EL CUBRIMIENTO DE UNA MUESTRA: ESTIMACIÓN Y PREDICCIÓN

Carlos N. Bouza[1], Facultad de Matemática y Computación, Universidad de La Habana

RESUMEN

El problema de estimar o predecir el número de clases en que se divide una población U está presente en muchas aplicaciones y plantea varias interrogantes a la estadística. Muchas de ellas no están resueltas Good (1953) y Esty (1986) desarrollaron Teoremas Centrales del Límite. Sus hipótesis son algo fuertes. En este trabajo proponemos el uso de un modelo superpoblacional del tipo regresión. Su aplicación en este caso nos lleva al Modelo Lineal Generalizado como descriptor del comportamiento del cubrimiento. Este no depende de la hipótesis de equiprobabilidad para garantizar la distribución normal asintótica del predictor. Para analizar el comportamiento de las diversas alternativas para caracterizar el cubrimiento de la muestra utilizamos datos de dos investigaciones. Una son los datos obtenidos del estudio de la infestación de campos de caña por una plaga en Cuba. El otro es un estudio de biodiversidad forestal en México.

ABSTRACT

The problem of estimating or predicting the number of classes in population U is present in many applications and it outlines several queries to statistics. Many of them have been not solved yet. Good (1953) and Esty (1986) developed Central Limit Theorems. Their hypothesis are somewhat strong. In this paper we propose the use of a regression type superpopulation model. Its application in this case leads to a Generalized Linear Model as a describer of the behaviour of the coverage. It does not depend on the equiprobability hypothesis to guarantee the asymptotic normal distribution of the predictor.

To analyse the behaviour of the diverse alternatives that characterize the coverage of the sample we use data of two investigations. One of them is the data obtained from the study of the infestation of cane fields by a plague in Cuba. The other one consists is a study of forest biodiversity in Mexico.

Key words: Asymptotic normality, stochastic models. Accuracy. Generalized Linear Model.

MSC: 62D05

1. INTRODUCCIÓN

El problema de estimar o predecir el número de clases en que se divide una población U está presente en muchas aplicaciones. Lo usual en la Teoría de Muestreo es que el decisor conoce las características que determinan las subpoblaciones y su cantidad. Este es el caso de la estratificación, la conglomeración y los dominios de estudio por ejemplo. Consideremos que U está formada por un número desconocido Δ de clases. Se toma un muestra $s \subset U$ usando muestreo simple aleatorio (MSA). Los objetos observados son clasificados en clases al ser identificados y considerando la similitud y diferenciación entre los grupos que determinan ellos. Este problema difiere del problema de la post estratificación sensiblemente. En la post estratificación conocemos los estratos y clasificamos los objetos después de tomar la muestra. Ahora lo que deseamos es conocer el número de clases. Podemos considerar que esta es una tarea de los métodos de aglomeración comunes en la paquetería estadística. Sin embargo el problema de fijar un valor Δ es da gran importancia en muchas aplicaciones. Si consideramos los problemas de la arqueología veremos que ese es el objetivo del investigador en ciertos momentos. Él analiza las piezas encontradas, determina grupos y piensa que en número de grupos formados es menor que los que existieron. En particular en los problemas medio ambientales es común una problemática similar pues se quiere conocer cuantos contaminantes están afectando el hábitat. Podemos hacer una parar algunos ejemplos con problemas de medio ambiente como:

- Identificar el número de especies en una región o los contaminantes presentes en los desperdicios que una empresa bota en un vertedero natural.
- Determinar el léxico personal de un autor o las enfermedades causadas por el nivel de toxicidad en el aire de una región de salud.

[1]**E-mail**: bouza@matcom.uh.cu

- Fijar los defectos presentes en las unidades de una línea de productos de una fábrica o las malformaciones en la vegetación circundante a una central átomo eléctrica.
- Establecer los tipos de plaga que están presentes en un cultivo o en las fuentes de los polutantes observados en un río.

El trabajo pionero en esta temática se remonta al trabajo del fundador de la estadística, vea Fisher **et al**. (1943) quien lo desarrolló estudiando un problema de Ecología aplicada. Posteriormente se han estudiado problemas diversos. Son muy conocidos los de Efron-Thisted (1976), quienes estudiaron el léxico de Shakespeare para establecer la autenticidad de ciertas obras, los de Engen (1978), Chao (1981) y Bernard (1982) quienes continuaron la investigación de modelos para hacer luna estimación de Δ en problemas biológicos. Otro enfoque es considerar la existencia de un proceso estocástico que explica la aparición de las clases en la muestra. Esto es más popular al considerar el arribo de especies a un punto de monitoreo, vea Chao-Bunge (2002) y Chao-Lee (1992).

El problema teórico de este trabajo plantea varias interrogantes a la estadística. Muchas de ellas no están resueltas. Una línea es determinar cotas para Δ o estimarlas. Otra es estudiar el cubrimiento obtenido, por la muestra seleccionada, de las clases desconocidas a priori. En este trabajo se abordará este último problema. Los estimadores propuestos por Good (1953) y por Esty (1986) serán estudiados y se proponen alternativas.

Tanto Good (1953) como Esty (1986) desarrollaron Teoremas Centrales del Límite para determinar intervalos confidenciales basados en la aproximación normal. Sus hipótesis son algo fuertes al basarse en que la observación de las clases es descrita por un modelo de equiprobabilidad. En este trabajo proponemos el uso de un modelo superpoblacional del tipo regresión. Su aplicación en este caso conlleva un Modelo Lineal Generalizado como descriptor del comportamiento del cubrimiento. Esto no depende de la hipótesis de equiprobabilidad para garantizar la distribución normal asintótica del predictor.

Para analizar el comportamiento de las diversas alternativas para caracterizar el cubrimiento de la muestra utilizamos datos de dos investigaciones. Una son los datos obtenidos del estudio de la infestación de campos de caña por una plaga en Cuba. El otro es un estudio de biodiversidad forestal desarrollado en la Cuenca del Río San Juan en Guerrero, México.

2. EL USO DE DISEÑOS MUESTRALES

Para modelar este problema hagamos de la población U una partición

$$U = \cup_{i=1}^{\Delta} U_i, U_i \cap U_j = \varnothing, \text{ si } i \neq j.$$

Al seleccionar una muestra s de tamaño n observamos D subpoblaciones. El modelo estocástico considera que hay Δ urnas y estas son ocupadas al azar observándose $D \leq \Delta$ ocupadas. Claramente, D es un estadígrafo suficiente para Δ. El cubrimiento de una muestra es definido por el parámetro

$$\theta = \sum_t I_t \pi_t$$

donde

$$\pi_t = \text{Prob } (U_t \text{ aparezca en la muestra})$$

y

$$I_t = \begin{cases} 1 \text{ si el número de unidades de } U_t \text{ en s es mayor que } 0 \\ 0 \text{ en otro caso} \end{cases}$$

Note que θ es un valor relacionado con la muestra y no con la población, pues es el cubrimiento de s. Este problema aparece también en los estudios de biodiversidad en lo que el índice está condicionado a la muestra observada.

Good (1953) propuso estimar θ mediante

$$\theta_1 = 1 - P_1,\ P_1 = n_1/n,$$

siendo n_1 el número de clases observadas una vez. Esto parece sorprendente a primera vista pero es el índice de cubrimiento usado como referencia por Engen (1978), Starr (1979) y Chao (1981).

Otra solución fue propuesta por Esty (1986) al considerar que el número de clases observadas y la real se relaciona mediante la ecuación

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

Y usar como estimador a

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

La conocida convergencia de ambos estimadores a θ se resume en el siguiente teorema

Teorema 1. (Engen, Starr, Esty). Si $\pi_t = \pi$ para toda clase U_t, $n \to \infty$, $n/\Delta \to \infty$, $n/\Delta \to \alpha$ para $\alpha \in\ ]0,1[$ entonces se cumplen

1. $n^{1/2}(\theta_1 - \theta)$ tiene como distribución aproximada la normal $N\left(0, \frac{\alpha^2(1-\alpha)\ln(1-\alpha)}{\alpha + (1-\alpha)\ln(1-\alpha)}\right)$

2. $n^{1/2}(\theta_2 - \theta)$ tiene como distribución aproximada la normal $N((0(1-\alpha)(\alpha - \ln(1-\alpha))$.

Norris-Metter (1993) estudiaron el caso en que se seleccionan muestras con el fin de estimar el cubrimiento. Los resultados observados en s generan las variables Bernoulli

$$Z_{it} = \begin{cases} 1 \text{ si la clase t - ésima se observa en la selección i} \\ 0 \text{ en otro caso} \end{cases}$$

y podemos calcular

$$n_t = \sum_{i=1}^{n} Z_{it}$$

Esta es una variable Binomial con parámetros n y π_t. Sea N_j el número de clases observadas j veces en la muestra. Analizando N_0, el número de clases no observadas, su esperanza es

$$E(N_0) = \sum_{t=1}^{\Delta} P(n_t = 0) = \sum_{t=1}^{\Delta} (1-\pi_t)^n \geq \sum_{t=1}^{D} (1-\pi_t)^n$$

Norris-Metter (1993) desarrollaron un análisis Bayesiano para derivar cotas inferiores para $E(N_0)$.

Nosotros tomaremos $p_t = n_t/n$, el que estima insesgadamente a π_t, podemos usar el predictor derivado de los resultados anteriores

$$N_0^* = \sum_{t=1}^{D} (1-p_t)^n$$

Es claro que $D + N_0^* \cong \Delta$ por lo que un predictor ingenuo, que subestima el número de clases es:

$$\hat{\Delta} = D + N_0^*$$

Este corrige el sesgo de D y tiene uno menor.

Partiendo de este resultado, si seguimos las ideas del procedimiento de Esty (1986), podemos utilizar

$$\theta_3 = \frac{D}{\hat{\Delta}_0} = \left(1 + \frac{\hat{\Delta}_0}{D}\right)^{-1}$$

para predecir el cubrimiento. Note que este predictor es una variante del estimador de Esty (1986). Sus propiedades son fijadas como sigue:

Proposición 2. Si $n \to \infty$, $(1 - \pi_t)^n \to \varpi_t$, $\varpi > 0$, para al menos un $t \in \{1,..,\Delta\}$, entonces θ_3 tiene una distribución aproximadamente normal con media $\left[\sum_{t=1}^{\Delta} (1-\pi_t)^n\right]^{-1}$ y varianza $V_3 = V(D) + V(\hat{\Delta}_0) - \rho\sqrt{V(D)V(\hat{\Delta}_0)}$ donde ρ es el coeficiente de correlación entre D y $\hat{\Delta}$.

Demostración:

Por las definiciones dadas de D y $\hat{\Delta}_0$ estos están correlacionados positivamente por lo que el sesgo de θ_3 puede considerarse pequeño a fines prácticos. En particular su aproximación a una distribución normal es garantizada por ser este resultado un Teorema Central del Límite para este estimador, ver Som (1996).

3. UN ENFOQUE SUPERPOBLACIONAL

Las hipótesis utilizadas en el Teorema 1 son muy fuertes en la mayor parte de las aplicaciones. En los ejemplos utilizados en la introducción la equiprobabilidad del fenómeno que genera la observación de las distintas clases es poco realista. Ciertos contaminantes y polutantes aparecen con menor frecuencia que otros, algunas plagas son más prolíferas que otras. Algunas dolencias respiratorias son más comunes ante el enrarecimiento del aire por una cierta emisión de gases, etc.

Es poco realista pensar que podremos acudir a los expertos que acepten que la equiprobabilidad es válida, que sean capaces de fijar un valor de α y que nos aseguren la convergencia de los parámetros envueltos para grandes valores de n. Es más sencillo o esperable que los expertos puedan dar ideas sobre la relación existente entre una variable que identifique la clase y la probabilidad de ser detectada. Un modelo superpoblacional suficientemente sencillo es tomar

$$Y_{it} = \begin{cases} 1 \text{ si la observación i pertenece a la clase t} \\ 0 \text{ en otro caso} \end{cases}$$

y valorar si es aceptable que esta se relacione con π_t mediante el modelo usado por Pothoof **et al**. (1992)

$$Y_{it} = \pi_t + \varepsilon_{it}$$

donde $E(Y_{it}) = \pi_t$, y

$$E(\varepsilon_{it}\varepsilon_{i't}) = \begin{cases} \sigma_t^2 \text{ si } Y_{it} = 1 \\ 0 \text{ en otro caso} \end{cases}$$

Ahora tenemos la variable Bernoulli

$$I(i,t|s) = \begin{cases} 1 \text{si } Y_{it} = 1 \\ 0 \text{ en otro caso} \end{cases}$$

y

$$\hat{\pi}_t = \frac{\sum_{i=1}^{n} Y_{it}}{n}$$

entonces tenemos que es modelo-insesgado el predictor

$$\theta_4 = \frac{\sum_{t=1}^{D} \sum_{i=1}^{n} Y_{it}}{n}$$

pues

$$E(\hat{\pi}_t) = \pi_t$$

dada la independencia de las observaciones

$$V(\theta_4) = \sum_{t=1}^{D} \pi_t(1-\pi_t)/n$$

Note que el nuevo estimador es una función lineal de variables Binomiales independientes. Siguiendo las ideas de Bouza (1996) bajo el modelo superpoblacional asumido podemos desarrollar un Teorema Central del Límite. Este es:

Teorema 3. Si $n \to \infty$ entonces $n^{1/2}(\theta_4 - \theta) \to N(0, \sum_{t=1}^{D} \pi_t(1-\pi_t)/n$,

Demostración:

$\hat{\pi}_t$ se distribuye $N\left(\pi_t, \frac{\pi_t(1-\pi_t)}{n}\right)$ aproximadamente y $V(\hat{\pi}_t) - \frac{}{n\to\infty} \to 0$ por lo que la sucesión $\{\hat{\pi}_t\}$ obedece la convergencia a la normal defiende en el Teorema. Ver Friedst-Gray (1997).

Note que este teorema es tan fuerte como el Teorema 1, en términos de convergencia, pero sus hipótesis son mucho más suaves. De hecho si $\pi_t = \pi$ para todo t la eficiencia de este predictor comparado con sus contrapartes del Teorema 1 son:

$$E_{14} = \frac{-\alpha^2(1-\alpha)\ln(1-\alpha)n}{\alpha + (1-\alpha)\ln(1-\alpha)D\pi(1-\pi)}$$

y

$$E_{24} = \frac{(1-\alpha)(\alpha - \ln(1-\alpha))n}{D\pi(1-\pi)}$$

Ambas razones son mayores que uno por lo que el predictor es más exacto.

4. EXPERIMENTOS DE MONTE CARLO

Un estudio de las plagas en la caña de azúcar fue desarrollado utilizando Minería de Datos, Bouza-Schubert (2004). Entomólogos del Centro de Sanidad de Cuba detectaron la existencia de plagas que eran resistentes a los tratamientos usuales. La estructura de las clases determinó el tomar mas de 3 000 muestras sistemáticamente durante más de dos años en todo el país. El número de ellas estaba entre 4 y 10.

Una investigación sobre biodiversidad forestal se desarrolló en el estado de Guerrero, México, Vea Bouza-Covarrubias (2005). Se estudió la población de árboles en la Sierra de Guerrero. Se evaluaron las especies existentes y el número de especimenes. Se hizo una definición de las clases tal que su número se movía entre 5 y 30.

Utilizamos las Bases de Datos de estas investigaciones para construir poblaciones artificiales. Esto permitió determinar los valores de los $\pi_t's$ y calcular θ. Se seleccionaron muestras de estas poblaciones artificiales con fracciones del 5%, 10% y 20%. Las estimaciones o predicciones fueron computadas a partir de los experimentos de Monte Carlo desarrollados. En cada muestra se computaron los $\theta_j's$, i = 1,..,4. Para evaluar el comportamiento de los diversos estimadores y predictores. Se generaron 100 muestras y se evaluaron las medidas

$$B_{i(e)} = \sum_{s=1}^{100} \left| \theta - \theta_i \right|_s / \sum_{s=1}^{100} \left| \theta - \theta_{(e)} \right|_s \text{ , } i \neq e, \; i, e = 1,..,4.$$

Los resultados para el estudio de las plagas se brindan en la Tabla 1.

Tabla 1. Desviación relativa en % ($100B_{i(e)}$). Infestación de la caña de azúcar.

Fracción de muestreo	$B_{1(2)}$	$B_{1(3)}$	$B_{1(4)}$	$B_{2(3)}$	$B_{2(4)}$	$B_{3(4)}$
0,05	88,6	99.6	121,4	95,7	134,1	142,3
0,10	79.3	94.1	114,8	91,2	123,4	135,4
0,20	65,8	89,4	108,3	87,6	111,6	127,5

Los resultados para el estudio de la biodiversidad forestal se brindan en la Tabla 2.

Tabla 2. Desviación relativa en % ($100B_{i(e)}$). Biodiversidad forestal.

Fracción de muestreo	$B_{1(2)}$	$B_{1(3)}$	$B_{1(4)}$	$B_{2(3)}$	$B_{2(4)}$	$B_{3(4)}$
0,05	45,7	101,1	127,9	143,5	182,6	112,4
0,10	35,8	100,5	119,2	138,7	175,2	109,4
0,20	31,9	100,8	120.6	135,3	175,1	110,0

Como se ve el predictor basado en el modelo superpoblacional posee un comportamiento mucho mejor que los estimadores. Entre los estimadores el mejor es θ_1.

RECONOCIMIENTOS

Este trabajo se llevó a cabo durante una visita del autor a la Universidad Autónoma de Guerrero dentro de un proyecto de colaboración entre esta institución y la Universidad de La Habana. Esta versión ha sido mejorada gracias a la incorporación de los criterios vertido por dos referees anónimos

REFERENCIAS

BOUZA, C. (1996): "Linear rank tests derived from a superpopulation model". **Biometrical J**, 38, 497-506.

BOUZA C. y D. COVARRUBIAS (2005): "Estimación del índice de diversidad de Simpson en m sitios de muestreo". **Inv. Operacional**, 26, 187-195.

BOUZA, C. and L. SCHUBERT (2004): The estimation of the biodiversity and the characterization of the dynamics: an application to the study of a pest (2003). **Revista de Matemática e Estatistica,** 21, 85-98.

BERNARD, B. (1982): "On the relation between the type-token and species problem". **J. Applied Prob**. 19, 785-793.

CHAO, A. (1981): "On estimating the probability of discovering a new species". **Ann. Statistics**, 9, 1339-1342.

CHAO, A. and T. BUNGE (2002): "Estimating the number of species in a stochastic abundance model". Biometrics. 58, 530-539.

CHAO, A. and S.M. LEE (1992): "Estimating the number of classes in sample coverage". **J. American Stat**. **Ass**. 87, 210-217.

EFRON, B. and R. THISTED (1976): “Estimating the number of unseen species: how many words did Shakespeare know?”. **Biometrika**, 63, 435-447.

ENGEN, S. (1996): **Stochastic Abundance Models**. Halted Press, New York.

ESTY, W.W. (1986): “The efficiency of Good’s nonparametric coverage estimator”. **Ann. of Statistics**. 14, 1-9.

FISHER, R.A.; A.S. CORBETT and S.C.B. WILLIAMS (1943): “The relation between the number of species and the number of individuals in a random sample of animal population”. **J. of Animal** E. 12, 42-58.

FRIEDST, B. and L. GRAY (1997): **A Modern Approach to Probability Theory**. Birkhäuser, Boston.

GOOD, I. (1953): “The population frequency of species and the estimation of the population parameter”. **Biometrika**, 43, 45-63.

POTHOFF, R.R.; M.A. WOODBURY and K.I.G. MANTON (1992): “Equivalent sample size and equivalent degrees of freedom for refinements weights under superpopulation models”. **J. Amer. Stat. Ass**. 87, 383-396.

SOM, R.K. (1996): **Practical Sampling Techniques**. Marcel Dekker, New York.

STARR, N. (1979): “Linear estimation of the probability of discovering a new species”. **Ann. of Statistics**. 7, 644-652.