Revista Colombiana de Estadística
Vol. **22** No. 2(1999) *págs 17 - 25*

El cubrimiento de una muestra: estimadores clásicos vs predicción

CARLOS BOUZA *

ABSTRACT.
The problem of determining the coverage of a sample is studied. A stratified model is worked out. An alternative to some classic estimators is proposed. A superpopulation model is used and a predictor is developed. Its asymptotic normallity depends only of the size of the sample. A Monte Carlo experiment evaluates the behavior of the different models.

Resumen.
El problema de determinar el cubrimiento de una muestra es estudiado. Un modelo estratificado es desarrollado. Una alternativa respecto a algunos estimadores clásicos es propuesta. Ella se basa en un modelo superpoblacional y un predictor es obtenido. Su normalidad asintótica depende sólo del tamaño de la muestra. Un experimento de Monte Carlo evalúa el comportamiento de los diferentes modelos.

Palabras claves: *Coverage, predictor, superpopulation model.*

1. Introducción

Considere una población con un número desconocido Δ de clases. Una muestra s es seleccionada mediante Muestreo Simple Aleatorio (MSA). Los objetos observados y las clases a que pertenecen son identificados. Éste es un problema de gran importancia en muchas aplicaciones usuales en diversas ciencias.

Contemporáneamente éste es planteado con frecuencia en estudios Medio Ambientales. Algunos ejemplos, en estos problemas y otros similares, son los siguientes:

1. Identificar el número de especies en una región o los contaminantes presentes en los desechos de una empresa situada en ella.

*Profesor. Facultad de Matemáticas y Computación. Universidad de la Habana. *E-mail*:bouza@matcom.uh.cu

2. La determinación del léxico personal de un autor o las enfermedades provocadas por el nivel de toxidad en el aire en una área de salud.
3. Los defectos presentes en las unidades de una línea de productos de una fábrica o las malformaciones en la vegetación circundante a una central atomoeléctrica.
4. Identificar los tipos de plagas que están presentes en un cultivo o las fuentes de los polutantes observados en un río.

Estudios de este tipo pueden remontarse al trabajo de Fisher-Corbert-William (1943) en Ecología Aplicada y son desarrollados en múltiples publicaciones posteriores con Efron-Thisted (1976), Engen (1978) y Chao (1981), por ejemplo.

Estos problemas plantean varias interrogantes. Una de ellas es la de determinar cotas para Δ o estimarle. Otro es fijar el cubrimiento obtenido por la muestra s de las clases desconocidas a priori. En este trabajo se abordará este último problema y se desarrollarán Teoremas Centrales del Límite (TCL) para los correspondientes estimadores. Para ello se analizará el comportamiento de los estimadores propuestos por Good (1953) y Esty (1986). Una alternativa estratificada es desarrollada y analizada en el marco de la Teoría Clásica.

Las hipótesis que sustentan los TCL para los estimadores son algo fuertes pues se basan en la equiprobabilidad de las clases. Un predictor es propuesto usando un modelo superpoblacional del tipo "Regresión Lineal", que en este caso es de la familia de los Modelos Lineales Generalizados. Este no depende de hipótesis sobre las probabilidades de observar una clase para garantizar la distribución normal asintótica del predictor.

Los datos de un estudio de la infestación de campos de caña son usados para estudiar el comportamiento de la eficiencia de los estimadores y del predictor analizados. Experimentos de Monte Carlo son desarrollados con este fin.

2. El problema del cubrimiento y el uso de diseños

Sea U una población particionada, a partir de un cierto criterio en clases. Este es un problema presente en muchos estudios Medio Ambientales y Ecológicos como los apuntados en la introducción.

El modelo estadístico se basa en establecer que existen Δ urnas. El *experimento muestral* es descrito por la ubicación aleatoria de las n unidades muestrales en las urnas. El resultado es el número de urnas D ocupadas. Éste es un estadígrafo suficiente para Δ.

El cubrimiento de la muestra s es definido como:

$$\theta = \sum_t \pi_t \, I_t = \sum_{t \geq s} \pi_t$$

donde π_t es la probabilidad de que la clase C_t sea detectada al evaluar la muestra, $t \geq s$ denota este hecho y $I_t = 1(0)$ si el número de unidades de C_t

en s es mayor que uno (en otro caso). El problema se plantea la estimación de θ es diferente del problema estadístico usual pués este no es un parámetro sino un valor relacionado con la muestra observada. Es decir, que está condicionado a la muestra s seleccionada. Good (1953) propuso como estimador de θ a:

$$\theta_1 = 1 - n_1/n$$

donden n_1 es el número de clases observada una vez. Este ha sido utilizado profusamente en el estudio de la abundancia de especies, Erigen (1978), Starr (1979) y Chao (1981).

Otra solución se asocia la estimación de Δ mediante la solución de la ecuación

$$D = \Delta^*(1 - \exp(-n/\Delta^*))$$

y usar

$$\theta_2 = D/\Delta^*$$

como estimador, Esty (1986).

El estudio de la convergencia de la ley de estos estimadores se lleva a cabo al fijar la hipótesis de que $\pi_t = \pi$ para todas las clases. El teorema siguiente fija la ley normal que es su límite.

Teorema.
Si $\pi_t = \pi$ para toda clase C_t, $n \to \infty$, $\Delta \to \infty$ y $n/\Delta \to -ln(1-\alpha), \alpha \in (0,1)$, entonces:

1. $(\theta - \theta_1)n^{1/2} \to^D N\{0, \alpha^2(-(1-\alpha)ln(1-\alpha)/(\alpha + (1-\alpha)ln(1-\alpha)))\}$
2. $(\theta - \theta_2)n^{1/2} \to^D N\{0, ((1-\alpha)(\alpha - ln(1-\alpha)))\}$

De este resultado se deduce facilmente que la eficiencia de θ_1 respecto a θ_2 es

$$E_{12} = (-ln(1-\alpha))\alpha^2/(\alpha + (1-\alpha)(\alpha - ln(1-\alpha)) > 0,85$$

En este modelo se asume que el diseño es el muestreo simple aleatorio. Sin embargo, en las aplicaciones éste es raramente empleado. Por tanto estos estimadores pueden ser utilizados como núcleo para el desarrollo de estrategias más complejas.

Si se usara el muestreo estratificado en el que la población es particionada en L estratos $U_1, \ldots, U_I$ y una muestra aleatoria independiente es seleccionada en cada uno, tenemos que cuando

$$W_j = \text{Prob } (i \in U_j) = N_j/N,$$

probabilidad marginal de que una unidad muestral pertenezca al estrato j–ésimo, $j = 1, \ldots, L$, y

$$\pi_{j(t)} = \text{Prob } (i \in C_t \mid U_j),$$

la probabilidad de hallar una unidad de la clase t–ésima en el estrato U_j, como

$$\sum_{j=1}^{L} W_t \pi_{j(t)} = \pi_t$$

podemos expresar el cubrimiento de s usando la relación

$$\theta = \sum_{t \approx > s} \sum_{j=1}^{L} W_j \pi_{jt}$$

Tomando $\theta_{h(j)}$ como la versión del estimador θ_h, $h = 1, 2$, en U_j obtenemos las contrapartidas de los estimadores del muestreo simple aleatorio.

$$\theta^*_{h(e)} = \sum_{t \approx > s} \sum_{j=1}^{L} W_j \theta_{h(j)}$$

cuya varianza es

$$V(\theta^*_{h(e)}) = \sum_{t \approx > a} \sum_{j=1}^{L} W_j^2 V(\theta_{h(e)})$$

en este diseño las hipótesis de equiprobabilidad pueden restringirse a un cumplimiento dentro de los estratos: $\pi_{jt} = \pi_{(t)}$ para todo $t = 1, \ldots, L$. Además debemos suponer que $n_j \to \infty$, $\Delta \to \infty$ y $n_j/\Delta \to -ln(1-\alpha_j), \alpha_j \in (0,1)$ son válidas también en todos los estratos. Esto es menos fuerte que usar las condiciones fijadas en el teorema sobre la población. Por ello debemos diseñar la construcción de estratos buscando la homogeneidad de las clases dentro de las subpoblaciones en el sentido de la equiprobabilidad. Por ejemplo, deberemos dividir los datos suministrados por las estaciones de una red de monitoreo buscando que la frecuencia de los polutantes sea similar y así determinar los $U'_j s$. Esto sería diseñable por los expertos estableciendo que sean agrupadas en el mismo estrato estaciones que monitoreen fuentes donde las fábricas producen desechos similares.

La normalidad de $\theta_{h(j)}$ en cada estrato garantiza la de $\theta^*_{h(e)}$.

Los errores correspondientes son:

$$V(\theta^*_{1(e)}) = \sum_{t \approx > s} \sum_{j=1}^{L} W_j \alpha_j^2 \left(-(1-\alpha_j) ln(1-\alpha)/n(\alpha_j + (1-\alpha_j) ln(1-\alpha_j))\right)$$

y

$$V(\theta^*_{2(e)}) = \sum_{t \approx > s} \sum_{j=1}^{L} W_j \left((1-\alpha_j)(\alpha_j - ln(1-\alpha_j))\right)/n$$

si utilizamos afijación proporcional ($n_j = nW_j$, para todo $j = 1, \ldots, L$).

Note que cuando $\alpha_j = \alpha$ para todo j este diseño es equivalente al simple aleatorio en términos del error. Sin embargo, cuando es esperable que, si los estratos son construidos adecuadamente, las unidades en ellos sean muy similares internamente una ganancia en precisión se considerable es obtenida gracias al uso de la estratificación utilizando las propuestas ya apuntadas.

3. Un enfoque superpoblacional

Las hipótesis utilizadas en los modelos clásicos son muy fuertes, en especial el de equiprobabilidad, en la mayoría de las aplicaciones. Pesnsando en los ejemplos citados en la introducción tenemos que ciertos contaminantes aparecen más raramente que otros, algunas plagas son más prolíferas o el habitat les es más propício, algunas dolencias respiratorias son más comunes ante el enrarecimiento del aire por una cierta emisión de gases, etc.

Una solución es utilizar la información de los expertos, no en fijar si la equiprobabilidad es válida y los valores de α sino en elicitar un cierto sentido una distribución a priori y utilizar el enfoque Bayesiano ecléctrico asociado al uso de una superpoblación. Nuestra propuesta es que el decisor puede modelar la relación entre la variable Y_i, generada por el ditribución a priori, y la probabilidad desconocida de que la unidad pertenezca a C_t mediante la relación

$$Y_i = \pi_t + \varepsilon_i$$

donde t representa la clase a la que la unidad i pertenece y los errores son independientes entre si.

Denotemos por μ el modelo superpoblacional (MSP) que genera la variable aleatoria Y_i. La estocasticidad que es modelada por el MSP permite estudiar la variabilidad de la medición en un cierto instante.

Suponiendo que la familia de distribuciones apriori es caracterizada por $E_\mu(\varepsilon_i) = 0$, y $V_\mu(\varepsilon_i) = \sigma_t^2 = \pi_t(1 - \pi_t)$, para todo $i \in C_t$, tanto σ_t^2 como π_t son ralizaciones condicionadas a un cierto instante. Omitimos el subindice correspondiente al "instante" dada su insignificancia en el análisis teórico que realizamos.

La definición inicial de θ establece que cada clase posee el mismo peso: $P_t = 1$ por lo que

$$\sum_{t=1}^{D} P_t = D$$

La muestra provee n valores de

$$Y_i = \begin{cases} 1, & \text{si } i \in C_t \\ 0, & \text{otro caso} \end{cases}$$

por tanto

$$\sum_{i \in C_t} Y_i/n = \pi_t^*$$

Entonces, un predictor del cubrimiento está dado por

$$\theta_3 = \sum_{t \in s} \sum_{i \in C_t} Y_i/n(t)D$$

donde

$$n(t) = \sum_{i \in C_t} P_t$$

Note que como $E_\mu(\theta_3) = \sum_{t \in s} \pi_t$

$$E_\mu(\theta_3) = \sum_{t \in s} \pi_t$$

y el predictor es modelo-insesgado para el cubrimiento. El error bajo el modelo es

$$V_\mu(\theta_3) = \sum_{t \in s} \pi_t(1 - \pi_t)/n(t) = \sum_{t \in s} \sigma_t^2/n(t)$$

Este es un caso particular del modelo desarrollado por Bouza (1996) como una particularización del propuesto por Pothoff et. al. (1992). En estos modelos $n(t)$ es conocido como tamaño de muestra equivalente.

Es claro que

$$Z_t = \sum_{i \in C_t} Y_i$$

sigue, condicionada a la muestra observada, una distribución Binomial $B\{n(t), \pi_t\}$. Por tanto, si $n(t)$ es suficientemente grande para la convergencia a la normal

$$Z_t^* = (Z_t/n(t) - \pi_t)n(t)^{1/2} \rightarrow N\{0, \sigma_t^2\}$$

Esto nos permite establecer la validez de la siguiente proposicin.

Proposición. *Si $n(t)$ es suficientemente grande cuando μ es válido entonces*

$$(\theta_3 - \theta) \rightarrow^D N\{0, \sum_{t \in s} \sigma_t^2/n(t)\}$$

Demostración.

Es claro que

$$E_\mu(\sum_{t=1}^{D} Z_t) = \sum_{t=1}^{D} n(t)\pi_t$$

y como

$$\sigma/n = (\sum_{t=1}^{D} n(t)\pi_t(1-\pi_t))^{1/2}/n \to 0$$

la sucesión obedece la Ley Central del Límite, Friedst-Gray (1997), de lo que se sigue el resultado enunciado. □

Si $\pi_t = \pi$ para todo t entonces $\sigma_t^2 = \pi_t(1-\pi_t)$ y

$$V(\theta_3) = \pi_t(1-\pi_t)\sum_{t=1}^{D} 1/n(t)$$

siendo la eficiencia del predictor θ_3 respecto a θ_1

$$E_{13} = \alpha^2(-(1-\alpha)ln(1-\alpha))(\sum_{t=1}^{D} 1/n(t))/(\alpha + (1-\alpha)ln(1-\alpha)\pi(1-\pi)n)$$

si se espera que $n(t) \cong \pi$ entonces el predictor propuesto es el más eficiente. Si $\alpha \downarrow 0$ entonces el valor del término a la izquierda de la inecuación decrece por lo que θ_3 será más recomendable cuando obtenemos un valor pequeño de α.

Al comparar el predictor con θ_2 obtenemos como expresión de la eficiencia a

$$E_{23} = (1-\alpha)(\alpha - ln(1-\alpha))/\pi(1-\pi)$$

Si $\alpha \cong \pi$ esta se reduce a

$$E_{23} \cong 1 - ln(1-\pi)/\pi$$

que está cerca de uno en general. La preferencia por el predictor es avalada en este caso por el hecho de que la equiprobabilidad de las clases es poco frecuente. Además, la hipótesis de que $\Delta \to \infty$ es también rara en la práctica.

4. Análisis de experimentos de Monte Carlo

Utilizamos los datos de un estudio de la infestación de campos de caña por plagas. Los agrónomos detectaron la existencia de plagas pero estas eran resistentes a los tratamientos usuales. Por ello acudieron entomólogos para establecer la estructura de los infestantes para diseñar una política de fumigación adecuada. El número de clases observadas estuvo entre 4 y 10 en los campos estudiados. Estos datos los usamos en nuestro experimento para confeccionar una población artificial de sitios de muestreo. Se generaron muestras con fracciones del 5%, 10% y del 20%. Las estimaciones fueron computadas a partir de los resultados de experimentos de Monte Carlo.

Los estratos fueron formados tomando en cuenta que se tuviese una proporción similar de las plagas, clases, lo más equiprobables posibles. Las muestras fueron seleccionadas mediante muestreo simple aleatorio con reemplazo de toda la población y se computaron θ_1, θ_2 y θ_3. De cada estrato se seleccionó una muestra usando la misma proporción y se computaron los estimadores correspondientes.

El número de muestras generado fue de 100 y se evaluaron

$$\delta_{i(e)} = \sum_{s=1}^{100} \mid \theta - \theta_i \mid_s \sum_{s=1}^{100} \mid \theta - \theta_{i(e)} \mid_s, \; i = 1, 2, 3$$

y

$$\delta_{12} = \sum_{s=1}^{100} \mid \theta - \theta_{1(e)} \mid_s / \sum_{s=1}^{100} \mid \theta - \theta_{2(e)} \mid_s .$$

Los resultados para los estimadores se tabulan en la tabla subsiguiente.

Tabla 1. Desviación Relativa en %: Métodos basados en diseños de muestreo

Fracción	$\delta_{1(e)}$	Máx($\delta_{1(e)}$)	$\delta_{2(e)}$	Máx($\delta_{2(e)}$)	δ_{12}	Máx(δ_{12})
0,05	88,2	93,9	82,6	88,6	103,7	144,9
0,10	80,3	94,0	75,2	90,7	117,5	132,9
0,20	62,1	63,3	55,1	74,3	11,7	112,2

Como se nota el uso de la estratificación fue más eficiente en todos los experimentos. Por otra parte la versión estratificada del estimador de Good fue la menos precisa.

Fueron computadas para las 100 muestras generadas las eficiencias

$$E_{f(ij)} = \sum_{s=1}^{100} E_{ij}/100, \; i, j = 1, 2, 3$$

$$E_{f(e(i)3)} = \sum_{s=1}^{100} V(\theta^*_{i(e)})_s / 100 V(\theta_3)_s \; i = 1, 2$$

Ellas nos permiten comparar el predictor con los estimadores. Los resultados obtenidos aparecen en la Tabla 2.

Tabla 2. Eficiencia Promedio en %: diseños de muestreo vs predictor

Fracción	$E_{f(13)}$	Mín$(E_{f(13)})$	$E_{f(23)}$	Mín$(E_{f(23)})$	$E_{f(e(1)3)}$	Mín $E_{(e1(3))}$	$E_{(e(2)3)}$	Mín $E_{f(e(2)3)}$
0,05	145,7	101,1	127,9	113,2	117,7	101,4	108,4	101,1
0,10	135,8	100,1	119,2	110,7	112,6	103,6	102,4	95,5
0,20	121,9	102,7	110,6	105,3	109,3	100,6	101,2	95,2

Note que el predictor posee mejor comportamiento que los estimadores basados en el muestreo simple en términos del promedio de la eficiencia. Sin embargo, es posible que una muestra estratificada sea más eficiente que el predictor como lo denota el análisis de la eficiencia mínima de los estimadores basados en el diseño en el que, para fracciones mayores que el 0,05, se obtuvieron mejores resultados para $\theta^*_{2(e)}$.

Reconocimientos: Este trabajo se llevó a cabo parcialmente mientras el autor desarrollaba una visita a la Universidad Veracruzana durante 1997 amparado por un Proyecto Fornes.

Referencias

1. Bouza C., *Linear rank tests derived from a superpopulation model*, Biometrical J. **37** (1995), 497–506.
2. Chao A., *On estimating the probability of discovering a new spacies*, Ann. Stat. **9** (1981), 1339–1342.
3. Efron, B. y Thisted, R., *Estimating the number of unseen species: how many words dis Shakespeare Know?*, Biometrika **63** (1976), 435–447.
4. Engen, S., *Stochastic abundance models*, Halsted Press, N. York (1978).
5. Esty, W. W., *The efficiency of Good's nonparametric coverage estimator*, Annals of Stat. **14** (1986), 1–9.
6. Fisher, R. A., Corbert, A. S. y William S. C. B., *The relation between the number of species and the number of individuals in a random sample of an animal population*, J. of Animal E. **12** (1943), 42–58.
7. Friedst, B. y Gray L., *A modern approach to probability theory*, Birkhauser, Boston (1997).
8. Good, I., *The population frequency of species and the estimator of the population parameter*, Biometrika **43** (1953), 45–63.
9. Pothoff, R. R., Woodbury, M. A. y Manton, K. G., *Equivalent samples size and equivalent degree of freedom for refinements for inference weight under super population models*, J. Amer. Stat. Ass. **87** (1992), 383–396.
10. Starr, N, *Linear estimation of the probability of discovering a new spacies*, Annals Stat **7** (1979), 644–652.