MUESTREO ADAPTATIVO BAJO CONTACTOS ALEATORIOS

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

RESUMEN

Se analizan estrategias bajo la existencia de conexiones entre las unidades que generan redes. Se modela el caso en el que las conexiones son aleatorias. El tamaño de la muestra varía en función de los contactos lo que establece la existencia de una etapa adicional de aleatorización. Se proponen estimadores y se obtienen sus errores. Este problema es estudiado para el muestreo simple aleatorio y el estratificado. Se propone como solución para computar los errores el uso de métodos de Bootstrap junto con algoritmos para computarlos.

Palabras clave: muestreo adaptativo, Bootstrap, redes.

ABSTRACT

Strategies are studied under the hyphotesis of random connections among the units. We model the case in which these conections are random. They fix networks and the sample size varies as a function of the contatcs. Hence an additional randomization stage is present. Estimators and their errors are developed. This problem is studied for simple random sampling and stratification. As a solution for computing the errors Bootstrap methods are proposed and computing algorithms are provided.

Key Words: adaptative sampling, Bootstrap, networks.

MSC: 62-DO5.

1. INTRODUCCION

Thompson [1990] y [1991] propuso modelos que incluyen el uso de relaciones entre las unidades muestreadas y aquellas con las que se relaciona. Bouza [1982] analizó modelos para el muestreo en poblaciones conexas. Las conexiones permiten derivar la estructura de redes en el sentido usual de la Teoría de Grafos. La estructura de la población $U = \{u_1,\ldots,u_N\}$ permite derivar estrategias llamadas de "muestreo adaptativo", ver Thompson-Seber [1996]. Desde el punto de vista práctico esto permite incrementar el número de unidades básicas con la característica de interés. Tal es el caso cuando se desea estudiar el comportamiento de una población de insectos, la preeminencia de una cierta enfermedad infecciosa o las características de ciertos prototipos en la producción de una cadena.

La poca abundancia de ciertos tipos de elementos con la propiedad deseada en la muestra, puede sugerir incrementar el número de ellas incluyendo otras, que se espera posean la característica, a partir de las observaciones en la muestra inicial. Ejemplos de este tipo de problema son:

- El agrónomo observa sitios de muestreo donde hay pocos insectos del tipo de interés. Sin embargo observa alguna evidencia de su existencia en sitios contiguos.
- El médico visita familias de una manzana y observa pocos portadores de un virus, pero obtiene información sobre la existencia de enfermos en las inmediaciones.
- El veterinario evalúa ejemplares de una cierta raza poco extendida y conoce que hay otras recién introducidas en fincas vecinas.
- El sociólogo entrevista a madres jóvenes solteras y selecciona familias de una barriada en la que le comunican que hay casas con estos casos en otras casas.

[1] Email:bouza@matcom.uh.cu

- Un ingeniero realiza el control de la calidad y observa que ciertos productos salen con defectos raros. Haciendo simples pruebas detecta si este problema afecta a otros lotes producidos por la línea que desea evaluar.

En todos los casos la selección de la muestra se lleva a cabo sin conocer qué unidades poseen la propiedad de interés. Las unidades son seleccionadas aleatoriamente. En general se utiliza muestreo simple aleatorio [msa] y se obtiene información sobre la abundancia de unidades con las características deseadas solo al hacer las visitas. Al observar pocas que satisfacen su interés el especialista desea ampliar la muestra inicial s incluyendo algunas unidades 'contiguas' donde espera existan otras unidades con la propiedad de interés. En ocasiones la similitud es detectable y no varía, por lo que podemos considerar que está determinada por 'conexiones' determinísticas. La muestra ampliada es aleatoria.

Consideremos que la existencia de estas conecciones [arcos entre los nodos representados por las unidades de U] es generada por algún mecanismo aleatorio. Tal es el caso de cuando los insectos se trasladan de un sitio a otro, los portadores pasan a enfermos, la manada de animales cambia de composición, hay mucha movilidad de los vecinos del barrio o si los defectos de los productos son afectados por la eficiencia de los operarios y ellos cambian de línea de producción. Este problema es modelado en este trabajo utilizando el marco propuesto por Thompson **op. cit** .

Es usual que el cliente desee obtener un modelo que le permita 'adaptarse' a la situación encontrada e incrementar las unidades con la característica de interés. Por ello es común que al observar la muestra trate de incrementar el número de observaciones en forma arbitraria. El estadístico usando un modelo adaptativo puede fijar criterios para incluir nuevas observaciones a partir de las características observadas en las unidades en la muestra inicial. Sin pérdida en generalidad podemos caracterizar el problema al fijar que se incluyan en la muestra s todas aquellas unidades para las cuales la variable de interés $Y \geq \alpha$. Entonces para cada u_i se determina el conjunto $\{u_j \mid Y_j \geq \alpha\} = C[i1]$. Si $u_j \in C[i1]$ y $Y_j \geq \alpha$ se repite el proceso hasta que ninguno de los nuevos vecinos genera la inclusión de nuevos elementos. La unión de todos ellos determina el 'conglomerado adaptativo' C[i].

La relación de vecindad utilizada debe ser simétrica y determina que cualquier subcolección $C'[i], \subseteq C[j]$ de C'[i] tal que si $u_j \in C'[i]$ entonces $C'[i], \subseteq C[j]$. Diremos que C'[i] es una red. Si $u_t \in C[i]$ y no existe C'[i] a la que pertenezca se dice que es una 'unidad frontera' [uf]. Utilizando este marco se desarrollan los modelos del muestreo adaptativo [ma], ver Thompson-Seber [1996] para más detalle.

Bouza [1982] por su parte estudió el problema de las conecciones no determinísticas en el marco de la existencia de una estructura poblacional que permite el diseño de un grafo. En este trabajo se desarrolla el estudio del msa cuando las conexiones en la red son modeladas por variables aleatorias con distribución Bernoulli. Estimadores de la media poblacional son obtenidos y sus errores son deducidos. Se propone realizar la estimación de éstos utilizando métodos de Bootstrap. Se caracteriza el problema de su convergencia utilizando condiciones de robustez del método de [BS], ver Jureckova-Sen [1996]. Se obtiene que los criterios de Liu-Singh [1987] son satisfechos y lo que sustenta que los algoritmos propuestos son convergentes. Queda pendiente un estudio computacional adecuado para determinar qué alternativas son las más recomendables para el caso del msa y del msa estratificado [msae].

2. CASO SIMPLE

Siguiendo los principios enunciados en la introducción la selección de una unidad u_i puede ser generada por:

1. Se observa una de las R_i redes en las que u_i es incluido.

2. Se observa al menos uno de los K_i conglomerados en los que u_i es una unidad frontera.

Supondremos que el tamaño de s es suficientemente grande como para que la distinción entre el uso o no del reemplazo sea de poco interés. Al utilizar msa para seleccionar la muestra, la probabilidad de observar u_i

es $P_i = [R_i + K_i]/N$. Note que al observar la muestra aparecen A_i redes conteniendo a u_i y que su esperanza es $E[A_i] = nR_i/N$. El enfoque tradicional utiliza, ver Cochran [1981]

$$m = \sum_{i=1}^{n} Y_i / n$$

para estimar μ, la media poblacional. Su error es

$$V[m] = \sigma^2/n = \sum_{i=1}^{n} [Y_i / \mu]^2 / Nn$$

si $n/N \cong 0$. Thompson [1990] propuso utilizar

$$m_T = \sum_{i=1}^{n} \sum_{j \in C'[i]}^{n} Y_i / nR_i = \sum_{i=1}^{n} Y_i' / n = \sum_{i=1}^{n} A_i Y_i / nR_i \qquad (2)$$

Este estimador es insesgado y su varianza es

$$V[m_T] = \sum_{i=1}^{N} \left[Y_i^{'} - \mu\right]^2 / Nn = \sigma_T^2 / n = \left[\sum_{i=1}^{N} (Y_{hi}^{'})^2 / R_i^2 - N\mu^2\right] / Nn \qquad (3)$$

pues m_T es la media de las $Y_i^{'}$.

Cuando los contactos son aleatorios tenemos que la conexión entre dos unidades es modelada por una distribución a priori. Esta es caracterizada por variables Bernoulli que establecen la existencia o no de arcos. Las denotaremos por $B_{ij} = 1$ (0) si $j \in C'[i]$ ($j \notin C'[i]$) donde $E[B_{ij}] = Q_{ij}$ es su valor esperado. Entonces

$$E[R_{ij}] = \sum_{i=1}^{N} Q_{ij} = Q_i$$

La simetría en la relación garantiza que $Q_{ij} = Q_{ji}$ por lo que no hay independencia entre estas variables. (2) continúa siendo insesgado pero la estructura de (3) determina que en estos casos R_i es una variable aleatoria. Cuando $R_i > 0$ y $E[R_{ij}^2] < \infty$, Bouza [1991] obtuvo que $Q'_{ij} = 1 - Q_{ij}$ y $Q_{ijj'} = Prob[B_{ij} = B_{ij'} = 1]$

$$E[R_i^{-2}] = \left[E(R_i^2)\right]^{-1} + \sum_{i=1}^{\infty} [-1]^t = \left[E(R_i^2 - E[R_i^2])^t / E(R_i^2)^{t+1} + 0[n^{-1}]\right]$$

Esto permite obtener una aproximación para la esperanza de $V[m_T]$ bajo el modelo superpoblacional que genera las conexiones. Esta es

$$EV[m_T] \cong \left[\sum_{i=1}^{N} \left[\left(\sum_{j \in C'[i]} Y_j\right)^2 / NQ_i'\right] - \mu^2\right] / n = V_T^* \qquad (4)$$

donde el desarrollo de la esperanza de $1/R_i^2$ en Series de Taylor se apoya en las propiedades de las variables Bernoulli y su distribución conjunta. Utilizando las propiedades de los momentos de esta distribución se obtiene, tras un trabajo algebraico arduo que

$$Q_i' = [H_i(1 - H_i) + J_i]/H_i \qquad (5)$$

donde

$$H_i = Q_i^2 + \sum_{j=1}^{N} Q_{ij} Q_{ij}' + \sum_{j \neq j'} \left[Q_{ijj'} - Q_{ij} Q_{ij'}\right] \qquad (6)$$

$$J_i = \sum_{i=1}^{N} Q_{ij}^4 + 2\sum_{j\neq j'} Q_{ij}^2 Q_{ij'}^2 + \sum_{j\neq j'\neq t} Q_{ij}^2 Q_{ij'} Q_{it} + 2\sum_{j\neq j'\neq t'} Q_{ij} Q_{it'}^2 Q_{it'} + \sum_{j\neq j'\neq t\neq t'} Q_{ij} Q_{ij'} Q_{it} Q_{it'} \quad (7)$$

Utilizando las propiedades de [3] la estructura de un estimador adecuado parece ser

$$S_T^2 = \sum_{h=1}^{n} \left[Y_h' - m_T \right]^2 / [n-1]$$

Note que podemos aceptar la validez de $E\left[S_T^2 / n\right] = V_T^* + o(n^{-1})$. Esto permite utilizar los resultados de Liu-Singh [1987] los que establecen que bajo esta hipótesis, se puede desarrollar S_T^2 / n en Series de Edgeworth [SE] garantizándose que el método de BS converge al funcional que estima. La complejidad del valor esperado sugiere que esta es la mejor opción para obtener una aproximación de [4] pues en general los parámetros del modelo superpoblacional descrito a través de [5], [6] y [7] son desconocidos.

Utilizando el principio de la repetición un algoritmo para obtener una estimación del error esperado es el siguiente:

Algoritmo 1. Bootstrap para la estimación del error esperado de m_T basado en la repetición.

Paso 1. Fijar M

Paso 2. Dada la muestra identificar C'[1],...,C'[n] y calcular $Y'_1,...,Y'_n$.

Paso 3. Seleccionar una muestra de tamaño n mediante msa con reemplazo de la muestra inicial y calcular

$$m_{TB} = \sum_{i=1}^{n} Y_{iB}' / n$$

para la muestra s_B.

Paso 4. Si B = M calcular

$$m_{TB}^* = \sum_{B=1}^{M} m_{TB} / M$$

$$S_{TB}^2[A1] = \sum_{B=1}^{M} \left[m_{TB} - m_{TB}^* \right]^2 / [M-1]$$

Paso 5. **FIN**

Por su parte el principio de la sustitución hace recomendable estimar los parámetros y sustituirles por sus estimaciones en el cálculo del error. Este es el estimador de BS del error. Un algoritmo que implementa la aproximación en este sentido es el siguiente:

Algoritmo 2. Boostrap para la estimación del error esperado de m_T basado en el principio de la sustitución.

Paso 1. Fijar M

Paso 2. Dada la muestra identificar C'[1],...,C'[n] y calcular $Y'_1,\cdots,Y'_n$.

Paso 3. Seleccionar una muestra de tamaño n mediante msa con reemplazo de la muestra inicial y evaluar en cada muestra las variables B_{ij}, y $B_{ijj'} = B_{ij}B_{ij'}$.

Paso 4. Estimar los parámetros $Q'_1,\cdots,Q'_n$ utilizando las M muestras de BS calculando (5), (6) y (7).

Paso 5. Utilizar las estimaciones obtenidas en vez de los parámetros en la fórmula dada en (4) y calcular S^2_{TB} [A2]

Paso 6. **FIN**

Este último algoritmo es más complejo que el primero pero su velocidad de convergencia generalmente es mayor dadas las relaciones de la regla de sustitución [plug-in-rule] con el método Delta y el de este con el BS, vea Jurĕckova-Sen [1996] para una detallada discusión.

3. CASO ESTRATIFICADO

El uso de la estructura del estimador del mase determina que si para el estrato U_h definimos C'[hi] como la red correspondiente a u_i, m_{Th} como el estimador de la media poblacional μ_h de U_h, $W_h = {}_hN/N$, siendo N_h su tamaño y n_h el de la muestra de éste.

Entonces si denotamos por N[hi] = |C'[hi]| al número de unidades de C'[hi] y

$$m_{e1} = \sum_{h=1}^{H} W_h m_{Th} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \sum_{j\in C'[hi]} Y_j / n_h R_{hi}$$

es un estimador insesgado de μ. La varianza de este está dada por

$$V[m_{e1}] = \sum_{h=1}^{H} W_h^2 = \sum_{i=1}^{N_h} [\, Y'_{hi} - \mu_h \,]^2 / N_h n_h = \sum_{h=1}^{H} W_h^2 \sigma^2_{Th} / n_h =$$

$$= \sum_{h=1}^{H} W_h^2 \left[\sum_{i=1}^{N_h} (Y'_{hi})^2 / R^2_{hi} - N_h \mu_h^2 \right] / N_h n_h$$

Bajo la existencia de conexiones aleatorias entre los elementos del estrato el error es:

$$EV[m_{e1}] = V^*_{e1} = \sum_{h=1}^{H} W_h^2 \sum_{i=1}^{N_h} \left[\left(\sum_{j\in C'[hi]} Y_{hj} \right)^2 / N_h Q'_{hi} \right] - \mu_h^2 / \Pi_h$$

Nuevamente la estimación del error puede ser llevada a cabo mediante el uso de BS haciendo uso de uno de los algoritmos propuestos en la sección anterior.

Algoritmo 3. Bootstrap para estimar el error esperado de m_{e1}.

Paso 1. Seleccionar p = 1,2.

Paso 2. Para cada h aplicar el algoritmo p.

Paso 3. Calcular $V_B[p] = \sum_{h=1}^{H} W_h^2 S_{TB}^2[Ap]$.

Paso 4. **FIN**

Thompson [1991] consideró la posibilidad de que existiese conexión entre estratos.

Tomando R_{khi} como el número de elementos de U_k que son incluidos en la red C'[hi] definida por la selección de $u_i \in U_h$ D_{khi} el número de unidades muestreadas en $U_k \cap C'[hi] = U'_{khi}$. Tomemos

T_{khi} como el total de Y en U_{khi} y

$$T_{hi} = f_h \sum_{k=1}^{H} T_{khi} / \sum_{t=1}^{H} f_t R_{thi}$$

$$T_h = \sum_{i=1}^{N_k} T_{hi} / N_h$$

donde f = n/N. Como

$$E[D_{khi}] = \sum_{t=1}^{H} f_t R_{thi}$$

el estimador

$$m_{e2} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} T_{hi} / n_h = \sum_{h=1}^{H} W_h T'_h$$

es insesgado. La medida de la precisión analizada por Thompson [1991] fue

$$V[m_{e2}] = \sum_{h=1}^{H} W_h^2 \sum_{i=1}^{N_h} [T_{hi} - T_h]^2 / N_h n_h = \sum_{h=1}^{H} W_h^2 V'^2_{T'h} / n_h$$

Cuando las conexiones son aleatorias las esperanzas correspondientes requieren del cálculo de aproximaciones de esperanzas. Tomando

$$r_{tt'hi} = \sum_{j \neq j'} Q_{tt'hijj'} + Q^2_{tt'hij} + Q_{tt'hij} Q'_{tt'hij} \tag{8}$$

$$r_{thi} = \left[\sum_{j=1}^{N_h} Q_{thij} \right]^2 + \sum_{j=1}^{N_h} Q_{thij} Q'_{thij} + \sum_{j \neq j'} [Q_{thijj'} - Q_{thij} Q_{thij'}] \tag{9}$$

$$r_{hi} = \sum_{h=1}^{H} f_t r_{thi} + \sum_{t \neq t'} f_t f_{t'} t_{tt'hi} \tag{10}$$

$$Q_{thijj'} = \text{Prob}\{j \wedge j' \in C'[hi] \cap U_{t'}\} \tag{11}$$

$$Q_{tt'hij} = \text{Prob}\{(i \wedge j \in C'[hi] \cap U_t) \wedge (i \wedge j \in C'[hi] \cap U_{t'})\} \tag{12}$$

$$Q_{tt'hijj'} = \text{Prob}\{(i \wedge j \wedge j' \in C'[ht]) \wedge (i \wedge j \wedge j' \in C'[ht'])\} \tag{13}$$

$$Q_{thij} = \text{Prob}\{j \in C'[hi] \cap U_t\} \tag{14}$$

$$Q'_{thij} = 1 - Q_{thij} \tag{15}$$

$$Z_{kh[1]} = \sum_{j=1}^{N_h} Q_{khij}^4 + 2\sum_{j\neq j'}\left[Q_{khi}^2 Q_{khij'} + \sum_{j\neq j'\neq t}\left[Q_{khij}^3 Q_{khij'} Q_{khit}\right] + \sum_{j\neq j'\neq t'} \right][Q_{khij'} Q_{khij'} Q_{khijt}] +$$

$$+ \sum_{j\neq j'\neq t\neq t'}\left[Q_{khij} Q_{khij'} Q_{khit'} Q_{khit}\right] \tag{16}$$

$$Z_{khi} = \sum_{j=1}^{N_h} Q_{khij} \tag{17}$$

$$^{hi}A_{kk'} = \sum_{j=1}^{N_h} Q_{khij} Q_{k'hij}[Q_{khij} - Q'_{khi}] + \left[\sum_{j=1}^{N_h} Q_{khij} Q'_{khij}\right]\left[\sum_{j=1}^{N_h} Q_{k'hij} Q_{k'hij}\right] \tag{18}$$

$$^{hi}B_{kk'} = \sum_{j=1}^{N_h} Q_{khij}^2 Q'_{khij} Q_{k'hij} \tag{19}$$

$$^{hi}S_{kk'} = \sum_{j=1}^{N_h} Q_{khij} Q_{k'hij}[Q'_{khij} - 2Q_{khij}] \tag{20}$$

podemos fijar a

$$\theta_{hi} = \sum_{h=1}^{H} f_k^4 \theta_{khij[1]} + \sum_{k\neq k'} f_k^2 f_{k'}^{2hi} A_{kk'} + \sum_{k\neq k'\neq t\neq t'} f_k f_{k'} f_t f_{t'} Z_{khi} Z_{k'hi} Z_{thi} Z_{t'hi}$$

$$- 6\sum_{k\neq k'} f_k f_{k'}^{hi} B_{kk'} + \sum_{k\neq k'\neq t} f_k f_{k'} f_t^{hi} S_{kk'} + \left[r_{hi}(1 - r_{hi}) f_{hi}^{-2}\right] \tag{21}$$

Entonces

$$EV[m_{e2}] = V_{e2}^* \cong \left[\theta_{hi} N_h n_h\right]^{-1}\left[\sum_{h=1}^{H} W_h^2 \sum_{i=1}^{N_h}\left(f_h \sum_{k=1}^{H} T_{khi} - \sum_{i=1}^{H} f_r \sum_{i=1}^{N_h} T_{rgi} / N_r\right)^2\right] \tag{22}$$

El resto de los términos son de orden n. De ahí que un estimador insesgado de la varianza dentro del estrato h-ésimo sea

$$s_h^2 = \sum_{i=1}^{n_h} \left[T_{hi} - T'_h\right]^2 / [n_h - 1]$$

por lo que

$$V_2^* = \sum_{h=1}^{H} W_h^2 S_h^2 / n_h$$

es desarrollable en SE y por lo que el uso de BS está respaldado por la existencia de una buena aproximación al verdadero error. El algoritmo siguiente permite implementar este método.

Algoritmo 4. Boostrap para la estimación del error esperado de m_{e2} basado en la repetición.

Paso 1. Fijar M.

Paso 2. Dada la muestra identificar C'[i1],…,C'[1n_1],…,C'[H1],…,C'[Hn_H] y calcular los totales T_{khi} y T_{hi} para todo k,h,i.

Paso 3. Seleccionar mediante m sacr muestras de BS de tamaño n_h en cada estrato y calcular para cada muestra de BS_{Sb}

$$m_{e2B} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} T_{hib} / n_h$$

Paso 4. Hallar

$$m_{e2B} = m_{e2} = \sum_{b=1}^{B} m_{e2b} / B$$

y

$$V_{2B}^{*} = \sum_{b=1}^{B} [m_{e2b} - m_{e2B}]^2 / B$$

Paso 5. **FIN**

El uso de reglas de sustitución sugiere el algoritmo siguiente:

Algoritmo 5. Método de Boostrap para la estimación del error esperado de m_{e2} usando la regla de sustitución.

Paso 1. Proceder con los pasos 1 y 2 del Algoritmo 4.

Paso 2. Seleccionar mediante m sacr muestras de BS de tamaño n_h en cada estrato y evaluar las conexiones observadas

Paso 3. Calcular las frecuencias relativas y efectuar las estimaciones de los parámetros (8) - (21).

Paso 4. Calcular (22) sustituyendo las estimaciones de los correspondientes θ_{hi}.

Note que esos algoritmos garantizan la convergencia a los parámetros estimados pero no la normalidad asintótica de los estimados.

AGRADECIMIENTOS

Este trabajo fue elaborado parcialmente durante una visita del autor a la Universidad Veracruzana amparada por un proyecto FOMES.

Agradezco las sugerencias de los refeerees, lo que ha permitido presentar una versión sensiblemente mejorada.

REFERENCIAS

BOUZA, C. [1982]: "El uso de paneles en estratos conexos", **Investigación Operacional**, 3, 109-116.

COCHRAN, W.G. [1981]: **Técnicas de muestreo**, Editorial CECSA, México.

JURĚCKOVA, J. y P.K. SEN [1996]: "Robust Statistical Procedures: Asymptotics and Interrelations", Wiley, New York.

LIU, R.Y. and K. SINGH (1990): "On partial correction by the boostrap", **Ann. Stat.** 1713-1718.

THOMPSON, S.K. [1990]: "Adaptative cluster sampling", **J. Amer. Stat. Ass.** 67, 224-227.

_____ [1991]: "Stratified cluster sampling", **Biometrika**, 78, 389-397.

_____ and G.A.F. SEBER [1996]: "Adaptative Sampling", Wiley, New York.