ESTRATIFICACIÓN POR TAMAÑO DE UNIDADES DE MUESTREO

BASILIO A. ROJAS
El Colegio de México

INTRODUCCIÓN

EN LA TEORÍA Y APLICACIÓN del muestreo el diseño estratificado tiene especial importancia por las múltiples circunstancias prácticas en que se utiliza con grandes ventajas para las estimaciones. Consiste esencialmente en que las unidades de muestreo se reparten en grandes grupos llamados estratos. Para reducir la varianza de la estimación, las unidades dentro de un estrato deberán ser más homogéneas en una o más características pertinentes a la o las estimaciones que se pretenden hacer. Además, se procura que los estratos, homogéneos dentro de sí, difieran lo más posible entre ellos. En esta forma la variabilidad total que posean las unidades de muestreo se puede reducir de manera considerable.

Con el criterio expresado a grandes rasgos arriba se definen los estratos y el problema siguiente es determinar el número de unidades en la muestra correspondiente a cada estrato. Una resolución en este problema es la afijación óptima de Neyman que minimiza la varianza del estimador de la población bajo ciertas restricciones.

El propósito de este estudio es establecer un método analítico para la formación de estratos cuando se conoce el tamaño de todas y cada una de las unidades de muestreo que componen a la población. Por tamaño entendemos el valor x de una característica de la unidad, el que suponemos está altamente correlacionado con la característica y que pretendemos estimar. Suponemos así que se saben las frecuencias absolutas o relativas de x o bien la densidad de probabilidades $f(x)$ en el caso de que x sea continua.

Kendall[1] examina brevemente la formación de estratos y expone los principios esenciales investigados por Dalenius y Cochran, entre otros. Consideramos que nuestro estudio difiere conceptualmente y además conduce a una metodología más simple.

AFIJACIÓN ÓPTIMA

El total de unidades de muestreo en la población es N. Tenemos p estratos: $i = 1, \ldots, p$. El total de unidades en el estrato i es N_i. Por lo tanto,

[1] Maurice G. Kendall y Alan Stuart, *The Advanced Theory of Statistics*, Vol. 3, Londres, Charles Griffin and Co., 1966.

$$N = \sum_{i=1}^{p} N_i \tag{1}$$

En el estrato i se observan n_i unidades, seleccionadas al azar de las N_i. El promedio de la característica x en este estrato es $\bar{x}_i$ y el estimador del total de la característica x en la población es X:

$$X = \sum_{i=1}^{p} N_i \bar{x}_i \tag{2}$$

Si σ_i^2 es la varianza de la característica x en las unidades del estrato i, la varianza de X será: [2]

$$V(X) = \frac{\sum_{i=1}^{p} N_i^2 \sigma_i^2}{n_i} \tag{3}$$

Si el costo de investigar u observar una unidad de muestreo es constante para todos los estratos, la afijación óptima de Neyman nos indica que los n_i deberán ser proporcionales a los productos $N_i \sigma_i$. Si el tamaño total de la muestra es n, entonces:

$$n = \sum_{i=1}^{p} n_i \tag{4}$$

y

$$n_i = n \frac{N_i \sigma_i}{\sum N_i \sigma_i} \tag{5}$$

Bien podemos teóricamente concebir un estrato e que tenga un promedio $\bar{x}_e$ muy pequeño, un número de unidades N_e grande y una σ_e^2 también grande. Este estrato, aun cuando requiera, de acuerdo con la fórmula (5), un n_e relativamente grande, es decir, que esté muy bien representado en la muestra, participaría con muy poco o nada en la estimación de X [fórmula (2)]. Esto nos indica que la afijación de la muestra por el método descrito será defectuosa y sospecharíamos que el calificativo de óptima no es el adecuado.

En diversas investigaciones se ha observado que las varianzas de grupos o estratos son función de los valores medios de los grupos. Si en la estratificación realizada se cumple la ley

$$\sigma_i^2 = k_1 \mu_i^2 \tag{6}$$

en la que μ_i es el valor medio verdadero del estrato i y k_1 es una constante, los n_i serán proporcionales a $N_i \mu_i$, es decir, a los totales de cada estrato y no se presentaría la dificultad anotada arriba. Pero la suposición (6) no es de ninguna manera una ley general.

[2] W. G. Cochran, *Sampling Techniques*, Nueva York, John Wiley, 1963, 2ª ed.

La fórmula (5) es válida cuando el costo de investigar una unidad de muestreo es el mismo para cualquier estrato. Si las unidades se forman con respecto al tamaño x, el costo es independiente del valor de x. Si, por ejemplo, la muestra se refiere a producción de trigo de predios agrícolas, el costo sería el mismo para investigar un predio de 100 hectáreas que para el de una hectárea. Efectivamente, el tiempo requerido para hacer la entrevista sería aproximadamente el mismo; sin embargo, la información de, digamos, 200 toneladas de trigo es muy diferente a la de 2 toneladas. Si el costo lo referimos no al de la entrevista por predio, sino a una tonelada de trigo, o en general al valor unitario de x, el costo de la entrevista será inversamente proporcional al valor de x que posea la unidad de muestreo. En esta forma el costo medio por unidad en el estrato i sería

$$C_i = \frac{k_2}{\mu_i} \tag{7}$$

Bajo esta suposición de costos diferenciales la afijación óptima de los n_i estará dada por

$$n_i = K_1 N_i \sigma_i \sqrt{\mu_i} \tag{8}$$

en la que K_1 es la constante de proporcionalidad, obtenida por la relación (4).

En las distribuciones estadísticas existe una relación, más o menos complicada, entre la desviación estándar σ_i y la amplitud (recorrido) A_i de los datos. Si

$$\sigma_i = k_3 A_i \tag{9}$$

la expresión (8) quedaría

$$n_i = K_2 N_i A_i \sqrt{\mu_i} \tag{10}$$

la cual tendría la ventaja de la simplicidad en la obtención del valor de A_i:

$$A_i = X_{i,\,\text{máx}} - X_{i,\,\text{mín}}$$

o sea la diferencia entre los valores máximos y los mínimos de las observaciones en el estrato i.

Puede demostrarse que la expresión (9) es más exacta a medida que el número de estratos es mayor y los A_i correspondientes más pequeños. El error Δn_i en el cálculo del n_i debido al error $\Delta\sigma_i$ en la determinación de σ_i es, según la fórmula (8):

$$\Delta n_i = K_1 N_i \sqrt{\mu_i}\, \Delta\sigma_i$$

si p (número de estratos) es grande y por lo tanto K es prácticamente independiente de σ_i.

El error relativo será

$$\frac{\Delta n_i}{n_i} = \frac{\Delta \sigma_i}{\sigma_i} \tag{11}$$

La distribución parcial $f(x)$ de un estrato, en el caso de variable continua, se puede aproximar a una distribución trapecial cuando el número de estratos es grande, que tiene como límites la rectangular, de varianza máxima, y la triangular, de varianza mínima. Por lo tanto, los límites superior e inferior de σ_i^2 son como sigue:

$$\frac{A_i^2}{18} \leqslant \sigma_i^2 \leqslant \frac{A_i^2}{12},$$

y por lo tanto,

$$0.236 \leqslant \frac{\sigma_i}{A_i} \leqslant 0.290 \tag{12}$$

El resultado anterior nos señala que el error relativo tendrá como límites

$$0 \leqslant \left| \frac{\Delta n_i}{n_i} \right| \leqslant \frac{0.054}{0.236} = 0.228 \tag{13}$$

En realidad el error relativo máximo de 0.228 muy difícilmente ocurrirá, pues la constante k_3 tomará un valor intermedio entre 0.236 y 0.290. Además, dicho error no será sistemático para los estratos. Por otra parte, si tenemos un número grande p de estratos los n_i serán relativamente pequeños y los errores en estos n_i, máximo de 22.8 %, tendrán escasa trascendencia. Pero esencialmente la exactitud en los valores de n_i es de validez dudosa ya que la estratificación con base en el tamaño x no será la óptima para la variable y que pretendemos estimar y que suponemos tiene alta correlación con la x, sin esperar que esa correlación sea uno.

Formación de estratos. Densidad continua

Sea la variable x tal que $x_{\text{mín}} \leqslant x \leqslant A + x_{\text{mín}}$ y cuya densidad de probabilidades $f(x)$ es también continua y conocida. Subdividamos la población en p estratos, $i = 1, 2, \ldots, p$, en tal forma que la amplitud (*range*) de cada estrato sea constante e igual a $A_e = A/p$. Los límites inferiores b_{i-1} y superior b_i en la variable x del estrato i se obtienen de la fórmula

$$b_i = i \cdot A_e + x_{\text{mín}} \tag{14}$$

La media aproximada del estrato i es

$$\tilde{\mu}_i = \frac{2i-1}{2} A_e + x_{\text{mín}}$$

El número N_i de unidades en el estrato i es proporcional a

$$N_i \propto \int_{b_{i-1}}^{b_i} f(x)\,dx$$

Aplicando estos resultados en la fórmula (10), encontramos que el tamaño óptimo de la muestra n_i correspondiente al estrato i es proporcional a

$$n_i \propto A_e \sqrt{\frac{2i-1}{2} \cdot A_e + x_{\text{mín}}} \int_{b_{i-1}}^{b_i} f(x)\,dx$$

la que puede expresarse como sigue:

$$n_i \propto \sqrt{\frac{2i-1}{2} + \frac{x_{\text{mín}}}{A_e}} \int_{b_{i-1}}^{b_i} f(x)\,dx$$

Si

$$K_i = \sqrt{\frac{2i-1}{2} + \frac{x_{\text{mín}}}{A_e}}, \quad I_i = \int_{b_{i-1}}^{b_i} f(x)\,dx$$

y n es el tamaño total de la muestra, tenemos:

$$n_i = n \frac{K_i I_i}{\Sigma K_i I_i} \tag{15}$$

Si el valor mínimo x_{min} de la variable x en la población es cero, K_i se reduce a

$$K_i = \sqrt{\frac{2i-1}{2}}$$

La fórmula (15) es válida cuando el radical es positivo. Para generalizar el procedimiento a una variable x que puede tomar valores negativos se dividiría la población en dos grandes subpoblaciones, una conteniendo los valores positivos de x y la otra los negativos. El proceso indicado sería tomar para cada una de las subpoblaciones los valores absolutos de x.

Ilustraciones del método. Variable continua

Presentamos dos ejemplos. El primero se refiere a una distribución rectangular y el segundo a una de tipo hiperbólico que sigue la ley

de Pareto sobre ingresos familiares. Para simplificar los cálculos se ha considerado un número reducido de estratos, ya que nuestro propósito es explicar el método.

1. DISTRIBUCIÓN RECTANGULAR. Sea la variable x tal que $0 < x < 1$, y $f(x) = 1$. Suponemos $p = 4$ estratos. Consideramos $n = 2$ y la amplitud de cada estrato $A_e = 1/4 = 0.25$. Aquí el valor mínimo de x es cero.

Tenemos

$$I_i = \int_{b_{i-1}}^{b_i} f(x)\,dx = \int_{b_{i-1}}^{b_i} 1 \cdot dx = x \Big|_{b_{i-1}}^{b_i} = b_i - b_{i-1} = A_e = 0.25$$

$$K_i = \sqrt{\frac{2i-1}{2}};$$

Los cálculos aparecen en el cuadro 1.

Cuadro 1

FORMACIÓN DE ESTRATOS Y AFIJACIÓN. DISTRIBUCIÓN RECTANGULAR. $n = 25$

Estrato	Valores de x	I_i	K_i	$K_i I_i$	n_i
1	$0 < x < 0.25$	0.25	0.71	0.1775	3
2	$0.25 < x < 0.50$	0.25	1.23	0.3075	6
3	$0.50 < x < 0.75$	0.25	1.58	0.3950	7
4	$0.75 < x < 1.00$	0.25	1.87	0.4675	8
Sumas				1.3475	24

Para este caso la afijación óptima simple, con la fórmula (5) da un valor n_i constante para todos los estratos, igual a $24/4 = 6$.

2. DISTRIBUCIÓN DE PARETO. Ésta tiene la forma $f(x) = \alpha x^{\beta}$, en la que x es el ingreso por familia, α y β son los parámetros de la distribución; β es propiamente el parámetro que caracteriza la forma de la distribución y α el de escala, obtenido éste por el requisito de que la integral de $f(x)$ sobre el aspecto de la variable x sea igual a uno. Para el cálculo de los n_i, fórmula (15), la constante α no es necesaria.

Una estimación tosca del parámetro β para la distribución del ingreso familiar mensual en la República Mexicana para 1963 es $\hat{\beta} = -1.062$. Si y es el número en miles de familias que poseen el ingreso mensual x, hemos estimado

$$y = 1.44 \times 10^6 \cdot x^{-1.062}; \qquad 1\,000 < x < 10\,000.$$

El ajuste lo suponemos válido para ingresos mensuales x comprendidos entre \$ 1 000 y \$ 10 000.

Nos proponemos estimar una característica altamente correlacionada con el ingreso como pueden ser los gastos familiares. Para ello usaremos como variable auxiliar para la formación de estratos a la x como el ingreso familiar y supondremos p igual a 3 estratos. En este caso, $A = 10\,000 - 1\,000 = 9\,000$; $A_e = 9\,000/3 = 3\,000$. Así también:

$$I_i = \int_{b_{i-1}}^{b_i} x^{-1.062}\,dx = -\frac{x^{-0.062}}{0.062}\Bigg]_{b_{i-1}}^{b_i} = \frac{b_{i-1}^{-0.062} - b_i^{-0.062}}{0.062}$$

Los valores de b_i son 1 000, 4 000, 7 000 y 10 000. $x_{\min} = 1\,000$. Los resultados de los cálculos se presentan en el cuadro 2.

Cuadro 2

FORMACIÓN DE ESTRATOS Y AFIJACIÓN. DISTRIBUCIÓN DE PARETO

Estrato $i=$	Valores de $x' = x$ en miles de pesos	$\frac{2i-1}{2}$	$\frac{2i-1}{2} + \frac{x_{\min}}{A_e}$	K_i	I_i	$K_i I_i$	n_i	n_i'
1	$1 < x' < 4$	0.50	0.83	0.91	0.87	0.84	515	622
2	$4 < x' < 7$	1.50	1.83	1.35	0.33	0.45	275	236
3	$7 < x' < 10$	2.50	2.83	1.68	0.20	0.34	210	142
Sumas					1.40	1.63	1 000	1 000

En la última columna del cuadro 2 hemos insertado los valores de la afijación óptima común, que hemos denominado n_i':

$$n_i' \propto I_i$$

Repetimos que el número de estratos utilizado en el ejemplo es pequeño por facilidad en la presentación y que las fórmulas empleadas son más aproximadas cuando p es grande.

Si la característica y que pretendemos estimar tiene el promedio $\bar{y}_i$ en el estrato i, el estimador de la media general $\bar{y}$ es

$$\bar{y} = \frac{\Sigma I_i \bar{y}_i}{\Sigma I_i} \qquad (16)$$

cualquiera que haya sido el método empleado para determinar el tamaño de la muestra en los estratos.

Formación de los estratos. Densidad discreta

Éste es el caso común. Tenemos a las unidades de población clasificadas, según la variable x, en intervalos de clase y se conoce la frecuencia relativa $f(x)$ correspondiente a cada intervalo. Si éstos tienen la misma amplitud (*range*) A_e, podemos utilizar como estratos a estos intervalos y hacer uso de la fórmula (15), en la que $I_i = f_i$, en donde f_i es la frecuencia relativa del intervalo (estrato) i. Si los intervalos A_i son de tamaño diferente empleamos la fórmula (10) para determinar los n_i. En esta fórmula (10) N_i es la frecuencia absoluta y μ_i es el valor medio del intervalo de clase. Bien puede ser que los estratos agrupen diversos intervalos de clase vecinos; el tratamiento sería el mismo.

Si tenemos conocimiento de las frecuencias absolutas o relativas para cada uno de los intervalos de clase que suponemos de igual amplitud A_e, nos podemos preguntar qué se gana o se pierde si formamos los estratos con los intervalos de clase originales, en comparación con estratos integrados por dos o más intervalos adyacentes. Sea el intervalo i con frecuencia absoluta N_i y el intervalo de clase siguiente $i + 1$ con frecuencia absoluta N_{i+1}, ambos con la misma amplitud A_e. Llamemos R a la eficiencia relativa de los dos estratos, con respecto al estrato compuesto. Se puede demostrar que, aproximadamente

$$R = 1 + \frac{12\,c}{(1 + c)^2} \tag{17}$$

en la que c es el cociente de las frecuencias: $c = N_i/N_{i+1}$.

Se puede ver que R es mayor que uno y sólo es uno cuando c es cero o infinito. El valor máximo de R es 4 y sucede cuando $c = 1$, o sea, $N_i = N_{i+1}$, es decir, la subdivisión en estratos es más valiosa cuando las frecuencias de los intervalos vecinos son aproximadamente las mismas. Para $R = 1.20$ se requiere un valor de c cercano a 60.

El resultado anterior nos lleva a la conclusión de que el número de estratos debe ser el máximo posible. Para la estimación del error dentro de cada estrato se exige un mínimo de dos unidades en la muestra, y para este caso el máximo de estratos sería $p = n/2$. En algunos diseños se hace $p = n$, es decir, se toma una muestra de una unidad por estrato, lo cual maximiza R y una sobrestimación del error de muestreo se obtiene componiendo estratos vecinos.[3]

Formación de estratos. Unidades ordenadas por tamaño

En diversas ocasiones se dispone de la ordenación de las unidades que componen a la población en forma creciente o decreciente con respecto a una variable x que tomamos como tamaño característico de la unidad. También podemos tener el tamaño de cada unidad en tarjetas y es fácil hacer su ordenación. Teniendo esta secuencia de unidades la formación de estratos es fácil siguiendo el procedimiento que describimos a continuación.

[3] M. H. Hansen, W. N. Hurwitz y W. G. Madow, *Sample Survey Methods and Theory*, Vol. II, Nueva York, John Wiley, 1953.

Supongamos tener p estratos y que en el estrato i hemos obtenido una muestra al azar de n_i unidades. La varianza de X, el total estimado de la población, será

$$V(X) = \Sigma N_i^2 \frac{\sigma_i^2}{n_i}, \quad i = 1, 2, \ldots, p$$

Si ahora

$$\sigma_i^2 \leqslant G,$$

tendremos

$$V(X) \leqslant G \Sigma \frac{N_i^2}{n_i}$$

Por otra parte, si hiciéramos una afijación proporcional, es decir,

$$n_i = \frac{N_i}{N} n$$

y sustituyendo

$$V(X) \leqslant \frac{GN^2}{n}$$

y, si $\bar{x}$ es la estimación de la media de la población,

$$V(\bar{x}) < \frac{G}{n}$$

Si la amplitud total (recorrido)) de la población es A y esto se divide en p intervalos iguales, correspondiendo cada uno de ellos a un estrato, el recorrido de cada estrato será el mismo e igual a $A_e = A/p$. La varianza máxima con un intervalo A_e es $A_e^2/2$ y tomaría este valor cuando el estrato tuviera dos unidades y con valores en los extremos del intervalo. Por lo tanto,

$$V(\bar{x}) < \frac{1}{2n} \left(\frac{A}{p} \right)^2$$

Pero si suponemos que el tamaño de muestra para cada estrato es cuando menos igual a 2, $n > 2p$, y así podemos escribir,

$$V(\bar{x}) < \frac{A^2}{4p^3},$$

de donde,

$$p < \left(\frac{A^2}{4V(\bar{x})} \right)^{1/3} = \left(\frac{\frac{A}{\mu}}{2 \frac{\sigma_{\bar{x}}}{\mu}} \right)^{2/3} \qquad (18)$$

En la última expresión μ es la media de la población, $\sigma_{\bar{x}}/\mu$ es el coefi-

ciente de variación de la media de la muestra y su valor mide el grado de aproximación que nos proponemos alcanzar.

La expresión (18) proporciona el límite superior en el número de estratos en que dividimos la población. Consideramos que en muchas situaciones prácticas el valor máximo de la varianza de un estrato es inferior al señalado de $A_e^2/2$, porque habrá un número N_i relativamente grande y con valores de x distribuidos dentro del intervalo correspondiente. Por ello recomendamos el uso de la siguiente expresión (19) para determinar el número de estratos,

$$p = \left(\frac{\dfrac{A}{\mu}}{3\,\dfrac{\sigma_{\bar{x}}}{\mu}} \right)^{2/3} \tag{19}$$

Obtenido el valor de p con la fórmula anterior, determinamos $A_e = A/p$. El estrato i estará formado por las unidades que posean valores x comprendidos en el intervalo

$$x_{\text{máx}} - (i-1)\,A_e < x < x_{\text{máx}} - i \cdot A_e \tag{20}$$

$x_{\text{máx}}$ es el valor máximo de x que posean una o más unidades en la población.

Una vez determinados los estratos, conocemos los N_i. Para un tamaño total n, podemos afijar los n_i proporcionalmente a los N_i, o bien de acuerdo con la fórmula,

$$n_i \propto N_i \sqrt{\mu_i} \tag{21}$$

en la que μ_i es la media del estrato i. Esta última fórmula para n_i es la correspondiente a la expresión (10) suponiendo que los A_i son constantes e iguales a A_e. En esta afijación n_i será siempre mayor o igual a 2.

En toda nuestra discusión no hemos considerado el ajuste en el tamaño de la muestra por finitud de la población. Si esto se presenta se haría el ajuste conocido.

No debe extrañarnos que después de construir los estratos por el procedimiento descrito lleguemos a un número de ellos distinto de p. La diferencia será despreciable para nuestro propósito.

RESUMEN

En el estudio se desarrollan conceptos teóricos que determinan la influencia de la formación de estratos en la aproximación de los estimados de una población. Se examina primeramente el caso de una población que sigue una función de probabilidades continua y en seguida el caso de una función discreta. En el último caso se señala el criterio de formación de estratos cuando las unidades que compo-

nen a la población estén ordenadas conforme a la magnitud de la variable x, definido como tamaño de la unidad.

Determinados los estratos, el aspecto siguiente es determinar el tamaño de muestra en cada estrato. En este estudio se introduce el criterio de considerar como costo de investigar a la unidad i un valor inversamente proporcional a la magnitud x_i de su tamaño correspondiente. Se señalan las razones para proceder en esta forma.