ACTA MEXICANA DE CIENCIA Y TECNOLOGÍA
Vol. IV, Núm. 15, Julio-Septiembre, 1986
págs. 11-16

ESTRATIFICACIÓN ÓPTIMA

FRANCISCO CASANOVA DEL ÁNGEL
Sección de Graduados, Escuela Superior de Ingeniería y Arquitectura (ESIA)
Instituto Politécnico Nacional. México, D.F.

Resumen: Se hace una introducción a la estrificación óptima, parte integral de la teoría del sondeo estadístico, haciendo notar que es fundamental la precisión en las estimaciones de las características de la población tratada. Se presenta una parte teórica para la determinación del número óptimo de estratos y para la estimación óptima de las unidades de sondeo y llegar a la presentación de una fórmula que expresa el límite superior y minimiza la varianza.

OPTIMAL STRATIFICATION

Abstract: An introduction is presented on optimal stratification, which is an integral part of the theory of statistiçal sampling; emphasis is made on the capital importance of attaining a high degree of precision when estimating the characteristics of the population dealt with. This paper includes a theoretical part for the determination of the units to be surveyed; finally, a formula is presented expressing the upper limit and minimizing the variance.

INTRODUCCIÓN

En 1950, Tore Dalenius presenta la estratificación óptima como una técnica basada en la representación de la población mediante una función de densidad $f(x)$ donde la varianza del estimador x, para un tipo específico de diseño de la muestra a E-estratos, es tomada como una función de los puntos x_i de la estratificación. Dicho concepto fue presentado en el *Diario de Actualidades Escadinavas.*

En 1962, G.J. Glasser presenta en la revista del Instituto Internacional de Estadística una técnica para la cobertura completa de grandes unidades de sondeo en un estudio estadístico basado en las principales fórmulas expresadas por Tore Dalenius.

A continuación veremos generalidades de la teoría de sondeos y la estratificación. Se presenta una sección teórica para determinar el número óptimo de estratos. Se ve también la estimación óptima de las unidades de sondeo ya que el conjunto a sondear (población, universo) es inicialmente fragmentado en elementos distintos, denominados *unidades de sondeo.* El tema central es el punto óptimo de estratificación.

La última fórmula del artículo ha sido obtenida a partir de las fórmulas de Tore Dalenius, que expresa el límite superior y minimiza la varianza. Por último, se presenta una aplicación que comprueba el buen funcionamiento de esta fórmula.

1. GENERALIDADES

Cuando se trabaja la teoría de sondeos para una determinada población, uno se pregunta: ¿Qué es un estrato? ¿Qué es una estrato óptimo? ¿Cómo puede uno obtener un intervalo de confianza o un intervalo de aceptación? ¿Qué es una muestra estratificada? Veamos a continuación las respuestas a manera de definición.

Sea N una población de unidades dividida en subpoblaciones de $N_1, \ldots, N_n$ unidades respectivamente, tales que

$$N = N_1 + \ldots + N_n \tag{1.1}$$

Definición 1.1. Se le llama estrato a cada una de las subpoblaciones de la población dada.

En relación al estrato óptimo, por el momento se puede decir que depende de la elección de la talla de la muestra que uno toma por cada estrato E. En la segunda sección veremos la forma en que uno puede determinar el número óptimo de estratos.

Toda estimación de una media a partir de una muestra tiene un error aleatorio, positivo o negativo:

$$e = \mu - M \tag{1.2}$$

donde:

e = error aleatorio
μ = media de la muestra
M = media verdadera de la población

Observación 1. El error aleatorio tiene media nula.

El margen de error o intervalo de confianza a 0.95 es el que se obtiene de la estimación superior a dos veces el valor de su desviación estándar.

En relación a las muestras estratificadas, uno puede decir que los valores medios para el conjunto de unidades son estimados con media de valores tales que

$$\mu = \frac{N_1 \mu_1 + \ldots + N_n \mu_n}{N_1 + \ldots + N_n} \tag{1.3}$$

con N_n el número de unidades en ese estrato.

La varianza de la estimación está dada por

$$\sigma^2 = \frac{1}{N_1 + \ldots + N_n} \left[N_1^2 \frac{\sigma_1^2}{k_1} + \ldots + N_n^2 \frac{\sigma_n^2}{k_n} \right] \tag{1.4}$$

con k_i el tamaño relativo del i-ésimo estrato; de aquí se prueba que una vez que la suma de observaciones $k_1, \ldots, k_n$, se fija al principio, la expresión comprendida en el paréntesis tiene su mínimo en

$$\frac{N_1 \sigma_1}{k_1} = \ldots = \frac{N_n \sigma_n}{k_n} \tag{1.5}$$

Los resultados de este sondeo, tratados según la fórmula (1.4), permiten determinar a posteriori la precisión de la estimación. Si ella es insuficiente, es necesario completar la muestra o volver a principiar el cálculo utilizando entonces las estimaciones $\sigma_1, \ldots, \sigma_n$ para aplicar (1.5). La estimación por muestreo de una proporción es menos precisa cuando es poco elevada.

Un último punto a mencionar en esta sección de generalidades es que se considera a la estratificación como una técnica que produce precisión en las estimaciones de las características de la población.

2. DETERMINACIÓN ÓPTIMA DEL NÚMERO DE ESTRATOS

Aquí la discusión se limitará a determinar el número óptimo, E_{op}, de estratos.

Se puede representar la población mediante una función de frecuencia, $f(x)\ \forall\, x \in [a, b]$, con media μ que es calculada mediante un estimador lineal de la forma $\bar{x} = \sum \left\{ \frac{N_i}{N} \bar{x}_i \mid i = 1, \ldots, n \right\}$ tal que cada x_i se basa en una muestra aleatoria y $n_i \to \frac{N_i}{N_i} \sigma_i$.

Se considera el caso en el cual la función de frecuencia es de la forma $f(x) = v\ (v = \text{cte})$, que representa una población con selección natural.

Para esta función, la relación entre varianzas del estimador variable con $E > 1$ y $E = v$ está dada por:

$$\sigma^2(\bar{x} \mid E > 1) = \frac{1}{\eta^2} \sigma^2(\bar{x} \mid E = v) \tag{2.1}$$

donde E es el estrato óptimo introducido. Pero podemos preguntarnos, ¿qué pasa si aumenta el número de estratos? En este caso la varianza es reducida por un factor, i.e

$$\sigma^2(\bar{x} \mid r E) = \left(\frac{1}{r}\right)^2 \sigma^2(\bar{x} \mid E) \tag{2.2}$$

Existen casos especiales de estratificación variable,[1] basados en una variable específica de estratificación y relacionados con el estimador variable, $\bar{x}$, mediante la ecuación.

$$x = \Phi(e) + \eta e \tag{2.3}$$

[1] Ver referencia 4, pág. 128.

donde el estrato formado por la cortadura y la distribución de e no afecta la varianza de ηe. Porque si E aumenta, la componente Φ de $\sigma^2(\bar{x} \mid E)$ decrece, pero ηe no.

Restrinjamos el análisis del problema de determinar el número óptimo, E_{op}, de estratos, al caso para el cual, por cada E, la estratificación es óptima.

Observación 2. El problema de la estratificación es equivalente a determinar la cortadura.

Una relación entre la varianza de un estimador variable y un número m de estratos que dé una medida razonable del crecimiento en dos estratos sucesivos (i.e: de E − 1 a E) es

$$\sigma^2(\bar{x} \mid E) = \left(\frac{E-1}{E}\right)^2 \sigma^2(\bar{x} \mid E-1) \qquad (2.4)$$

Usando esta relación es posible hacer una inclusión de una muestra de medida N en la relación si el análisis se basa en (2.2). Si tomamos un valor E_o al que $E - 1 = E_o$ ó $E > E_o$ con (2.4) válida, entonces

$$\sigma^2(\bar{x} \mid E) = \frac{E_o^2}{E^2}\sigma^2(\bar{x} \mid E_o) \qquad (2.5)$$

con $\sigma^2(\bar{x} \mid E)$ representando la varianza de un estimador variable $\bar{x}$ en correspondencia a un número de estratos.

Cuando uno construye la tabla de valores $E = 1$ a E_o, aparece un valor constante como en (2.2), r_o, es decir

$$\sigma^2(\bar{x} \mid E_o) \doteq \frac{1}{r_o}\sigma^2(\bar{x} \mid E = 1) \qquad (2.6)$$

obteniendo de la ecuación (2.2)

$$\sigma^2(\bar{x} \mid E) = \sigma^2(\bar{x} \mid r E_o) = \frac{1}{r_o^2}\frac{1}{r_o^2}\sigma^2(\bar{x} \mid E = 1) \qquad (2.7)$$

Para el caso donde exista una función varianza es posible escribir la varianza de un estimador variable en relación a un gran número de estratos $E = r E_o$ de la forma

$$\sigma^2(\bar{x} \mid r E_o) = \frac{1}{r^2}\frac{1}{r_o^2}\frac{\sigma^2}{N_o} + R \qquad (2.8)$$

donde R no depende ni del número de estratos ni de la medida de la muestra.

Para completar este pequeño análisis se requiere que la función costo, C, dependa de E

$$C(\bar{x} \mid r E_o) = r E_o C_s + n C_n \qquad (2.9)$$

de la cual puede obtenerse el número óptimo de estratos E_{op} de la forma (2.10) con $C_s \geqslant 2 C_n$

$$E_{op} = \eta_{op} + 2 C_n / C_s \qquad (2.10)$$

3. ESTIMACIÓN ÓPTIMA DE UNIDADES EN UN ESTUDIO ESTADÍSTICO

La técnica de la estratificación óptima está basada en la representación de la población mediante una función de densidad $f(x)$, se puede ver que la talla total de la población es

$$\nu \int_a^b f(x)\,dx = \nu \qquad (3.1)$$

con ν las unidades de los estratos y una media de la forma

$$\mu = \int_a^b x f(x)\,dx \qquad \forall x \in [a, b] \qquad (3.2)$$

tal, que x divide a la población en dos estratos. ¿Pero qué pasa si la población es muy asimétrica?

Consideremos

$$\nu_1 = \nu \int_a^x f(x)\,dx \qquad (3.3)$$

y

$$\nu_2 = \nu - \nu_1 \qquad (3.4)$$

con medias μ_1 y μ_2 respectivamente. De aquí se tiene que:

$$\mu = w_1\mu_1 + \mu_2 w_2 = \frac{\nu_1}{\nu}\mu_1 + \left(1 - \frac{\nu_1}{\nu}\right)\mu_2 \qquad (3.5)$$

Si tomamos a i como las grandes unidades y a $\nu - i$ como el residuo, las ecuaciones (3.3) y (3.4) se pueden reescribir como

$$\nu - i = \nu \int_a^{x \leqslant b} f(x)\,dx \qquad (3.6)$$

y

$$\nu_2 = \nu - \nu \int_a^x f(x)\,dx = \nu\left(1 - \int_a^x f(x)\,dx\right) \qquad (3.7)$$

$$= \nu - \nu + i = i$$

con v_2 las unidades grandes y $w_1 = \frac{v-i}{v}$ y $w_2 = \frac{i}{v}$, de donde la media (3.5) se puede escribir como

$$\mu = \frac{v-i}{v}\mu_{v-i} + \frac{i}{v}\mu_i \tag{3.8}$$

De la misma manera, la varianza tiene la forma

$$\sigma^2(\bar{x} \mid E) = w_1^2 \frac{v-n_1}{v_1-1} \frac{\sigma^2_{v-i}(\bar{x} \mid E)}{n_1}$$

$$= \left(\frac{v-i}{v}\right)^2 \frac{v-i-n+i}{v-i-1} \frac{\sigma^2_{v-i}(\bar{x} \mid E)}{n-i}$$

$$= \left(\frac{v-i}{v}\right)^2 \frac{v-n}{v-i-1} \frac{\sigma^2_{v-i}(\bar{x} \mid E)}{n-i} \tag{3.9}$$

que son los mismos resultados que derivan Tore Dalenius y G.J. Glasser en artículos individuales, pero éste con aplicación específica a poblaciones muy asimétricas.

Si hacemos $i = 0$ en la ecuación (3.9), obtenemos

$$\sigma^2(\bar{x} \mid E) = \frac{\sigma^2_v(\bar{x} \mid E)}{n} \frac{v_1-n}{v_1-1} \tag{3.10}$$

que es la varianza reducida a una muestra aleatoria sin restricción y el mismo resultado es deducido por Tore Dalenius con una condición de aproximación.

Al multiplicar la población y la varianza (3.8) se obtiene un estimador para la población total

$$\sigma^2(v\mu) = v^2\sigma^2(\mu) \tag{3.11}$$

4. PUNTO ÓPTIMO DE ESTRATIFICACIÓN

El punto óptimo, $\bar{x}$, de estratificación es el punto en el cual $\sigma^2(\bar{x} \mid E)$ es un mínimo. Tore Dalenius utiliza la ecuación (3.10), es decir

$$\mu_1 \sigma_1^2(\bar{x} \mid E) = \mu_1^2 \frac{v-n}{v-1} \frac{\sigma^2_v(\bar{x} \mid E)}{n} \tag{4.1}$$

hace la diferenciación usando la aproximación $\frac{v-n}{v-1} = \frac{v-n}{v}$ y llega a la condición de estratificación óptima

$$(x-\mu_1)^2 = \frac{\frac{v_1}{v}\sigma^2_v(\bar{x} \mid E)}{\frac{n}{v} - 1 + \frac{v_1}{v}} \tag{4.2}$$

G. J. Glasser usa la ecuación (3.9) (es la misma que la ecuación (3.10) con $i \neq 0$) con la condición $i = m-1$ o $i = m+1$ donde m es el número óptimo de estratos, obteniéndose

$$\overset{*}{x} = \mu_{v-m} + \sqrt{\frac{v-m}{n-m}}\,\sigma_{v-m}(\bar{x} \mid E) \tag{4.3}$$

Existe otra interpretación al problema, la que da W. G. Cochran, que difiere muy poco de las interpretaciones dadas anteriormente.

En una muestra aleatoria estratificada con una función de costo de la forma

$$C = c_o + \Sigma c_n n_n \tag{4.4}$$

la varianza de la media estimada $\bar{x}_{est}$ tiene un mínimo cuando el número de unidades de la muestra n_n es proporcional a $N_n S_n / \sqrt{c\,n}$, donde la varianza verdadera S_n^2 es

$$S_n^2 = \frac{\Sigma\left\{(x_{n_i} - \bar{x}_n)^2 \mid i = 1, \ldots, N_n\right\}}{N_n - 1} \tag{4.5}$$

con N_n el número total de unidades.

El problema es minimizar $\sigma^2(\bar{x}_{est})$, es decir

$$\sigma^2(\bar{x}_{est}) = \Sigma\left\{\frac{w_n^2 S_n^2}{n_n} \mid n = 1, \ldots, N_n\right\} - \Sigma\left\{\frac{w_n S_n^2}{N_n} \mid n = 1, \ldots, N_n\right\} \tag{4.6}$$

con la reducción $c_1 n_1 + \ldots + c_n n_n = C - c_o$. Usando el método de Lagrange se obtiene

$$\frac{n_n}{n_o} = \frac{N_n S_n / \sqrt{cn}}{\Sigma\left\{N_n S_n / \sqrt{cn} \mid n = 1, \ldots, N_n\right\}} = \frac{N_n S_n}{\Sigma\left\{N_n S_n \mid n = 1, \ldots, N_n\right\}} \tag{4.7}$$

Si consideramos el costo por unidad igual en cada estrato, entonces, para una talla total de muestra n se tiene

$$n_n = n_o \frac{N_n S_n}{\sum \{ N_n S_n \mid n = 1, \ldots, N_n \}} \tag{4.8}$$

que es el valor que minimiza la varianza del estimador, $\bar{x}_{est}$

Ahora bien, una forma de expresar el límite superior necesario e importante, porque da la cortadura, es

$$\sigma^2_{\min}(\bar{x} \mid E) = \frac{N_n - n_n}{n_n N_n}$$
$$\sum \{ N_n \sigma_n (\bar{x} \mid E) \mid n = 1, \ldots, N_n \} \tag{4.9}$$

A partir de esta fórmula es posible obtener, mediante algunos pasos algebraicos, las fórmulas para minimizar la varianza, con sus respectivas restricciones, que Tore Dalenius y G.J. Glasser han mostrado en diferentes artículos.

5 APLICACIÓN

Daremos aquí un ejemplo que describe una muestra de población estabular en el país con un total de 196 establos y 56 472 becerreros. La estratificación se basa en el empleo de σ_n. Los primeros cinco estratos se construyen sobre la talla de los establos y el último pertenece a un establo piloto. (Tabla 5.1.)

TABLA 5.1. Datos para la estimación de la talla de la muestra

Estratos	N_n	σ_n	$N_n\sigma_n$	n_n
1	13	325	4225	9
2	18	190	3420	7
3	26	189	4914	10
4	42	82	3444	7
5	73	86	6278	13
6	24	190	4560	10
	196		26841	56

Como los valores N_n, σ_n y $N_n\sigma_n$ son conocidos, es posible determinar n. El error estándar es

$$(0.05)\ (56\ 472) = 2\ 823.6 \approx 2\ 824$$

de donde la varianza es

$$s = (2\ 824)^2 = 7\ 974\ 976$$

utilizando la ecuación (4.7) con un costo igual en todos los estratos tal que una primera utilización a n_o es

$$n_o = \frac{(\Sigma N_n S_n)^2}{s} = \frac{(26\ 841)^2}{7\ 974\ 976} = 90.34$$

de donde la n_n óptima es

$$n_n = \frac{n_o}{1 + \frac{\sum N_n S_n^2}{s}} = \frac{90.34}{1 + \frac{4\ 640\ 387}{7\ 974\ 976}} = 57.19$$

Es decir, la talla mejor para la muestra es 57.

Basando los cálculos para minimizar la varianza en los primeros tres estratos con los datos de la tabla 5.2 para la muestra estratificada (cálculos en la tabla 5.3), las fórmulas (4.6) y (4.8) dan los resultados

$$\sigma^2(\bar{x}_{est}) = \Sigma \frac{w_n^2 S_n^2}{n_n} - \Sigma \frac{w_n S_n^2}{N_n}$$

$$= 0.0118 - 0.00742 = 0.0044$$

y

$$\sigma^2_{min}(\bar{x} \mid E) = \frac{N_n - n_n}{n_n N_n} \Sigma \{ N_n \sigma_n^2 \}$$

$$= (0.0020)\ (24131) = 0.0504$$

La estritificación aparece en la reducción de la varianza en la relación a 0.088.

TABLA 5.2

Estratos	N_n	n_n	$\bar{x}_n$	σ_n^2
1	13	9	2.200	1.615
2	18	7	1.688	0.063
3	26	10	0.992	0.077
	57	26		1.755

TABLA 5.3 Cálculos.

Estratos	w_n	$w_n S_n^2$	$w_n S_n^2/n_n$	$w_n^2 S_n^2/n_n$	$w_n \bar{x}_n$
1	0.227	0.368	0.0409	0.0093	0.5016
2	0.315	0.019	0.0028	0.0009	0.5176
3	0.457	0.035	0.0035	0.0016	0.4524
	0.999	0.422	0.0472	0.0118	1.4716

6. CONCLUSIONES

Aunque el tema presentado es de interés teórico se ve que en estudios estadísticos para poblaciones muy asimétricas es bueno incluir unidades de sondeo y elegir una muestra aleatoria de unidades.

La fórmula presentada expresa el límite superior de la cortadura y no tiene relación directa con el promedio de la población; a pesar de todo da una apreciable mejoría a la varianza y expresa el límite superior de la cortadura en términos del promedio de la población.

BIBLIOGRAFÍA

1. **Cochran, W. G.**, 1961. *Sampling Techniques*, p 106-108.
2. **Dalenius, Tore**, 1952. The problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*, p. 61-70.
3. _____, 1957. Sampling in Sweden. Contributions to the methods and theories of sample survey practice, UPPSALA.
4. **Glasser, G. J.**, 1962. One the complete converage for large units in a statistical study. *Review of the International Statistical Institute*, Vol. 30-1, p. 28-32.
5. **Herniaux, G.**, 1971. Initiation aux sondages. Masson et Cie.
6. **Thionet, Pierre**, 1970. Analyse des distributions par sondage. *Journal de la Sicietè de Statistique de Paris*, cent onzieme annèe. 3-eme trimestre.
7. _____, 1969. Sur les sondages avec probabilitès inègales. *Revue de Statistique Appliqueè*, vol. XVII, No. 4.
8. _____, 1977. Apercu sur la thèorie statistique des sondages. Universite Paris IX Dauphine, fasicule 4.
9. _____, 1965. Comment Choiser un èchantillon dans une population ou les sujets sont d'importance tres diffèrents. Facultè des Sciences de Poitiers.