TRABAJOS DE ESTADISTICA Y
DE INVESTIGACION OPERATIVA
Vol. 33, Núm. 3, 1982, pp. 30 a 37

# AFIJACION DEL TAMAÑO DE MUESTRA EN ESTRATEGIAS $\pi$ PXM - ESTRATIFICADAS BAJO UN MODELO SUPERPOBLACIONAL

*C. N. Bouza*
*Universidad de La Habana*

**SUMARIO**

El problema de la afijación de los tamaños de muestra en estratos es abordado. Los sistemas de probabilidades desiguales se basan en una variable auxiliar $X^t$. La existencia de un modelo superpoblacional permite el desarrollo de criterios de afijación óptima. Las propiedades de los tamaños de muestra obtenidos son similares a los clásicos.

Palabras Claves: Muestreo $\pi$PXM; Modelo superpoblacional; Afijación óptima; Muestra balanceada.

## 1. Introducción

Es usual que una población actual $U$ esté particionada en estratos. En estos casos se denota la población mediante:

$$U = \sum_{i=1}^{K} U_i; \quad U_i \cap U_j = \phi, \quad \forall\, i \neq j$$

siendo $U_i = (u_{i1}, u_{i2}, ..., u_{iN_i})$ un conjunto de individuos. La selección de una muestra $s \subset U$ se hace utilizando esta estructura por lo que se representa como una partición

$$s = \sum_{i=1}^{K} s_i; \quad s_i \subset U_i$$

La selección de cada $s_i$ se lleva a cabo independientemente de las demás mediante una medida de probabilidad $d_i$ conocida como diseño muestral. Generalmente todos los diseños utilizados para seleccionar las $s_i$ pertenecen a una misma familia. Esto es, se usa muestreo aleatorio simple (mas), de probabilidades desiguales, etc. para seleccionarles. Los diseños caracterizan las probabilidades de inclusión de un $u_{ij}$ en la muestra mediante la relación

$$\pi_j(i) = \sum_{s_i \ni u_{ij}} d(s_i); \quad i = 1, ..., K; \quad j = 1, ..., N_i \qquad (1)$$

Las probabilidades de inclusión conjunta de $u_{ij}$ y $u_{ip}$ también son calculables a partir del diseño. En este caso esta es

$$\pi_{jp}(i) = \sum_{s_i \supset (u_{ij},\, u_{ip})} d(s_i) \qquad (2)$$

A la población $U$ se asocia un vector de parámetros $Y$ en el que cada coordenada es el valor de la característica de interés $Y$ en un individuo. Este valor se puede denotar por $Y(u_{ij}) = Y_{ij}$ al tomar en cuenta la pertenencia de cada individuo a un estrato. Si se quiere caracterizar el comportamiento de $Y$ es usual el empleo del total dado por la función paramétrica

$$T_Y = \sum_{i=1}^{K} \sum_{j=1}^{N_i} Y_{ij}$$

Cuando es conocido el vector paramétrico de una variable auxiliar cuya correlación con $Y$ es alta es recomendable hacer que las probabilidades de inclusión sean proporcionales a esta variable. En el caso de poblaciones no estratificadas se ha estudiado ampliamente el comportamiento de diseños que utilizan este método. Cuando la expresión de (1) es

$$\pi_j(i) = n_i X_{ij}/T_{X_i};\quad T_{X_i} = \sum_{j=1}^{N_i} X_{ij}$$

dentro de cada $U_i$ se dice que el diseño es $\pi$ PX-estratificado como extensión de la definición de diseño $\pi$ PX utilizada en poblaciones no particionadas en estratos. El comportamiento de estos diseños ha sido poco estudiada. Rao (1968) estudió el problema de la afijación en la estratificación al usar diseños $\pi$ PX en cada $U_i$.

La estimación de $T_Y$ al usar diseños $\pi$ PX se vincula con estimadores del tipo Horvitz-Thompson (1952). Estos son insesgados pero además poseen otras propiedades estadísticas que le hacen atrayente. Una buena caracterización de este estimador es hecha por Sampford (1975).

En el presente trabajo se analiza el comportamiento de estrategias muestrales que utilizan como probabilidades de inclusión en cada estrato $U_i$ a

$$\pi_j(i|t) = n_i X_{ij}^t/T_{X_i}(t);\quad T_{X_i}(t) = \sum_{j=1}^{N_i} X_{ij}^t$$

$j = 1, ..., N_i;\ i = 1, ..., K;\ t \in \mathrm{IR}$

Los diseños asociados a estos sistemas de probabilidades se conocen con el nombre de $\pi$ PX de tamaño modificado o sencillamente $\pi$ PXM.

Partiendo de la familia de distribuciones a priori que genera $Y$ se puede hacer un estudio del comportamiento de los $n_i$ óptimos. En este trabajo se utiliza el modelo superpoblacional

$$M(g,\ p\ i):\ E_M(Y_{ij}|X_{ij}) = a_i X_{ij}^p$$

$$\mathrm{Cov}(Y_{ij},\ Y_{ij'}|X_{ij}\ X_{ij'}) = \begin{cases} \sigma_i^2\ X_{ij}^g & \text{si}\quad j = j' \\ 0 & \text{si}\quad j \neq j' \end{cases} \qquad (4)$$

dentro de cada estrato. Se obtienen tamaños óptimos para la muestra y un estudio sobre los valores adecuados de $t$ y $p$ es realizado.

## 2. Afijación del tamaño de la muestra

En el uso de la estratificación un problema de especial importancia es la determinación de los tamaños de las muestras $s_i$. Es común que el criterio de afijación utilizado se base en la minimización de la varianza del estimador dado un costo fijo. Neymann (1934) obtuvo fórmulas para obtener los valores óptimos de los $n_i$ bajo este criterio al utilizar el diseño masa y el estimador media ponderada.

$$\bar{y}_e = \sum_{i=1}^{K} \left(N_i/N\right) \sum_{u_{ij} \in s_i} Y_{ij}/n_i$$

como estrategia muestral para $T_Y/N = \bar{Y}$.

Al utilizar diseños $\pi$ PXM en cada estrato tiene sentido seguir la costumbre de hacer uso de estimadores de Horvitz-Thompson para el total en cada $U_i$. Un estimador insesgado de $T_Y$ es

$$y = \sum_{i=1}^{K} \sum_{u_{ij} \in s_i} Y_{ij}/\pi_j(i|t)$$

donde $\pi_j(i|t)$ es calculable mediante (3).

Y es un estimador lineal y los diseños $d_i$ son independientes de ahí que el error de éste sea

$$V = V(y) = \sum_{i=1}^{K} \left[ \sum_{j=1}^{N_i} \frac{(1 - \pi_j(i|t))\, Y_{ij}^2}{\pi_j(i|t)} + \right.$$

$$\left. + \sum_{j \neq j'} \frac{[\pi_{jj'}(i|t) - \pi_j(i|t)\, \pi_{j'}(i|t)]\, Y_{ij}\, Y_{ij'}}{\pi_{jj'}(i|t)} \right] = \sum_{i=1}^{K} V_i$$

donde $\pi_{jj'}(i|t)$ es la probabilidad de la inclusión conjunta de $u_{ij}$ y $u_{ij'}$ al usar el diseño $\pi$ PXM con parámetro $t$. Si hay motivos para aceptar que en cada $U_i$ el modelo (4) ha generado a $\underline{Y}_i = (Y_{i1}, \ldots, Y_{iN_i})$

es lógico fijar los $n_i$ óptimos utilizando el error esperado de $y$. Denotando $E_M$ como la esperanza bajo ese modelo

$$E_M\left[\frac{(1-\pi_j(i\,|\,t))\,Y_{ij}^2}{\pi_j(i\,|\,t)}\right]=\frac{(1-\pi_j(i\,|\,t)}{\pi_j(i\,|\,t)}\,(a_i^2\,X_{ij}^p+\sigma_i^2\,X_{ij}^g) \qquad (5)$$

es la esperanza de cada sumando en la sumatoria de los $Y_{ij}^2$. La de los términos cruzados está dada por:

$$E_M\left[\frac{(\pi_{jj'}(i\,|\,t)-\pi_j(i\,|\,t)\,\pi_{j'}(i\,|\,t))\,Y_{ij}\,Y_{ij'}}{\pi_{jj'}(i\,|\,t)}\right]=$$

$$=\left[\frac{\pi_{jj'}(i\,|\,t)-\pi_j(i\,|\,t)\,\pi_{j'}(i\,|\,t)}{\pi_{jj'}(i\,|\,t)}\right][a_i^2\,X_{ij}^p\,X_{ij'}^p] \qquad (6)$$

Sumando estos resultados y tras alguna manipulación algebraica se obtiene que:

$$E_M(V_i)=\sigma_i^2\sum_{j=1}^{N_i}(1-\pi_j(i\,|\,t)\,X_{ij}^g/\pi_j(i\,|\,t)+a_i^2\,\mathrm{Var}\,[\sum_{u_{ij}\in s_i}X_{ij}^p/\pi_j(i\,|\,t)]=\bar{V}_i$$

Sustituyendo $\pi_j(i\,|\,t)$ por su valor en la estrategia

$$\bar{V}_i=\sigma_i^2\,T_{X_i}(t)\,[\sum_{j=1}^{N_i}X_{ij}^{g-t}/n_i-\sum_{j=1}^{N_i}X_{ij}]+$$

$$+\,[a_i^2\,T_{X_i}(t)/n_i]\,[\mathrm{Var}\sum_{u_{ij}\in s_i}X_{ij}^{p-t}] \qquad (7)$$

Una función de costo adecuada para este problema es:

$$C=c_0+\sum_{i=1}^{K}c_i\,n_i$$

donde $c_0$ es un costo general de las operaciones de muestreo y $c_i$ el de evaluar un individuo en $U_i$. Al fijar un presupuesto $C'$ para la encuesta se quiere minimizar el error esperado

$$\bar{V} = \sum_{i=1}^{K} \bar{V}_i$$

Tomando $\lambda$ como un multiplicador de Lagrange y derivando con respecto a cada $n_i$ la función:

$$F = \bar{V} - \lambda \left(C' - c_0 - \sum_{i=1}^{K} c_i n_i\right)$$

se obtienen $K$ relaciones del tipo

$$\lambda n_i c_i = \sigma_i^2 \left[\sum_{j=1}^{N_i} X_{ij}^{t}\right] \left[\sum_{j=1}^{N_i} X_{ij}^{g-t}\right] - T_{x_i}(t)\, a_i^2 \operatorname{Var}\left[\sum_{u_{ij} \in s} X_{ij}^{p-t}\right] = Z_i^2$$

por lo que:

$$n_i = Z_i \left(\lambda c_i\right)^{-1/2}; \quad i = 1, \ldots, K$$

Si el tamaño total de la muestra $s$ se fija como $n$ se obtiene que:

$$\lambda^{1/2} = \sum_{i=1}^{K} Z_i / n\, c_i^{1/2}$$

y $\lambda^{1/2}$ es despejado fácilmente. Su sustitución en (8) fija como tamaños óptimos de muestra las $K$ expresiones:

$$n_i = \frac{n c_i^{-1/2} \left[T_{X_i}(t)\, \sigma_i^2 \sum_{j=1}^{N_i} X_{ij}^{g-t} + a_i^2 \operatorname{Var}\left(\sum_{u_{ij} \in s_i} X_{ij}^{p-t}\right)\right]^{1/2}}{\sum_{i=1}^{K} c_i^{-1/2} \left[\sigma_i^2\, T_{X_i}(t) \sum_{j=1}^{N_i} X_{ij}^{g-t} + T_{X_i}(t)\, a_i^2 \operatorname{Var}\left[\sum_{u_{ij} \in s_i} X_{ij}^{p-t}\right]\right]^{1/2}} \tag{9}$$

Note que Var $[\sum_{u_{ij} \in s_i} X_{ij}^{p-t}] = 0$ si $p = t$ lo que simplifica la expresión de los $n_i$ óptimos. Usando los resultados de Callebout (1965) se tiene que el producto

$$[\sum_{j=1}^{N_i} X_{ij}^{t}] [\sum_{j=1}^{N_i} X_{ij}^{g-t}]$$

crece como función de $|(g - 2t)/2|$. De ahí que cuando $t = g/2$ el error esperado en $U_i$ tenga un menor peso en la afijación de $n_i$ que los costos de evaluar cada $u_{ij}$ en $s_i$.

Un análisis de (9) denota que los estratos con mayor representación en $s$ son los más variables internamente, los más baratos de muestrear y los de mayor peso relativo en sus totales $T_{X_i}(t)$. Estos resultados son cualitativamente similares a los que obtuvo Neymann en poblaciones actuales al recordar que los totales se relacionan en gran medida con el tamaño del estrato.

La selección del grado $t$ del sistema de probabilidades de inclusión lo fija el estadístico. Tomar $p = t$ o $g/2 = t$ puede llevar a distorsiones graves en la representación de los estratos. Una solución más aceptable que hacer $p = t$ es trabajar con muestras balanceadas en el sentido de que:

$$\sum_{u_{ij} \in s_i} X_{ij}^{p-t} = \sum_{u_{ij'} \in s_i'} X_{ij'}^{p-t}; \quad s_i \neq s_i' \text{ de probabilidad no nula}$$

Esto garantizaría que (10) se cumpla aproximadamente. Si los estratos deben ser construidos estos resultados sugieren que puede usarse la característica $X^t$ para garantizar que las muestras dentro de $U_i$ sean aproximadamente balanceadas.

## BIBLIOGRAFIA

Bouza C.N. (1978): Criterios de preferencia para el uso de conjuntos de probabilidades de inclusión poissonianas. Trab. de Estad. y de Inv. Operativa. 39, 34-9.

Callebout P.K. (1965): Generalization of Cauchy - Schwarz inequallity. J. of Math. Analysis and App. 12, 491-4.

Horvitz D.G. and Thompson D.J. (1952): A generalization of sampling without replacement from a finite universe. J. Amer. Stat. Ass. 47, 663-85.

Neymann J. (1934): On the two different aspects of the representative method. J. Royal Stat. Soc. 97, 558-625.

Rao T.J. (1968): On the allocation of sample size in strati fied sampling. Annals of the Inst of Stat. Math. 20, 159-66.

Sampford M.R. (1975): The Horvitz –Thompson method in theory and practice– an historical survey. Presentado en la 40ma sesión del ISI. Varsovia.