Estadística Española • Volumen 60, número 196 / 2018, pp. 151-157

Muestreo doble óptimo

Mariano Ruiz Espejo
Universidad Católica San Antonio de Murcia

Resumen

En este artículo ofrecemos un estimador óptimo en muestreo doble usando muestreo aleatorio simple con reemplazamiento de tamaño fijo en cada fase, observando una variable auxiliar en la primera fase, y accediendo a la variable de interés en la segunda fase de submuestreo. Hacemos uso de estimadores óptimos para distribución libre y del ajuste óptimo del modelo general lineal en poblaciones finitas.

Palabras clave: estimador óptimo, muestreo doble, muestreo aleatorio simple con reemplazamiento

Clasificación AMS: 62D05, 62H12, 62J12

Optimum double sampling

Abstract

In this article we offer an optimum estimator in double sampling using simple random sampling with replacement of fixed size in each phase, observing an auxiliary variable in the first phase, and accessing to the interest variable in the second phase of subsampling. We make use of optimum estimators for free distribution and of optimum fit of the general linear model in finite populations.

Key words: optimum estimator, double sampling, simple random sampling with replacement

AMS classification: 62D05, 62H12, 62J12

1 Introducción

Presentamos una solución al problema de estimación insesgada óptima de la media poblacional en muestreo doble con submuestreo, usando muestreo aleatorio simple con reemplazamiento en ambas fases de muestreo, y observando una variable auxiliar en la primera fase. Este es un problema que estaba abierto y al que hemos dado una solución usando tres métodos de optimización.

Veamos con detalle cómo hemos resuelto este problema del que no tenemos información de una solución satisfactoria hasta este artículo.

2 Muestreo doble

Entendemos por muestreo doble aquel procedimiento de muestreo que se desarrolla en las siguientes dos fases.

En una *primera fase* se selecciona una muestra aleatoria simple con reemplazamiento $\boldsymbol{s}$ de tamaño fijo n, a partir de una población finita U de tamaño N, y se observa la variable auxiliar x en las unidades seleccionadas. Sea el vector reordenado $(x_1, x_2, \dots, x_n)$ obtenido a partir de la muestra ordenada de datos $\big((k, x_k): k \in \boldsymbol{s}\big)$ de la variable auxiliar x. En dicho vector pueden aparecer observaciones repetidas ya que el muestreo es con reemplazamiento.

En una *segunda fase* se selecciona una muestra aleatoria simple con reemplazamiento de tamaño fijo n', a partir del vector reordenado $(x_1, x_2, \dots, x_n)$ que contiene las n observaciones de la variable auxiliar, y por tanto tiene un tamaño efectivo fijo que es un número natural "mayor o igual que 1" y "menor o igual que n" de unidades de la población finita. En esta submuestra de tamaño fijo n' obtenida en la segunda fase observamos la variable de interés y que recogemos en el vector reordenado que denotamos $(y_1, y_2, \dots, y_{n'})$ que puede contener observaciones de unidades repetidas también.

La teoría básica para los conceptos usados en esta sección pueden encontrarse en Cassel *et al.* (1977) y en Ruiz Espejo (2017).

3 Estimador propuesto óptimo

Queremos estimar la media poblacional de la variable de interés,

$$\bar{y} = \frac{1}{N}\sum_{i \in U} y_i$$

Para ello proponemos en el muestreo doble anteriormente descrito el estimador

$$\bar{y}_d = \hat{\boldsymbol{a}}\boldsymbol{A}^{-1}\bar{\boldsymbol{x}}^t$$

Este estimador ha sido propuesto por Ruiz Espejo (2015, 2018), y ahora el vector $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_{1\times(m+1)}$ no está tomado a partir de la población finita porque la variable auxiliar solo se conoce dentro de la muestra de tamaño fijo n, sino que está tomado del vector reordenado $(x_1, x_2, \dots, x_n)$ como población de referencia obtenida en la primera fase. El número $m + 1$ es el número de parámetros a ajustar en el modelo lineal

$$y_i = k_0 + \sum_{r=1}^{m} k_r f_r(x_i) + e_i$$

Aquí $k_0, k_1, \ldots, k_m$ son las constantes o parámetros a ajustar, $f_r(x)$ es la función real de variable real r-ésima usada en el ajuste, y e_i es el error del ajuste en la unidad poblacional $i \in U$.

En concreto,

$$\boldsymbol{a} = \boldsymbol{a}_{1\times(m+1)} = (A_{1;y} \quad A_{1,1;y,f_1(x)} \quad \cdots \quad A_{1,1;y,f_m(x)})$$

Siendo

$$A_{1,y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

Y para $r = 1, 2, \ldots, m$,

$$A_{1,1;y,f_r(x)} = \frac{1}{n}\sum_{j=1}^{n} y_j f_r(x_j)$$

Que son estimables insesgadamente y de mínima varianza para distribución libre respectivamente por

$$a_{1,y} = \frac{1}{n'}\sum_{i=1}^{n'} y_i$$

Y para $r = 1, 2, \ldots, m$,

$$a_{1,1;y,f_r(x)} = \frac{1}{n'}\sum_{i=1}^{n'} y_i f_r(x_i)$$

La matriz cuadrada $\boldsymbol{A} = \boldsymbol{A}_{(m+1)\times(m+1)}$ resulta ser

$$\boldsymbol{A} = \begin{pmatrix} 1 & A_{1;\,f_1(x)} & \cdots & A_{1;\,f_m(x)} \\ A_{1;\,f_1(x)} & A_{1,1;\,f_1(x),f_1(x)} & \cdots & A_{1,1;\,f_1(x),f_m(x)} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1;\,f_m(x)} & A_{1,1;\,f_m(x),f_1(x)} & \cdots & A_{1,1;\,f_m(x),f_m(x)} \end{pmatrix}$$

Esta matriz $\boldsymbol{A}$ depende exclusivamente de la información auxiliar de las variables explicativas del modelo de regresión lineal multivariante. Por ejemplo,

$$A_{1,1:\,f_r(x),f_s(x)} = \frac{1}{n}\sum_{j=1}^{n} f_r(x_j) f_s(x_j)$$

Para todo $r, s = 1, 2, \ldots, m$.

Finalmente, el vector $\overline{\boldsymbol{x}} = \overline{\boldsymbol{x}}_{1\times(m+1)}$ resulta ser el siguiente

$$\bar{\boldsymbol{x}} = (1 \quad A_{1;\, f_1(x)} \quad A_{1;\, f_2(x)} \quad \cdots \quad A_{1;\, f_m(x)})$$

Donde para $r = 1, 2, \ldots, m$,

$$A_{1;\, f_r(x)} = \frac{1}{n}\sum_{j=1}^{n} f_r(x_j)$$

En las referencias de Ruiz Espejo (2015, 2018) se prueba que el ajuste proporciona un estimador insesgado para $\bar{y}_n$, que es la media muestral de la variable de interés en la primera fase. Además "la varianza de tal estimador" en la segunda fase, $V_2(\bar{y}_d)$, minimiza para distribución libre la varianza vectorial de cualquier estimador del vector $(m+1)$-dimensional $\boldsymbol{a}$, que permite estimar el ajuste de modo insesgado y de mínima varianza para distribución libre del modelo general lineal propuesto.

Es sencillo comprobar entonces que el estimador $\bar{y}_d$ es insesgado para la media poblacional $\bar{y}$ de la variable de interés:

$$E(\bar{y}_d) = E_1[E_2(\bar{y}_d)] = E_1(\bar{y}_n) = \bar{y}$$

Además, teniendo en cuenta el teorema de Madow, su varianza verifica que

$$V(\bar{y}_d) = E_1[V_2(\bar{y}_d)] + V_1[E_2(\bar{y}_d)] =$$

$$E_1[V_2(\bar{y}_d)] + V_1(\bar{y}_n) = E_1[V_2(\bar{y}_d)] + \frac{\sigma_y^2}{n}$$

Aquí la media muestral en la primera fase $\bar{y}_n$ es estimador insesgado para $\bar{y}$, y de mínima varianza para distribución libre (Zacks, 1971, p. 150; Ruiz Espejo *et al.*, 2013; 2016), σ_y^2 es la varianza poblacional para la variable de interés, $E_1[V_2(\bar{y}_d)]$ minimiza el valor esperado de las posibles varianzas con dicho ajuste lineal para distribución libre, y por tanto $V(\bar{y}_d)$ alcanza el mínimo de cualquier estimador con el ajuste lineal dado para distribución libre. Es decir, el estimador $\bar{y}_d$ es óptimo en este sentido para distribución libre en el muestreo doble usando muestreo aleatorio simple con reemplazamiento en ambas fases y con submuestreo en la segunda fase. También la media muestral es insesgada y de varianza mínima para estimar la media poblacional en el muestreo aleatorio simple con reemplazamiento de tamaño muestral fijo, entre los estimadores lineales, cuando la población tiene varianza finita (Ruiz Espejo, 1997), como es el caso de cualquier variable de interés uniforme discreta asociada a una población finita.

4 Discusión y conclusiones

Una consecuencia directa de este artículo es que el estimador propuesto para la media poblacional a partir del modelo general lineal propuesto aquí, o a partir de m variables auxiliares conocidas de antemano como hicimos en Ruiz Espejo (2015, 2018), es estimador

insesgado de mínima varianza para distribución libre con el estimador insesgado de mínima varianza ajustado al modelo lineal. Es decir, es óptimo en dicho sentido descrito.

Una crítica que se hacía al muestreo aleatorio simple con reemplazamiento de tamaño muestral fijo n era que el muestreo aleatorio simple sin reemplazamiento de tamaño efectivo fijo n proporciona al estimador media muestral insesgación y más precisión que el anterior. Esta crítica carece de interés práctico al observar que con muestreo aleatorio simple con reemplazamiento se obtienen muestras con menor coste esperado que con muestreo aleatorio simple sin reemplazamiento del mismo tamaño muestral n, ya que las unidades pueden aparecer repetidas en el diseño muestral con reemplazamiento y como consecuencia el tamaño efectivo no es fijo sino menor igual a n, y por tanto el coste esperado es menor o igual al coste del tamaño efectivo fijo n.

El modelo general lineal que hemos propuesto con una variable auxiliar no tiene que tener solo dos parámetros de ajuste a minimizar, como sería el caso de ajustar una recta de y sobre x, ni siquiera tiene que ser un polinomio necesariamente. Un ejemplo teórico supuesto distinto a estos podría ser el siguiente

$$y_i = k_0 + k_1 e^{x_i} + k_2 \frac{1}{x_i + 2} + e_i$$

Donde en este modelo lineal con $m = 2$ tenemos las funciones

$$f_1(x) = e^x$$

Y

$$f_2(x) = \frac{1}{x + 2}$$

También es posible ajustar un modelo lineal multivariante si en la primera fase se hubieran observado m variables auxiliares en la muestra aleatoria simple con reemplazamiento de tamaño fijo n. Incluso pueden ajustarse otras posibilidades funcionales a partir de las variables auxiliares observadas en la primera fase del muestreo doble. El razonamiento es similar al propuesto en este artículo o bien en líneas semejantes a las expuestas por Ruiz Espejo (2015, 2018).

Obviamente el modelo concreto que se proponga en cada caso tiene que tener una gran fiabilidad basada en la experiencia, es decir, que ha de ser propuesto por expertos en el tipo de datos manejados y con experiencia en el área de trabajo al que se aplica el modelo. En este sentido, existe la posibilidad de que dos o más expertos distintos propongan distintos modelos lineales concretos, y es entonces cuando se tiene que llegar a un consenso o acuerdo del modelo más conveniente para el fin que nos proponemos. Una propuesta de solución de consenso es que cada experto aporte su función de ajuste según su conocimiento, y que el modelo con los datos se encargue de seleccionar el mejor ajuste lineal de dichas funciones.

El hecho de habernos referido al "muestreo de poblaciones finitas" se debe a que es en poblaciones finitas donde tiene sentido hablar de muestreo aleatorio simple con

reemplazamiento con unidades reales identificadas y accesibles. Hablar de muestreo aleatorio simple con reemplazamiento de una población infinita limita a muestras artificiales obtenidas con un ordenador y sin salir del mismo, es decir las unidades no pueden estar identificadas todas con el medio o los medios de acceso físico para observar los datos auxiliares o de interés en todas sus unidades.

La conclusión final es que queda resuelto un problema de optimización en la estimación de la media poblacional en muestreo doble con muestreo aleatorio simple con reemplazamiento en las dos fases y submuestreo en la segunda fase, haciendo uso de tres procedimientos de optimización, dos de ellos en estimación para distribución libre (opcionalmente de estimación insesgada de mínima varianza para la media poblacional con población de varianza finita, entre todos los estimadores lineales) y el otro en el ajuste del modelo general lineal en poblaciones finitas.

Una posible crítica a esta resolución del problema de optimización es que si el tamaño poblacional N es conocido, el conjunto de poblaciones finitas con distribución uniforme discreta con N valores posibles de la variable de interés es mucho más concreta que el conjunto de todas las posibles poblaciones teóricas como presupone el método de optimización para distribución libre, por lo que un estimador óptimo para distribución libre puede no serlo para el conjunto reducido de poblaciones finitas de distribución uniforme discreta con N valores de la variable de interés. De hecho, como justifica Ruiz Espejo (1987), no existe tal estimador "insesgado y uniformemente de mínima varianza", ni siquiera "uniformemente de mínimo error cuadrático medio", en el modelo de población finita fijada, que es un modelo diferente pero más próximo al modelo de muestreo doble con submuestreo, con observación de una variable auxiliar en la primera fase de muestreo.

Referencias

CASSEL, CLAES-MAGNUS; SÄRNDAL, CARL-ERIK; & WRETMAN, JAN HAKAN (1977). «*Foundations of Inference in Survey Sampling*». New York: Wiley.

RUIZ ESPEJO, MARIANO (1987). «Sobre estimadores UMV y UMECM en poblaciones finitas». *Estadística Española* 29 (115), 105-111.

RUIZ ESPEJO, MARIANO (1997). «Optimalidad insesgada de la media muestral». *Revista de la Academia de Ciencias Exactas, Físico-Químicas y Naturales de Zaragoza (2)* 52, 81-82.

RUIZ ESPEJO, MARIANO (2015). «Regresión lineal multivariante objetiva en poblaciones finitas». *Statistical Reports* 21, 1-12.

RUIZ ESPEJO, MARIANO (2017). «*Ciencia del Muestreo*». Madrid: Bubok.

RUIZ ESPEJO, MARIANO (2018). «Recientes frutos en bioestadística». *Estadística Española* 60 (195), 61-84.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2013). «Optimal unbiased estimation of some population central moments». *Metron* 71, 39-62.

RUIZ ESPEJO, MARIANO; DELGADO PINEDA, MIGUEL; & NADARAJAH, SARALEES (2016). «Optimal unbiased estimation of some population central moments». *Metron* 74, 139.

ZACKS, SHELEMYAHU (1971). «*The Theory of Statistical Inference*». New York: Wiley.