delta''}.$$

4. SELECCION DE LAS MUESTRAS

Suponiendo fijados los tamaños muestrales n y m, faltaría por resolver el problema de cómo seleccionar las unidades que constituyen s_1 y s_2.

Puesto que el coste de observación c_1 es fijo para todas las unidades de s_1, un criterio útil para decidir la elección de la muestra consiste en minimizar el valor esperado de la varianza de $f(\bar{y}/\vec{x}_{s_1})$. Se comprueba que dicho valor esperado es constante para toda muestra s_1 de tamaño

n, por lo que dicho criterio es compatible con la selección de s_1 mediante un diseño aleatorio simple.

Razonando de igual manera con la muestra s_2 debe calcularse el valor esperado respecto de $f(\vec{y}_{s_2}/\vec{x}_{s_1})$ de la varianza de $f(\bar{y}/\vec{x}_{s_1}, \vec{y}_{s_2})$. Su valor viene expresado por:

$$k_1 + k_2 \sum_{i \in s_1 - s_2} x_i + k_2 \frac{((N-n)\mu'' + \sum_{i \in s_1 - s_2} X_i)^2}{h' + \sum_{i \in s_2} x_i}$$

donde

$$k_1 = \frac{N-n}{N^2} h''\mu'' + \frac{(N-n)^2}{N^2}\left(\alpha'^2 + \frac{b'}{h'(a'-2)} \operatorname{Var}(\bar{X}_{s_i'}/\vec{x}_{s_1})\right)$$

y

$$k_2 = \frac{b'}{N^2(a'-2)}.$$

Se observa que dicho valor esperado alcanza su mínimo cuando la muestra s_2 se forma con aquellas unidades de s_1 con los valores de x_i más grandes. Esta elección extrema, que es óptima, no es por otra parte robusta frente a violaciones en el modelo. Una forma de elección alternativa que ofrece ciertas garantías de robustez es la que toma muestras «equilibradas» en el sentido que definen Royall y Herson (1973a) de que $\bar{x}_{s_1} = \bar{x}_{s_2}$. A la varianza de la distribución $f(\bar{y}/\vec{x}_{s_1}, \vec{x}_{s_2})$ para este tipo de muestras las denotaremos por Veq.

5. DETERMINACION DE LOS TAMAÑOS MUESTRALES

Si como sugieren los resultados establecidos en la sección 4 se decide seleccionar la muestra s_1 mediante un diseño aleatorio simple y la s_2 de forma que sea equilibrada, faltará por determinar los tamaños n y m de ambas muestras.

Sea $L = K\,\text{Veq}\,(\bar{Y}/\vec{x}_{s_1}, \vec{y}_{s_2}) + c_1 n + c_2 m + c_0$ una función de pérdida en la que c_0 es un coste fijo y K una constante de proporcionalidad que permite medir el error de estimación en términos de coste monetario.

Para terminar m se procederá a minimizar el valor esperado respecto

de $f(\vec{y}_{s_2}/\vec{x}_{s_1})$. Considerando un valor elevado para N, dicho mínimo se obtiene para:

$$m = \begin{cases} 0 & \text{si} \quad p \leqslant 0 \\ np & \text{si} \quad 0 \leqslant p \leqslant 1 \\ n & \text{si} \quad p \geqslant 1 \end{cases}$$

donde

$$p = \left(\frac{\mu''}{\bar{x}_{s_1}} \sqrt{\frac{k\bar{x}_{s_1}}{c_2}} \sqrt{\frac{b'}{a'-2}} - \frac{h'}{\bar{x}_{s_1}} \right) / n$$

Para determinar el valor de n debe sustituirse el valor calculado para m en la pérdida esperada y de nuevo calcular el valor esperado en este caso respecto de los datos $\vec{x}_{s_1}$. Desgraciadamente el procedimiento presenta dificultades analíticas importantes. Una solución aproximada al problema se obtiene al imponer las restricción $n = m$. El valor óptimo es entonces:

$$n = \sqrt{\frac{K\left(\frac{b'\mu'}{a'-2} + \left(\alpha'^2 + \frac{b'}{h'(a'-2)} \right) \frac{v'}{n'-2} \right)}{c_1 + c_2}}$$

6. DISCUSION

El modelo de superpoblación propuesto en la sección 2 puede no ser el más adecuado para el estudio de cierta población. En tal caso sería deseable disponer de recursos que permitieran una inferencia robusta frente a las posibles violaciones del modelo. En nuestro caso, si admitimos que el modelo de regresión propuesto no pasa por el origen, razonamientos análogos a los de la sección 3 permiten concluir que la media de $\bar{Y}$ conocidos los datos $\vec{x}_{s_1}$ y $\vec{y}_{s_2}$ es:

$$\bar{y}_{s_2} + (\bar{x}_{s_1} - \bar{x}_{s_2}) \frac{\bar{y}_{s_2} \sum_{i \in s_2} x_i^{-1} - \sum_{i \in s_2} y_i/x_i}{\bar{x}_{s_2} \sum_{i \in s_2} x_i^{-1} - m}$$

Es inmediato comprobar que ambos estimadores coinciden si se selecciona s_2 de manera que sea «equilibrada». Esta es la razón por la que se ha optado por dicho diseño.

Las conclusiones establecidas pueden extenderse fácilmente a poblaciones estratificadas. Los modelos a utilizar entonces presentarían analogías con respecto a los estudiados por Royall y Herson (1973b) y Murgui (1984) para diseños en una fase.

El análisis de diseños muestrales en dos fases mediante los modelos de superpoblación ha sido anteriormente realizado por otros autores. Singh y Sedransk (1976) consideran una variable X que toma un número finito de valores y Ericson (1967) y Basulto y Murgui (1982) particularizan a un caso en el que X sólo toma dos valores únicos. Resulta interesante por otro lado, la comparación de los resultados aquí expuestos con los obtenidos por Murgui (1983) en el supuesto de que los valores x_i, $i = 1, ..., N$ fuesen conocidos.

REFERENCES

BASULTO, J., y MURGUI, J. S. (1982): «Diseño Muestral óptimo en el caso de no respuesta», *Trabajos de Estadística y de Investigación Operativa*, vol. 33, núm. 2, pp. 3 a 15.

COCHRAN, W. G. (1978): *Sampling Techniques*, Wiley, New York.

DEGROOT, M. H. (1970): *Optimal Statistical Decision*, McGraw-Hill Book Co., Inc., New York.

ERICSON, W. A. (1967): «Optimal sample design with non-response», *J. Amer. Statist. Ass. March*, pp. 62 a 78.

MURGUI, J. S. (1983): «Estimadores de razón y de regresión en poblaciones finitas: modelos de superpoblación», *Estadística Española*, núm. 99, pp. 61 a 72.

MURGUI, J. S. (1984): «Un modelo de regresión para poblaciones finitas estratificadas», *Estadística Española*, núm. 101.

ROYALL, R. M., y HERSON, J. (1973a): «Robust Estimation in finite populations I», *J. Amer. Statist. Ass.*, Vol. 68, pp. 880 a 889.

ROYALL, R. M., y HERSON, J. (1973b): «Robust estimation in finite populations II: Stratifications on a size variable», *J. Amer. Statist. Ass.*, vol. 68, pp. 890 a 893.

SINGH, B., and SEDRANSK, K. (1976): «A two phase sample design for estimating the finite population mean: post stratification», *Technical report*, núm. 39, Statistical Science Division, Suny at Barffalo.

SUGDEN, R. A. (1978): «Exchangeability and the foundations of survey sampling», *Unpublished Ph. D. Thesis*, University of Southampton.