PONDERACIONES OPTIMAS: OTRA MIRADA A LA MINIMIZACION DEL ERROR EN LA ESTRATIFICACION

Carlos Bouza y Sira Allende, Departamento de Matemática Aplicada, Facultad de Matemática y Computación, Universidad de La Habana

RESUMEN
La minimización de la varianza es analizada buscando determinar ponderaciones óptimas de los estratos. Esta aparece como una opción adecuada para obtener valores de la varianza menores que las asociadas a las derivadas clásicamente. El uso de modelos superpoblacionales permite obtener valores aproximados de los parámetros desconocidos lo que hace posible hacer estimados asociados al estimador de mínima varianza. Varios ejemplos son desarrollados para ilustrar el comportamiento de esta propuesta.

ABSTRACT
The determination of optimal strata weights for minimizing the variance is analysed. It appears as an adequate option for deriving smaller variance´s values than with the classic approach. The use of a superpopulation model enhances to obtain proxys to the unknown parameters needed for calculating the minimum variance estimates. Some examples are worked out for illustrating the behavior of the proposal.

Key words: Minimum variance, Neymann allocation, superpopulation model.

MSC: 62D05

1. INTRODUCCION

La teoría del muestreo de poblaciones finitas está íntimamente ligada al modelo estratificado. Neymann (1934) utilizó el Teorema de Gauss-Markov para fijar la posibilidad de minimizar el error de estimación cuando el muestreo simple aleatorio (MSA) es utilizado, para seleccionar muestras dentro de los estratos. En su estudio, un estimador lineal fue propuesto . La minimización del error fue llevada a cabo determinando tamaños óptimos de muestra a los estratos más variables, en términos de la varianza, más baratos de muestrear y mayores. El estimador utiliza pesos que son funciones del tamaño del estrato.

En este trabajo se propone utilizar ponderaciones óptimas. Estos también buscan minimizar la varianza del estimador de la media poblacional.

En la sección 2 se analiza el problema de la afijación y se obtienen las expresiones de los pesos (ponderaciones) óptimos. Estos le asignan mayor importancia a los estratos con mayor variabilidad relativa. Como en la afijación de Neymann ellos dependen de parámetros desconocidos de la variable de interés Y. En la sección 3 se propone el uso de un modelo superpoblacional . Este permite calcular las ponderaciones utilizando la información que proporciona una variable auxiliar X, conocida y ligada con Y a través de este modelo. En la última sección se utilizan algunos ejemplos para ilustrar la ganancia en precisión asociada a la propuesta. En ellos se evidencia que el uso del enfoque propuesto es una alternativa para disminuir el error obtenido al afijar los tamaños de muestra usando el criterio de Neymann o cualquier otro de los métodos comúnmente utilizados con este fin.

2. PONDERACIONES OPTIMAS

Consideraremos el marco usual en el que se estudia una población finita $U = \{i_1, i_2, ..., i_N\}$. Un parámetro $Y = (Y_1, Y_2, ..., Y_N)$ se asocia a ella. Y_j se identifica con la unidad i_j de U y es desconocido su verdadero valor hasta medirle. El interés usual es el estudio de una función paramétrica $\theta(Y)$. Generalmente esta es la media.

Una muestra de tamaño n es seleccionada de U utilizando un mecanismo que implementa un diseño muestral d(s). Este no es más que una cierta medida de probabilidad determinada por el estadístico. Un

estimador es utilizado para obtener un valor aproximado $\theta(s)$ (proxy) del parámetro utilizando los datos de la muestra s. Es popular esperar que este sea insesgado. La precisión es medida mediante el error cuadrático medio

$$E(\theta(s) - \theta)^2 = V(\theta(s)) + B(\theta(s))^2$$

si el estimador es insesgado no hay necesidad de preocuparse por el segundo término. Si el estimador no es insesgado se espera que este sea pequeño [$E(\theta(s) \cong \theta$] o que $|B(\theta.(s))| \diagup V(\theta)^{1/2}$ lo fuere.

El principio del muestreo repetido sustenta este análisis. Este es el marco basado en el trabajo de Neymann (1934) al justificar el uso del muestreo aleatorio frente a los defensores del muestreo al juicio. Cuando θ es una función lineal puede asegurarse la normalidad asintótica de esta, ver Hájek (1960).

El trabajo de Neymann se basó en el estudio del estimador proveniente de una muestra estratificada. U está dividida en K subpoblaciones disjuntas que le cubren. Esto es

$$U = \bigcup_{1 \le h \le H} U_h$$

Cada subpoblación es llamada estrato y se selecciona una muestra de cada uno en forma independiente. Tomando como parámetros del estrato h-ésimo a

$$\mu_h = \frac{1}{N_h} \sum_{j \in U_h} Y_j$$

$$\sigma_h^2 = \frac{1}{N_h} \sum_{j \in U_h} (Y_j - \mu_h)^2$$

$$Q_h = \frac{\mu_h n_h}{\sqrt{V(m_h)}} = \left(\frac{\mu_h}{\sigma_h} \right) n_h$$

donde m_h es el estimador de la media del estrato U_h . El objetivo es estimar la media poblacional

$$\theta(Y) = \mu = \frac{1}{N} \sum_{i \in U} Y_i = \sum_{h=1}^{H} W_h \mu_h$$

Un problema de particular importancia es determinar el tamaño óptimo de la muestra. Este fue estudiado en el mencionado trabajo de Neymann (1934). Al usar el muestreo simple aleatorio con reemplazo (MSACR) para la selección de las H muestras

$$M_e = \sum_{h=1}^{H} N_h m_h / N = \sum_{h=1}^{H} W_h m_h$$

es un estimador insesgado de μ y

$$V(M_e) = \sum_{h=1}^{H} W_h^2 V(m_h) = \sum_{h=1}^{H} W_h^2 \sigma_h^2 / n_h$$

es la varianza.

El problema de optimización que plantea la obtención de tamaño óptimo de muestra es:

$$\text{P1: Min}\left\{\sum_{h=1}^{H} W_h^2\sigma_h^2/n_h \mid C = C_0 + \sum_{h=1}^{H} c_h n_h\right\}$$

Su solución establece la relación

$$n_h = \frac{(C - C_0)W_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} W_h\sigma_h/\sqrt{c_h}}$$

C_0 representa costos generales, c_h es el correspondiente a evaluar una unidad en U_h, y C es el presupuesto asignado y

$$n = \sum_{h=1}^{H} n_h$$

Este es el resultado brindado en los libros de texto, ver Cochran (1977). Consideraremos que el tamaño de la muestra total n es fijado y que los costos de muestrear en dentro de los distintos estratos es igual a c. Entonces esta expresión se simplifica y obtenemos, ver op. cit.:

$$n_h^0 = \frac{nW_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} W_h\sigma_h/\sqrt{c_h}} \tag{2.1}$$

Una discusión del vínculo de los resultados con la moderna teoría del muestreo puede verse en Chaudhuri-Vos (1988). Sin embargo esta solución obvia el hecho de que los tamaños de muestra son enteros. Por tanto la solución obtenida al calcular (2.1) y hacer un redondeo es realmente una solución no optimal. Una solución a este problema es dada en Allende-Bouza (1993) utilizando Optimización Estocástica en Enteros.

Nuestro interés es estudiar el uso de otro criterio para minimizar la varianza. Partiendo de que deseamos trabajar con un estimador insesgado de mínima varianza podemos utilizar una clase más amplia que la que estudió Neymann. Nosotros analizaremos el efecto de buscar pesos óptimos. Anteriormente se asignó un peso a cada U_h que es la probabilidad de que al seleccionar aleatoriamente una unidad ésta pertenezca a este estrato. Nosotros trataremos de minimizar el error buscando otros pesos.

Sea el problema

$$\text{P2: Min}\left\{\sum_{h=1}^{H} \beta_h^2\sigma_h^2/n_h \mid \sum_{h=1}^{H} \beta_h\mu_h - \mu = 0,\ \beta_h > 0, h = 1,...,H\right\}$$

El mínimo es obtenido al resolver el sistema

$$\frac{\partial\left(\sum_{h=1}^{H} \beta_h{}^2\sigma_h^2/n_h - \lambda * \left(\sum_{h=1}^{H} \beta_h\mu_h/\mu\right)\right)}{\partial\beta_h} = 0$$

que nos determina que

$$\beta_h = -\frac{\lambda\mu_h n_h}{\sigma_h^2}$$

donde $\lambda = \lambda^*/2$

Sin perder en generalidad podemos fijar que

$$\sum_{h=1}^{H} \beta_h = 1$$

Entonces

$$\beta_h^0 = \frac{\mu_h n_h / \sigma_h^2}{\sum_{h=1}^{H} \mu_h n_h / \sigma_h^2} = \frac{Q_h}{\sum_{h=1}^{H} Q_h}$$

Son los pesos que minimizan la varianza teniendo que la expresión de esta es:

$$V(m_\beta) = \sum_{h=1}^{H} \beta_h^2 \sigma_h^2 / n_h \qquad (2.2)$$

Note que esta afijación establece que estarán más representados aquellos estratos con mayor variabilidad relativa medida por Q_h.

3. UNA SOLUCION BASADA EN UN MODELO SUPERPOBLACIONAL

El uso de modelos superpoblacionales se remonta a trabajos desarrollados por los fundadores de la Teoría del Muestreo. Cochran (1939) la utilizó, Godambe (1955) recomendó su uso para hacer inferencias al considerar que Y es generado por un mecanismo aleatorio denominado superpoblación. Este enfoque es considerado como parcialmente Bayesiano pues a pesar de utilizar como principio el inferir a partir de una distribución a priori no busca la minimización del riesgo aposteriori. Un modelo popular es el dado por asumir que una variable conocida X permite establecer la familia a la que pertenece la distribución de Y que a su vez es caracterizada por un modelo superpoblacional $\wp$.

El modelo clásico para la estratificación, ver Font (1999) es

$$\wp : Y_{hj} = \alpha_{h0} + \alpha_{h1} X_{hj} + \varepsilon_{hj}, j = 1,...,N_h, h = 1,...,H$$

donde

$$E_\wp(\varepsilon_{hj}) = 0$$

y

$$Cov_\wp(\varepsilon_{hj}, \varepsilon_{h'j'}) = \begin{cases} \sigma_h^2 g(X_{hj}) & \text{si} \quad h \neq h' \quad y \quad j \neq j' \\ \rho'_h \sigma_h^2 \sqrt{g(X_{hj}) g(X_{hj'})} & \text{si} \quad h = h' \quad y \quad j \neq j' \\ \text{o en otro caso} \end{cases}$$

α_{ho} α_h y $g(X_{hj})$ son estrictamente positivos y ρ es el coeficiente de correlación entre X y Y. Simplificaremos este modelo haciendo $\rho_h = \alpha_{ho} = 0$ y $g(X_{hj}) = 1$. Esto nos permite modelar problemas en los que se espera una expansión de Y como función de X. En problemas como los asociados a estudios del crecimiento en que X mide los resultados ´antes y la Y ´después´ es clásico.

Podemos esperar suavizar la restricción de insesgadez de $m_ß$ por la de que esta se cumpla respecto al modelo. Entonces tendremos el problema de optimización

$$P3: \text{Min}\left\{ \sum_{h=1}^{H} \beta_h^2 \frac{\sigma_h^2}{n_h} \,\middle|\, \sum_{h=1}^{H} \beta_h \mu_{X(h)} - \mu_X = 0, \quad \beta_h > 0, \quad h = 1,..,H \right\}$$

Los pesos óptimos se definen convenientemente a partir de los resultados de X y

$$\beta_h^* = \frac{Q_h^*}{\sum_{h=1}^{H} Q_h^*}$$

donde

$$Q_h^* = \frac{n_h \mu_{X(h)} / \sigma_h^2}{\sum_{h=1}^{H} n_h \mu_{X(h)} / \sigma_h^2}$$

y el error está dado por

$$V(m_\beta) = \sum \frac{\beta_h^{*2} \sigma_h^2}{n_h} \tag{3.1}$$

Aceptando que E(A/B) ≅ E(A)/E(B) tendremos que $E_\wp(Q_h) \cong Q_h^*$ y $E_\wp(ß_h) \cong \beta_h^*$

De ahí que

$$E_\wp(E(m_ß)) \cong E_\wp\left(\sum_{h=1}^{H} \beta_h \mu_h \right) \cong \sum_{h=1}^{H} \beta_h^* \mu_h$$

por lo que aceptaríamos que

$$\sum_{h=1}^{H} \beta_h^* m_h$$

es aproximadamente insesgado. Entonces usando (3.1) tendríamos un valor aproximadamente igual al obtenible al computar (2.2).

4. COMPARACION DE LOS PESOS OPTIMOS

En muchas aplicaciones los estratos están definidos a partir de intervalos de una variable continua. Tal es el caso cuando se usan variables como el área de las fincas, los ingresos de las familias, etc. En ellos es muy común que la media y la varianza crezcan en forma no proporcional y que una medida relativa sea más adecuada para medir la precisión de los estimadores.

4.1. Población fija

Analizamos algunos ejemplos de los libros de texto para fijar el comportamiento de nuestra propuesta. En algunos casos los datos considerados como muestrales los tomamos como una población artificial.

Analizamos la ganancia de precisión debida al uso del método propuesto respecto al uso de la Afijación de Neymann (2.1), y la proporcional $n_h^P = nW_h$. Para calcular las ponderaciones óptimas son usadas n_h^0 y n_h^P.

Ejemplo 4.1. T. Yamane (1970), página 31.

Tabla 4.1.1. Parámetros del Ejemplo 4.1.

h	W_h	σ_h^2	c_h	n_h^0	μ_h	Q_h^0	β_h^0
1	0,5	3,2	1	2	5	3,1	0,6
2	0,5	14,8	4	2	15	2,39	0,4

Tabla 4.1.2. Eficiencia de las Ponderaciones Optimas en el Ejemplo 4.1.

Criterio	Varianza	$V(m_\beta)$/Varianza
Afijación de Neymann	1,50	1,17
Afijación Proporcional	1,50	1,17
Ponderación Optima	1,76	1,00

En este caso el uso del método propuesto de las ponderaciones óptimas (PO) tiene un comportamento peor que la afijación de Neymann (AN) y que el de la proporcional (AP). Si tuviéramos que $\mu_1 = 15$ y $\mu_2 = 5$ los resultados cambian y por tanto son otros los valores de los parámetros, vea Tabla 4.13.

Tabla 4.1.3. Parámetros al variar los valores de la media en el Ejemplo 4.1.

h	W_h	σ_h^2	μ_h	Q_h^0	β_h^0
1	0,5	3,2	15	9,40	0,92
2	0,5	14,8	5	0,78	0,08

Ahora hay una mayor eficiencia al tener una relación diferente entre varianza y media: el método es mejor que la AP.

Tabla 4.1.4. Eficiencia del método de las proporciones óptimas con los parámetros variados para el Ejemplo 4.1.

Criterio	Varianza	$V(m_\beta)$/Varianza
Afijación de Neymann	1,50	1,17
Afijación Proporcional	2,75	0,64
Ponderación Optima	1,40	1,00

Ejemplo 4.2. T. Yamane (1970), página 132.

Tabla 4.2.1. Parámetros para el Ejemplo 4.2…n = 100

h	W_h	σ_h^2	c_h	n_h^0	n_h^P	μ_h	Q_h^0	β_h^0
1	0,16	144	9	16	16	50	55,5	0,78
2	0,32	64	4	32	32	30	13,3	0,20
3	0,52	16	1	52	52	10	1,1	0,02

Los resultados de la eficiencia están en la Tabla 4.2.2. Ellos establecen que el método propuesto es muy malo para estos datos. Sin embargo si cambiamos la media del primer estrato con la del tercero tenemos el nuevo juego de parámetros que aparece en la Tabla 4.2.3 Los resultados de la eficiencia se brinda en la Tabla 4.2.4. Note que se establece la preferencia de la PO.

Tabla 4.2.2. Eficiencia del método de las proporciones óptimas para el Ejemplo 4.2.

Criterio	Varianza	$V(m_\beta)$/Varianza
Afijación de Neymann	0,41472	13,40
Afijación Proporcional	0,41472	13,40
Ponderación Optima	5,5556	1

Tabla 4.2.3. Parámetros al variar los valores de las medias del Ejemplo 4.2. n = 100.

h	W_h	σ_h^2	c_h	n_h^0	n_h^P	μ_h	Q_h^0	β_h^0
1	0,16	144	9	16	16	10	1,11	0,006
2	0,32	64	4	32	32	30	15,00	0,084
3	0,52	16	1	52	52	50	162,50	0,910

Tabla 4.2.4. Eficiencia del método de las proporciones óptimas con los parámetros variados para el Ejemplo 4.2.

Criterio	Varianza	$V(m_\beta)$/Varianza
Afijación de Neymann	0,41472	0,65
Afijación Proporcional	0,41472	0,65
Ponderación Optima	0,2692	1

Ejemplo 4.3. Ardilly (1994).En este ejemplo el comportamiento de las distintas afijaciones: AN, AP y la arbitraria es estudiado. Los resultados iniciales están en la Tabla 4.3.1. Para cada afijación se computa el vector de PO. Determinamos un valor para la varianza utilizando cada uno de ellos. Estas aparecen en la Tabla 4.3.2. Vea en la Tabla 4.3.3 como se comporta cada PO con respecto a los errores asociados a los métodos tradicionales. Así si usamos las PO´s determinadas usando la afijación arbitraria o la AP la AN es mejor. Si utilizamos la determinada por esta última, nuestra propuesta siempre es peor.

Tabla 4.3.1. Parámetros para el Ejemplo 4.3. n = 300.

	W_h	σ_h^2	n_h	n_h^0	n_h^P	μ_h	Q_h^a	β_h^0	Q_h^0	β_h^0	Q_h^P	β_h^P
1	0,47	1,5	130	71	142	5	433,3	0,462	236,7	0,303	473,3	0,509
2	0,28	4,0	80	70	85	12	240,0	0,256	210,0	0,268	255,0	0,275
3	0,14	8,0	60	49	42	30	225,0	0,241	183,8	0,234	157,5	0,170
4	0,10	100,0	25	100	28	150	37,5	0,040	150,0	0,192	42,0	0,045
5	0,01	2500,0	5	10	3	600	1,2	0,001	2,4	0,003	0,71	0,001

Tabla 4.3.2. Varianzas para los distintos vectores de Ponderaciones Optimas en el Ejemplo 4.3.

Varianzas	$V(m_\beta^*)$	$V(m_\beta^*)/V(m_\beta^a)$	$V(m_\beta^*)/V(m_\beta^0)$	$V(m_\beta^*)/V(m_\beta^P)$
$V(m_\beta^a)$	0,020	1,00	8,00	0,95
$V(m_\beta^0)$	0,160	0,125	1,00	0,12
$V(m_\beta^P)$	0,019	1,05	8,40	1,00

Tabla 4.3.3. Eficiencia del método de las proporciones óptimas respecto al uso de las afijaciones usuales en el Ejemplo 4.3.

Criterio	Varianzas	$V(m_\beta^a)/V$	$V(m_\beta^0)/V$	$V(m_\beta^P)/V$
Afijación Arbitraria	0,055	0,36	2,91	0,34
Afijación de Neymann	0,010	2,00	16,00	1,90
Afijación Proporcional	0,086	0,23	1,86	0,22

4.2. Superpoblaciones

El efecto de utilizar la información adicional brindada por X cuando el modelo superpoblacional genera Y es evaluado usando una contrapartida de los ejemplos analizados. Tomemos ε_{hj} como una variable normal con varianza h y media cero. El valor correspondiente se le sumó a X ($Y = X + \varepsilon_{hj}$). Esta variable era el valor obtenido en uno de los ejemplos 4.1 y 4.2. Generamos 100 poblaciones con la estructura dada en el ejemplo por las ponderaciones $W_h \cdot N$ =10 000 fue fijado. Utilizando esta distribución computamos en cada una (3.1)

y la eficiencia correspondiente. El promedio de estas fue calculada y los resultados son los dados en la Tabla 4.4.1. Note que el uso de la PO basada en X fue como promedio muy similar a la obtenible utilizando Y. En la columna variación aparece la variabilidad de la eficiencia en las poblaciones generadas.

Tabla 4.4.1. Eficiencia de las ponderaciones óptimas bajo Y = X + ε en 100 poblaciones usando la N(0,h) en los Ejemplos 4.1 y 4.2.

Criterio	Ejemplo Original	4.1 Variación	Ejemplo Original	4.2 Variación
Mínimo	1,03	0,87	19,5	0,40
Máximo	2,92	1,05	9,4	0,82
Promedio	1,25	0,98	10,7	0,72
Población Fija	1,17	0,93	13,4	0,65

Tabla 4.4.2. Eficiencia de las ponderaciones óptimas bajo Y = X + ε en 100 poblaciones usando la N(0,h), en el Ejemplo 4.3.

Criterio		$V(m_\beta^a)/V$	$V(m_\beta^0)/V$	$V(m_\beta^P)/V$
Afijación Arbitraria	Máximo	0,42	4,81	0,64
	Mínimo	0,21	1,94	0,21
	Promedio	0,38	2,66	0,33
	Población Fija	0,36	2,91	0,34
Afijación de Neymann	Máximo	3,27	19,5	2,31
	Mínimo	2,11	13,4	1,36
	Promedio	2,19	16,8	2,10
	Población Fija	2,00	16,0	1,90
Afijación Proporcional	Máximo	1,13	2,32	0,39
	Mínimo	0,21	1,45	0,38
	Promedio	1,03	1,81	0,23
	Población Fija	0,23	1,86	0,22

Note que el análisis del Ejemplo 4.3 establece un resultado parecido sobre la similitud del promedio para la AP.

RECONOCIMIENTOS

Este trabajo fue desarrollado durante una beca postdoctoral de uno de los autores auspiciada por el Programmabteilung Süd Referat 14 del Deutscher Akademischer Austauschdienst sobre el tema `Optimization in Statistics` que se desarrolla dentro de un proyecto con la Humboldt Universitäat zu Berlin. Agradecemos a los revisores las sugerencias hechas a una versión inicial de este trabajo lo que permitió mejorarle considerablemente .

REFERENCIAS

ALLENDE, S. and C. BOUZA (1993): "Stochastic Programming approaches for the estimation of the mean in stratified populations", **Inv. Operacional**, 13, 109-118.

ARDILLY, P. (1994) : **Les Téchniques des Sôndages**, Editions Technic, Paris.

COCHRAN, W. (1939): "The use of the Analysis of Variance in enumeration by sampling", **J. American Stat. Ass**., 34, 429,-510.

CHAUDHURI, A. y J.W.E. VOS (1988): **Unified Theory and Strategies in Survey Sampling**, N. Holland, Amsterdam.

COCHRAN, W. (1977) : **Sampling Techniques**, Wiley, New York.

FONT, B. (1999) : "Una revisión de diferentes aportaciones al diseño en poblaciones finitas", **Questiio**, 23, 3-35.

GODAMBE, V.P. (1955) : “A unified theory of sampling”, **J. Royal Stat. Soc**.-B. 17, 269-278.

HAJEK, J. (196?): “Limiting distributions in simple random sampling from a finite population“, **Pub. Math. Inst. Hungarian Academy of Science**, 5, 361-374.

NEYMANN, J. (1934): “On the two different aspects of the representative method: the method of statistical sampling and the methodd of purposive selection”, **J. Royal Stat. Soc**, 97, 558-625.

YAMANE, T. (1970) : **Elementary Sampling Theory**, Instituto Cubano del Libro, Habana.