Revista Colombiana de Estadística

Volumen 24 (2001) N^{o} 2, páginas 121 *a* 140

MUESTREO DE CONGLOMERADOS CON MULTIPLICIDAD: ESTIMACIÓN DEL TOTAL EN POBLACIONES RARAS

PEDRO ALEJO TORRES SAAVEDRA[*]
DAVID OSPINA BOTERO[**]

Resumen

El muestreo de conglomerados con multiplicidad utiliza el concepto de regla de conteo de multiplicidad para incluir en la encuesta individuos no pertenecientes a la muestra. Se derivan las fórmulas para la ganancia de eficiencia y para los componentes de varianza de dos estimadores de multiplicidad del total poblacional. Mediante simulación se ilustra la magnitud de la ganancia de eficiencia con el muestreo con multiplicidad en función del tamaño de muestra y la amplitud de la regla de conteo.
***Palabras claves**: Población escasa, Muestreo de conglomerados, Multiplicidad, Eficiencia, Encuesta de hogares, Muestreo de redes, Estimador de multiplicidad.*

Abstract

Cluster sampling with multiplicity uses the concept of counting rule of multiplicity in order to include in the survey those individuals not belonging to the sample. Additionally, some results to improve the efficiency and for the variance components from multiplicity estimators of the population total

[*]Estadístico, Universidad Nacional de Colombia, E-mail:pa_torress@yahoo.es.
[**]Profesor Asociado, Departamento de Estadística, Universidad Nacional de Colombia; E-mail: dospina@matematicas.unal.edu.co.

are derived. Finally, a simulation exercise to illustrate the magnitude of the gain in the efficiency with the sampling with multiplicity in function of the sample size and the extent of the counting rule is presented.
***Key words**: Scarce population, Cluster sampling, Multiplicity, Efficiency, Household survey, Network sampling, Multiplicity estimator.*

1. Introducción

Una población rara, según Graham y Dallas (1986), se define como un subconjunto *pequeño* de la población total. Algunos autores, como Czaja, et. al. (1992) consideran una población rara aquella presente en menos del 3% del universo de estudio. Las poblaciones denominadas raras, escasas o evasivas, tratadas en Graham y Dallas (1986) y Sudman et al. (1988), tienen como casos típicos personas con enfermedades con bajas tasas de prevalencia (cáncer, epilepsia, entre otras), adicciones (alcohol, droga, etc.), víctimas de crímenes (violación, robo, secuestro, etc.), y en general, todos aquellos grupos con baja frecuencia de unidades que poseen una característica determinada.

En la estimación de parámetros asociados con características escasas, los diseños muestrales tradicionales no ofrecen las mejores condiciones metodológicas, principalmente por la dificultad de localizar los elementos con la característica deseada. Es por ello que debe recurrirse a otras técnicas especiales, tales como: el muestreo de multiplicidad o redes, el muestreo de marcos múltiples, el muestreo de captura-recaptura, el muestreo secuencial, las muestras geográficamente agrupadas y el muestreo inverso o binomial negativo.

Las técnicas convencionales de muestreo en encuestas de hogares relacionan cada individuo a un único hogar. De acuerdo con Sirken (1998), con el desarrollo en la década de los 70's del muestreo de redes o de multiplicidad y el concepto de regla compuesta o de conteo de multiplicidad, los investigadores empezaron a utilizar en los diseños muestrales una serie de reglas que permitían a los miembros de una unidad de enumeración informar acerca de otros individuos no pertenecientes a dicha unidad.

El muestreo de multiplicidad o de redes reduce en gran parte el número de contactos necesarios para detectar los miembros de una población rara, principal problema en estos casos. Para el muestreo de conglomerados, se inicia con una muestra aleatoria simple de conglomerados y posteriormente se ubican los individuos pertenecientes a esta muestra. Con la metodología de multiplicidad se pregunta a cada elemento de la muestra sobre la característica de estudio y, a su vez, se le indaga acerca de otros individuos relacionados con él bajo un criterio específico, denominado *regla de conteo de multiplicidad.*

El concepto de regla de conteo de multiplicidad se maneja como un sistema

que define y establece relaciones entre individuos de la población residentes en diferentes hogares. La regla de conteo distribuye los individuos de la población entre los hogares de tal manera que todo individuo se relacione por lo menos con un hogar y además, que varios hogares se relacionen con más de un individuo. Los grupos de individuos relacionados a un hogar determinado se denominan *conglomerados* y los conjuntos de hogares que relacionan un individuo particular se llaman *redes*.

En general, las reglas de conteo de multiplicidad se basan en relaciones consanguíneas, sociales o espaciales entre individuos. Sirken (1970) define la *multiplicidad* de un individuo como el número total de hogares que están relacionados a él mediante una regla de conteo específica. Sirken (1970 y 1975) y Czaja et al. (1986), reportan en sus estudios reglas de conteo de hermanos, hijos, amigos cercanos, vecinos o familiares, entre otras. En las encuestas convencionales todos los individuos tienen multiplicidad igual a 1 (las reglas De Jure y De Facto son ejemplos de reglas de conteo convencionales).

Adicional a los supuestos necesarios para llevar a cabo un diseño muestral tradicional, la aplicación del muestreo de redes debe considerar dos aspectos importantes: la percepción social de la característica a ser estudiada y el nivel de conocimiento entre los miembros de la red. En el primer caso, si se estudia una incapacidad física esta tiene mayores posibilidades y mejores condiciones de ser reportada que una característica evasiva, tal como el alcoholismo, la drogadicción o alguna otra que discrimine socialmente. En la segunda situación, se debe buscar una regla de conteo que garantice al máximo un conocimiento entre los miembros de la red. Las reglas de conteo de multiplicidad amplias reducen los errores muestrales pero son propicias para aumentar los sesgos y errores no muestrales en las estimaciones.

La teoría del muestreo de multiplicidad o redes se remonta a comienzos de los años 60's. Esta técnica tuvo su origen en la necesidad de resolver los problemas de reportes múltiples con pacientes que habían recibido tratamiento de fibrosis quística, enfermedad genética, en estudios con marcos muestrales conformados por centros médicos.

A mediados de los 60's, Sirken (1970) publicó un trabajo en la serie *Vital and Health Statistics del National Center for Health Statistics* (NCHS), en el cual se desarrollaron tres estimadores insesgados de multiplicidad para estimar la prevalencia de enfermedades raras mediante encuestas muestrales. La diferencia entre estos estimadores radica en el tipo de información requerida acerca del tamaño de las redes de la población. Nathan (1976) evaluó diferentes reglas de conteo y procedimientos de ponderación en este tipo de encuestas. Con base en un estudio de matrimonios y nacimientos en Israel, se mostró que el número de hogares relacionados a cada individuo, mediante ciertas reglas de conteo de multiplicidad basadas en encuestas de hogares, se distribuye aproximadamente como una variable aleatoria Poisson con parámetro igual al número promedio de hogares rela-

cionados a los individuos. En las dos reglas adoptadas, los valores del estadístico de la prueba de bondad de ajuste ji-cuadrado mostraron resultados significativos. Sirken (1970) dió a conocer la teoría de redes o de multiplicidad a las encuestas de hogares mediante muestreo aleatorio simple. El resultado de este artículo establece una comparación teórica del muestreo tradicional con el de multiplicidad mediante el uso de variables indicadoras y modelos basados en agrupamientos de hogares. Dos años más tarde, Sirken (1972a) muestra los componentes de varianza del estimador de multiplicidad, con lo cual deja ver el aporte de las reglas de conteo de multiplicidad y los errores muestrales en la varianza del estimador del total poblacional. Este mismo año sirvió para la divulgación de un documento donde se presenta el uso del muestreo de multiplicidad o redes al muestreo estratificado aleatorio (Sirken, 1972b). Posteriormente, Levy (1977b) halló una afijación óptima del tamaño de muestra para muestreo estratificado aleatorio, basada en los costos de la encuesta y los componentes de varianza del estimador.

Dentro de los usos del muestreo de multiplicidad o de redes, se encuentra el control de calidad de los reportes estadísticos publicados en *Vital and Healths* del NCHS por Levy y Sirken (1972). Consecuente con este trabajo, Sirken y Levy (1974) propusieron un estimador de multiplicidad basado en razones de variables aleatorias. Más tarde, Levy (1977a) presentó el estimador de multiplicidad para el caso de un muestreo bietápico en muestreo por conglomerados.

El concepto de redes fue aprovechado por Granovetter (1977) para establecer un método de muestreo en la estimación del número promedio de personas que conocen determinado individuo. Trabajos similares como "Snowball Sampling" en muestreo de redes sociales de Goodman (1961) y "Sampling Personal Network Structures" trabajado por Spreen (1999), entre otros, expanden la teoría de redes a casos generales y aplicaciones de teoría de grafos a este contexto, citados en Bouza (1999).

Tratando de resolver el problema de sobre-cobertura en marcos muestrales duales o incompletos, Sirken (1979) construyó un estimador de redes basado en la combinación de la información de los dos marcos mediante reglas de conteo disyuntas, estimador trabajado también por Casady, et al. (1985). Es precisamente Sirken (1983) quien presentó el muestreo de redes como una herramienta para el manejo de datos faltantes debido al uso de marcos muestrales incompletos. En la búsqueda continua de otras aplicaciones, Shimizu y Sirken (1998), utilizan las encuestas con establecimientos (Population Based Establishment Surveys - PBES). Estas son encuestas de negocios donde la muestra se selecciona a través de las transacciones que los establecimientos han tenido con determinados hogares. El estimador usado en estas situaciones es el de Horvitz-Thompson propuesto por Sirken y Shimizu (1999).

El muestreo de redes, ocupa hoy importantes lugares en las investigaciones realizadas a nivel mundial. El NCHS(1999), por ejemplo, incluye en su metodología de la Encuesta Nacional de Salud para el período 1995-2004 este método mues-

tral para la estimación del número de habitantes pertenecientes a grupos étnicos minoritarios.

2. Muestreo aleatorio simple de conglomerados con multiplicidad

2.1. Definiciones preliminares

Para una regla de conteo de multiplicidad específica, se tiene:

Multiplicidad del individuo α en la i-ésima UPM; es decir, el número de hogares relacionados al individuo α en el i-ésimo conglomerado:

$$s_{\alpha i} = \sum_{j=1}^{L_i} \delta_{\alpha ij}$$

Multiplicidad del α-ésimo individuo en la encuesta, equivalente al número total de hogares relacionados al individuo α

$$s_\alpha = \sum_{i=1}^{M} s_{\alpha i}$$

$$t_{\alpha i} = \begin{cases} 1 & \text{si } s_{\alpha i} > 0 \\ 0 & \text{si } s_{\alpha i} = 0 \end{cases}$$

Número de unidades primarias en las que existen uno o más hogares relacionados al individuo α:

$$t_\alpha = \sum_{i=1}^{M} t_{\alpha i}$$

2.2. Ponderaciones

El estimador de multiplicidad es una transformación del estimador convencional mediante la inclusión de factores de ponderación que ajustan el efecto de los individuos reportados no muestreados. Estas ponderaciones, denotadas por $Z_{\alpha i}$ para todo (α, i), tal que $t_{\alpha i} = 1$ (si el evento se relaciona al conglomerado i), deben cumplir la siguiente propiedad:

$$\sum_{i=1}^{M} Z_{\alpha i} s_{\alpha i} = 1, \qquad (\alpha = 1, ..., N) \tag{1}$$

Las opciones de ponderación más comunes son (Levy, 1977a):

1. El inverso de la multiplicidad de un individuo, $Z_{\alpha i} = 1/s_\alpha$ y
2. $Z_{\alpha i} = 1/(s_{\alpha i} t_\alpha)$

En ambos casos se cumple la propiedad (1).

La población Θ se encuentra agrupada en M unidades primarias de muestreo ($UPM'S$), cada una de las cuales contiene L_i $(i = 1, 2, ..., M)$ unidades secundarias de muestreo ($USM'S$), en este caso particular, *hogares*. En la población existen N individuos con el atributo o eventos relacionados a los hogares mediante una regla de conteo específica. Los eventos se identifican con la variable l_α $(\alpha = 1, 2, ..., N)$.

2.3. Parámetros en unidades primarias y secundarias

Una variable indicadora importante en la construcción del estimador es (Levy, 1977a):

$$\delta_{\alpha ij} = \begin{cases} 1 & \text{si el individuo } I_\alpha \text{ se relaciona al hogar } j \text{ en la UPM } i \\ 0 & \text{en cualquier otro caso} \end{cases} \tag{2}$$

donde $\alpha = 1, ..., N$; $i = 1, ..., M$ y $j = 1, ..., L_i$.

De igual manera, en las unidades primarias y secundarias se definen las siguientes variables auxiliares:

Número ponderado de individuos reportados por el j-ésimo hogar en el conglomerado i:

$$\lambda'_{ij} = \sum_{\alpha=1}^{N} Z_{\alpha i} \delta_{\alpha ij} \quad (j = 1, ..., L_i) \tag{3}$$

Número ponderado de individuos reportados por la unidad primaria i:

$$Y_i^* = \sum_{j=1}^{L_i} \lambda'_{ij} \quad (j = 1, ..., L_i)$$

Generalización del término encontrado por Sirken y Levy (1974) el cual va implícito en la varianza del estimador de multiplicidad:

$$E_i = \frac{\sum_{\alpha=1}^{N} Z_{\alpha i}^2 s_{\alpha i}}{Y_i^*}$$

Número total de hogares en la población:

$$M_0 = \sum_{i=1}^{M} L_i$$

Con base en las definiciones anteriores, se construyen las siguientes medias poblacionales que serán utilizadas en las definiciones de los estimadores y las varianzas:

$$\bar{Y}^* = \frac{1}{M} \sum_{i=1}^{M} Y_i^* = \frac{N}{M}$$

$$\bar{\lambda} = \frac{1}{M_0} \sum_{i=1}^{M} Y_i^* = \frac{1}{M_0} \sum_{i=1}^{M} \sum_{j=1}^{L_i} \lambda'_{ij}$$

$$\bar{Y}^* = \frac{1}{L_i} \sum_{j=1}^{L_i} \lambda'_{ij} = \frac{Y_i^*}{L_i}$$

2.4. Estimador del total poblacional

Dada una muestra aleatoria simple sin reemplazamiento de conglomerados, el estimador de multiplicidad del total poblacional es:

$$\widehat{N} = \frac{N}{M} \sum_{i=1}^{m} \sum_{j=1}^{L_i} \lambda'_{ij} \qquad \text{donde } \lambda'_{ij} \text{ ha sido definido en (3)} \tag{4}$$

Teorema *El estadístico descrito en (4) es un estimador insesgado para el total poblacional.*

Demostración: ver (Torres, 2001).

Teorema *La varianza del estimador de multiplicidad para el total poblacional está dada por:*

$$Var(\hat{N}) = \frac{M(M-m)}{m(M-1)}\left(M\bar{Y}^*(E_k - \bar{Y}^*) + \sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right) \tag{5}$$

donde

$$E_k = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N} Z_{\alpha i}^2 s_{\alpha i}^2}{N}$$

y

$$V_{\alpha\beta}\sum_{i=1}^{M} Z_{\alpha i}\, Z_{\beta i}\, s_{\alpha i}\, s_{\beta i}$$

Demostración: ver (Levy, 1977a).

2.5. Ganancia de eficiencia en encuestas con multiplicidad

Se definen las variables indicadoras $\mu_{\alpha ij}$ y $v_{\alpha ij}$ como:

$$\mu_{\alpha ij} = \begin{cases} 1 & \text{si el individuo } I_\alpha \text{ reside en el hogar } j \text{ de la } i\text{-ésima } UPM \\ 0 & \text{en cualquier otro caso} \end{cases}$$

$$v_{\alpha ij} = \begin{cases} 1 & \text{si el individuo } I_\alpha \text{ se relaciona al hogar } j \text{ en la } i\text{-ésima } UPM \\ & \text{pero no reside en él} \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Con base en estas variables, se definen los estimadores del total poblacional mediante los dos tipos de encuesta:

Número ponderado de individuos relacionados al j-ésimo hogar en el conglomerado i en la encuesta con multiplicidad:

$$\lambda'_{ij} = \sum_{\alpha=1}^{N} Z_{\alpha i}(\mu_{\alpha ij} + v_{\alpha ij})$$

Siendo $\mu_{\alpha ij} + v_{\alpha ij} = \delta_{\alpha ij}$.

Número ponderado de individuos relacionados al j-ésimo hogar del conglomerado i en la encuesta convencional, pero no residentes en él:

$$\pi'_{ij} = \sum_{\alpha=1}^{N} \mu_{\alpha ij}$$

2.5.1. Estimador convencional del total poblacional

El estimador del total poblacional, con base en la información de la encuesta convencional, suponiendo una muestra aleatoria simple sin reemplazamiento de conglomerados de tamaño m, está dado por:

$$\widehat{N}_\pi = \frac{M}{m}\sum_{i=1}^{m} X_i^* \quad \text{donde} \quad X_i^* = \sum_{j=1}^{L_i} \pi'_{ij}$$

Con varianza:

$$Var(\widehat{N}_\pi) = M(M-m)\frac{\sigma^2_{X^*}}{m} \tag{6}$$

siendo

$$\sigma^2_{X^*} = \frac{1}{M-1}\sum_{i=1}^{M}(X_i^* - \bar{X}^*)^2 \quad y \quad \bar{X}^* = \frac{1}{M}\sum_{i=1}^{M} X_i^*$$

2.5.2. Estimador de multiplicidad del total poblacional

Para la encuesta con multiplicidad, el estimador del total se define con base en la variable ponderada, así:

$$\widehat{N}_\lambda = \frac{M}{m}\sum_{i=1}^{m} Y_i^*$$

La varianza de este estimador está dada por:

$$Var(\widehat{N}_\lambda) = M(M-m)\frac{\sigma^2_{Y^*}}{m} \tag{7}$$

siendo

$$\sigma^2_{Y^*} = \frac{1}{M-1}\sum_{i=1}^{M}(Y_i^* - \bar{Y}^*)^2 \quad y \quad \bar{Y}^* = \frac{1}{M}\sum_{i=1}^{M} Y_i^*$$

2.6. Comparación entre los estimadores del total poblacional

Para determinar la ganancia de eficiencia del estimador de multiplicidad con respecto al estimador convencional, se realiza a continuación la comparación de las varianzas de estos estimadores.

De (6) se tiene que:

$$\frac{Var(\widehat{N}_\pi)}{\sigma^2_{X^*}} = \frac{M(M-m)}{m} \tag{8}$$

Reemplazando la igualdad (8) en (7), resulta:

$$Var(\widehat{N}_\lambda) = Var(\widehat{N}_\pi)(1-\delta)$$

donde

$$\delta = \frac{\sigma^2_{X^*} - \sigma^2_{Y^*}}{\sigma^2_{X^*}} = \frac{Var(\widehat{N}_\pi) - Var(\widehat{N}_\lambda)}{Var(\widehat{N}_\pi)} \tag{9}$$

δ indica la pérdida o ganancia de eficiencia cuando se lleva a cabo una encuesta mediante muestreo aleatorio de conglomerados con multiplicidad en comparación con la técnica clásica de conglomerados.

El efecto del diseño en la encuesta con multiplicidad es:

$$deff = \frac{Var(\widehat{N}_\lambda)}{Var(\widehat{N}_\pi)} = \frac{\sigma^2_{Y^*}}{\sigma^2_{X^*}} \tag{10}$$

De lo anterior, se concluye que $\delta = 1 - deff$ y $Var(\widehat{N}_\lambda) = Var(\widehat{N}_\pi)deff$. La expresión (10) se puede modificar utilizando los resultados de Levy (1977a), tal como se muestra a continuación:

$$\sigma^2_{Y^*} = \frac{1}{M-1}\left(M\bar{Y}^*(E_k - \bar{Y}^*) + \sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)$$

siendo

$$E_k = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N} Z^2_{\alpha i}s^2_{\alpha i}}{N} \quad \text{y} \quad N = \sum_{i=1}^{M}\sum_{j=1}^{L_i}\sum_{\alpha=1}^{N} Z_{\alpha i}\delta_{\alpha i}$$

Bajo la encuesta convencional, todo individuo con el atributo reportado por un hogar se pondera con el mismo factor,$Z_{\alpha i} = 1$, cuando se cumple que $t_{\alpha i} = 1$, por lo tanto, E_k se convierte en:

$$E_k = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N} s^2_{\alpha i}}{N} = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N}\left(\sum_{j=1}^{L_i}\delta_{\alpha ij}\right)^2}{N}$$

El término principal de E_k se descompone de la siguiente manera teniendo en cuenta (2):

$$\left(\sum_{j=1}^{L_i}\delta_{\alpha ij}\right)^2 = \sum_{j\neq k}^{L_i}\delta_{\alpha ij}\delta_{\alpha ik} + \sum_{j=1}^{L_i}\delta^2_{\alpha ij} = \sum_{j=1}^{L_i}\delta_{\alpha ij}$$

Por lo tanto:

$$E_k = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N} s_{\alpha i}^2}{N} = \frac{\sum_{i=1}^{M}\sum_{\alpha=1}^{N}\sum_{j=1}^{L_i} \delta_{\alpha ij}}{N} = 1$$

y

$$\sigma_{X^*}^2 = \frac{1}{M-1}\left(M\bar{Y}^*(1-\bar{Y}^*) + \sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)$$

Remplazando en (9):

$$\begin{aligned}\delta &= \frac{\left(M\bar{Y}^*(1-\bar{Y}^*) + \sum\limits_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right) - \left(M\bar{Y}^*(E_k-\bar{Y}^*) + \sum\limits_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)}{\left(M\bar{Y}^*(1-\bar{Y}^*) + \sum\limits_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)} \\ &= \frac{M\bar{Y}^*(1-E_k)}{\left(M\bar{Y}^*(1-\bar{Y}^*) + \sum\limits_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)} \qquad (11)\end{aligned}$$

El efecto del diseño debido a la encuesta con multiplicidad esta dado por:

$$deff = \frac{M\bar{Y}^*(E_k-\bar{Y}^*) + \sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}}{\left(M\bar{Y}^*(1-\bar{Y}^*) + \sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right)} \qquad (12)$$

2.7. Componentes de varianza del estimador de multiplicidad del total poblacional

Retomando (5) y haciendo $M/(M-1) \cong 1$, se tiene:

$$Var(\widehat{N}) = \frac{M(M-m)}{m}\left(\frac{1}{M-1}\sum_{i=1}^{M}\sum_{\alpha=1}^{N} Z_{\alpha i}^2 s_{\alpha i}^2 - \bar{Y}^{*2} + \frac{1}{M-1}\sum_{\alpha\neq\beta}^{N} V_{\alpha\beta}\right) \qquad (13)$$

2.7.1. Estimador ponderado por el inverso de la multiplicidad

El estimador de multiplicidad se pondera por $Z_{\alpha i} = 1/s_\alpha$ para todo $(i.\alpha)$, obteniéndose el estimador del total como:

$$\widehat{N} = \frac{M}{m}\sum_{i=1}^{m}\sum_{j=1}^{L_i}\lambda'_{ij} \quad \text{siendo} \quad \lambda'_{ij} = \sum_{\alpha=1}^{N}\frac{\delta_{\alpha ij}}{s_\alpha}$$

Reemplazando $V_{\alpha\beta}$ y $Z_{\alpha i}$ en (13)

$$Var(\widehat{N}) = \frac{M(M-m)}{m}\left(\frac{1}{M-1}\sum_{\alpha=1}^{N}\frac{1}{s_{\alpha i}^2}\sum_{i=1}^{M}s_{\alpha i}^2 - \bar{Y}^{*2} + \frac{1}{M-1}\sum_{\alpha\neq\beta}^{N}\sum_{i=1}^{M}\frac{s_{\alpha i}s_{\beta i}}{s_\alpha s_\beta}\right)$$

pero

$$\sum_{i=1}^{M}s_{\alpha i}^2 = \sum_{i=1}^{M}\left(\sum_{j=1}^{L_i}\delta_{\alpha ij}\right)^2 = \sum_{i=1}^{M}\left(\sum_{j\neq k}^{L_i}\delta_{\alpha ij}\delta_{\alpha ik} + \sum_{j=1}^{L_i}\delta_{\alpha ij}\right) = \sum_{i=1}^{M}\sum_{j\neq k}^{N_i}\delta_{\alpha ij}\delta_{\alpha ik} + s_\alpha$$

Por tanto:

$$Var(\widehat{N}) = \frac{M(M-m)}{m}\left(\frac{1}{M-1}\sum_{\alpha=1}^{N}\sum_{i=1}^{M}\sum_{j\neq k}^{L_i}\frac{\delta_{\alpha ij}\delta_{\alpha ik}}{s_{\alpha i}^2} + \underbrace{\frac{1}{M-1}\sum_{\alpha=1}^{N}\frac{1}{s_\alpha}}_{2^0} - \bar{Y}^{*2} + \frac{1}{M-1}\sum_{\alpha\neq\beta}^{N}\sum_{i=1}^{M}\frac{s_{\alpha i}s_{\beta i}}{s_\alpha s_\beta}\right) \tag{14}$$

Para descomponer el segundo término del paréntesis en (14) se define la variable auxiliar:

$$\gamma_{\alpha ij} = \frac{\delta_{\alpha ij}}{s_\alpha} \quad \text{con } \alpha = 1,2,...,N \text{ y } j = 1,2,...,L_i \tag{15}$$

El valor esperado de γ está dado por $E(\gamma) = N/R$, donde R es el número total de relaciones entre hogares e individuos, el cual se define como:

$$R = \sum_{i=1}^{M}\sum_{j=1}^{L_i}\sum_{\alpha=1}^{N}\delta_{\alpha,i} = \sum_{i=1}^{M}\sum_{\alpha=1}^{N}s_{\alpha i} = \sum_{\alpha=1}^{N}s_\alpha$$

La esperanza de γ se minimiza cuando los individuos en la población se relacionan con todos los hogares, es decir, $R = M_0 N$. De esta manera:

$$\frac{1}{M_0} \leqslant E(\gamma) < 1$$

Igualmente:

$$E(\gamma^2) = \frac{1}{R} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha}$$

Ahora, usando la definición de varianza:

$$var(\gamma) = E(\gamma^2) - E^2(\gamma) = \frac{1}{R} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} - \left(\frac{N}{R}\right)^2$$

Partiendo de la definición de $E(\gamma^2)$ y con base en los resultados previos, se tiene que:

$$\frac{R}{M-1} E(\gamma^2) = \frac{1}{M-1} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} = \frac{1}{M-1} \sum_{\alpha=1}^{N} \frac{s_\alpha - 1}{s_\alpha} + \frac{N}{M-1}$$

Combinando los hallazgos anteriores:

$$\frac{1}{M-1} \sum_{\alpha=1}^{N} \frac{1}{s_\alpha} = \left[\frac{R}{M-1} E(\gamma^2) - \bar{Y}^*\right] + \bar{Y}^* = \left\{\frac{\bar{Y}^* Var(\gamma)}{E(\gamma)} - \bar{Y}^*\big(1 - E(\gamma)\big)\right\} + \bar{Y}^* \quad (16)$$

Reemplazando (16) en (14) se obtiene:

$$Var(\hat{N}) = \frac{M(M-m)}{m} \left(\frac{1}{M-1} \sum_{\alpha \neq \beta}^{N} \sum_{i=1}^{M} \frac{s_{\alpha i} s_{\beta i}}{s_\alpha s_\beta} + \frac{1}{M-1} \sum_{\alpha=1}^{N} \sum_{i=1}^{M} \sum_{j \neq k}^{L_i} \frac{\delta_{\alpha ij} \delta_{\alpha ik}}{s_\alpha^2} \quad + \right.$$
$$\left. \frac{\bar{Y}^* Var(\gamma)}{E(\gamma)} - \bar{Y}^*\big(1 - E(\gamma)\big) + \bar{Y}^*(1 - \bar{Y}^*) \right) \quad (17)$$

2.7.2. Estimador ponderado según la multiplicidad del individuo en un conglomerado y el número de conglomerados relacionados con él

La ponderación del número total de individuos reportados por cada hogar de la población se hace mediante los factores $Z_{\alpha i} = 1/(t_\alpha s_{\alpha i})$ para todo (i, α). Usando esta ponderación, el estimador del total poblacional está dado por:

$$\widehat{N} = \frac{M}{m} \sum_{i=1}^{m} \sum_{j=1}^{L_i} \lambda'_{ij}$$

Siendo

$$\lambda'_{ij} = \sum_{\alpha=1}^{N} \frac{\delta_{\alpha ij}}{t_\alpha s_{\alpha i}}$$

Con varianza:

$$Var(\hat{N}) = \frac{M(M-m)}{m} \left(\frac{1}{M-1} \sum_{\alpha=1}^{N} \frac{1}{t_\alpha} - \bar{Y}^{*2} + \frac{1}{M-1} \sum_{\alpha \neq \beta}^{N} \sum_{i=1}^{M} \frac{t_{\alpha i} t_{\beta i}}{t_\alpha t_\beta} \right) \quad (18)$$

El primer término del paréntesis en (18) se puede descomponer utilizando una variable auxiliar.

Sea $\eta_{\alpha i} = \dfrac{t_{\alpha i}}{t_\alpha}$ para $(i = 1, 2, ..., N$ y $\alpha = 1, 2, ..., N)$ y $Q = \sum\limits_{\alpha=1}^{N} t_\alpha$ (número total de relaciones entre individuos y conglomerados, definidas mediante la variable $t_{\alpha i}$).

Los valores esperados de η y η^2 están dados por:

$$E(\eta) = \frac{N}{Q}$$

$$E(\eta^2) = \frac{1}{Q} \sum_{\alpha=1}^{N} \frac{1}{t_\alpha}$$

Por lo tanto,

$$\frac{1}{M-1} \sum_{\alpha=1}^{N} \frac{1}{t_\alpha} = \left\{ \frac{\bar{Y}^* Var(\eta)}{E(\eta)} - \bar{Y}^* [1 - E(\eta)] \right\} + \bar{Y}^* \quad (19)$$

Reemplazando (19) en (18):

$$Var(\widehat{N}) = \frac{M(M-m)}{m}\left(\frac{1}{M-1}\sum_{\alpha\neq\beta}^{N}\sum_{i=1}^{M}\frac{t_{\alpha i}t_{\beta i}}{t_\alpha t_\beta} \quad + \right.$$
$$\left. \frac{\bar{Y}^* Var(\eta)}{E(\eta)} - \bar{Y}^*\big(1-E(\eta)\big) + \bar{Y}^*(1-\bar{Y}^*)\right)$$

3. Ilustración del efecto de la amplitud de la regla de conteo en la eficiencia de un estimador de multiplicidad

Para la aplicación de la teoría desarrollada, se considera la información correspondiente a los registros de personas residentes en viviendas particulares censadas en Pereira en 1993 mediante el formulario número 1, según el Departamento Administrativo Nacional de Estadísticas (DANE). La base final depurada constaba de 2.583 manzanas o conglomerados, en los cuales habitaban 306.744 personas. Se trabajó con la variable ceguera del capítulo de limitaciones físicas, la cual presentaba una proporción poblacional de 0.74%, correspondiente a 2.267 individuos con ceguera. Tanto la metodología utilizada como los resultados se presentan seguidamente.

3.1. Metodología

Los pasos seguidos en el proceso de simulación fueron:

- Selección de una muestra aleatoria simple de manzanas de tamaño m ($m = 400, 600, 800, 1000y1500$), usando el paquete estadístico SAS®.

- Asignación a cada individuo seleccionado en la muestra del número de hogares relacionados con él, mediante la generación de un número pseudo-aleatorio proveniente de una variable aleatoria Poisson con media $\lambda(\lambda = 0.2, 0.4, 0.6, 0.8, 1, 1.5)$.

- Generación de otro número pseudo-aleatorio correspondiente a una variable aleatoria Binomial con parámetros n y 0.0074 para cada individuo de la muestra, donde n es el número generado de la variable aleatoria Poisson. Este último proceso se realizó para determinar aleatoriamente si los individuos relacionados a la persona muestreada poseían o no el atributo de estudio.

- Construcción de las variables para el cálculo de las estimaciones en cada una de las 1.000 simulaciones realizadas con base en la información de los individuos dentro de cada uno de los hogares de la muestra.

3.2. Resultados

La tabla 1 muestra el efecto del diseño y la ganancia de eficiencia del muestreo por conglomerados con multiplicidad con respecto al muestreo convencional por conglomerados. Siguiendo los resultados teóricos, se aprecia una mayor ganancia de eficiencia del estimador en reglas de conteo simuladas con mayor amplitud (mayor valor de λ).

Tabla 1: ***Errores estándar, efecto del diseño y ganancia de eficiencia en conglomerados con multiplicidad***

		Nº de manzanas m				
λ	***Estimador***	400	600	800	1000	1500
0.2	Convencional	271.76	204.53	177.75	144.60	101.97
	Multiplicidad	258.71	195.60	170.04	138.43	97.99
	$deff$	0.91	0.91	0.92	0.92	0.92
	$\%\delta$	9.4%	8.5%	8.5%	8.4%	7.7%
0.4	Convencional	272.22	228.69	180.62	145.06	96.96
	Multiplicidad	253.98	210.76	165.37	133.93	90.05
	$deff$	0.87	0.85	0.84	0.85	0.86
	$\%\delta$	12.9%	15.1%	16.2%	14.4%	13.7%
0.6	Convencional	286.30	211.17	177.32	149.95	98.74
	Multiplicidad	252.62	187.16	155.91	133.33	88.93
	$deff$	0.78	0.79	0.77	0.79	0.81
	$\%\delta$	22.1%	21.4%	22.7%	20.9%	18.9%
0.8	Convencional	275.43	209.00	178.40	147.22	101.08
	Multiplicidad	237.34	180.79	154.79	127.77	89.09
	$deff$	0.74	0.75	0.75	0.75	0.78
	$\%\delta$	25.7%	25.2%	24.7%	24.7%	22.3%
1	Convencional	276.06	207.82	175.40	146.61	98.35
	Multiplicidad	233.94	178.82	148.82	124.11	84.47
	$deff$	0.72	0.74	0.72	0.72	0.74
	$\%\delta$	28.2%	26.0%	28.0%	28.3%	26.2%
1.5	Convencional	270.73	206.67	173.02	149.13	99.66
	Multiplicidad	213.72	164.51	141.83	120.78	80.13
	$deff$	0.62	0.63	0.67	0.66	0.65
	$\%\delta$	37.7%	36.6%	32.8%	34.4%	35.3%

Para un tamaño de muestra de 400 manzanas, el efecto del diseño con una regla de conteo simulada bajo una variable aleatoria Poisson con media 1, puede alcanzar una ganancia de eficiencia del 28,2% en la varianza del estimador, aumentado notablemente para el caso de reglas de conteo con media igual a 1.5. Con tamaños de muestra más grandes, el comportamiento es similar. El caso contrario sucede con reglas de baja amplitud (l cercano a 0) en donde la ganancia de eficiencia del estimador de multiplicidad es cercana al 9%.

Los resultados evidencian el impacto de la regla de conteo y la importancia de este aspecto en la utilización del muestreo de redes o multiplicidad. Una regla de conteo de multiplicidad adecuada garantiza estimadores de multiplicidad más eficientes.

4. Conclusiones

Del ejercicio de simulación, con una tasa de prevalencia de 0.74%, se puede observar el efecto de la amplitud de las reglas de conteo en la eficiencia del estimador de multiplicidad. A pesar de que la simulación se realizó bajo ciertas particularidades (modelo específico de regla de conteo de multiplicidad y ausencia de errores no muestrales), se pudo observar que a medida que el parámetro de la distribución Poisson se alejaba de cero (regla de conteo más amplia), el efecto del diseño se hacía más significativo, mostrando una mayor eficiencia del muestreo por conglomerados con multiplicidad sobre la encuesta de hogares convencional.

El mayor interés de la técnica muestral se presenta con reglas de conteo de multiplicidad amplias, las cuales permiten a un individuo relacionarse con "múltiples" hogares. La regla de conteo ideal es aquella que permite a un individuo relacionarse con todos los hogares de la población. Es importante notar que la encuesta con multiplicidad esta sujeta a producir mayores errores de respuesta. En la encuesta convencional las respuestas obtenidas corresponden directamente al encuestado, mientras que en la encuesta con multiplicidad, las respuestas de la variable objetivo pueden provenir directamente del encuestado y de reportes del encuestado acerca de individuos relacionadas con él mediante alguna regla de conteo.

En esta recolección indirecta de información sobre otros miembros de la población escasa, tomando como canal los individuos encuestados en la muestra, existen mayores posibilidades de inconsistencia en los datos suministrados; por ejemplo, individuos que se reportan con el atributo y en realidad no lo poseen, u omisión de individuos relacionados al encuestado, entre otros.

Generalmente una encuesta con multiplicidad incrementa los costos de recolección y manejo de la información. Es por ello, que se debe estudiar detalladamente el problema sobre el cual se quiere aplicar esta técnica muestral para controlar el factor costo-eficiencia.

Con relación a la varianza de los dos estimadores de multiplicidad del total poblacional, según los resultados presentados en (3.6.1. y 3.6.2.), se pudo determinar el efecto de la regla de conteo de multiplicidad reflejado en cada uno de los componentes de varianza de los estimadores.

Si bien es cierto, esta técnica muestral es casi desconocida en el ámbito investigativo local, se percibe la aplicabilidad a muchas investigaciones actualmente demandadas, en las cuales el muestreo convencional presenta problemas por marcos muestrales incompletos, características evasivas o raras. Mas aún, en estudios donde además de requerir estimar el total poblacional, se desea tener un número adecuado de contactos para llevar a cabo entrevistas que conduzcan a estimaciones de características relacionadas con el grupo especial. Mediante el método convencional, generalmente se contactan pocos miembros de dichas poblaciones haciendo difícil la tarea de los investigadores.

References

[1] BAUZA, C. (1999) *Diseños muestrales no tradicionales para estimar parámetros de interés en problemas de control de calidad, en: Memorias del Simposio de Estadística de Control Estadístico de la Calidad, Rionegro (Antioquia). Agosto de 1999. Universidad Nacional de Colombia, Bogotá.*

[2] CARY, N.C., SAS Institute Inc., USA, SAS for Windows Release 8.0

[3] CASADY, R., NTHAN, G, and SIRKEN, M. (1985) *Alternative Dual System Network Estimators.* International Statistical Review, vol. 53, N°2, pag. 183-197.

[4] CZAJA, R., SNOWDEN, C. B., andCASADY, R. (1986) *Reporting Bias and Sampling Errors in a Survey of a Rare Population Using Multiplicity Counting Rules*,Journal of the American statistical Association, Vol. 81, 411-419.

[5] CZAJA, R., TRUNZO, D.,ROYSTON, P. (1992) *Response Effects in a Network Sampling*, Sociological Methods & Research, Vol. 20, N° 3, 340-366.

[6] GRAHAM, K., DALLAS, A. (1986) *Sampling Rare Populations*Journal of the Royal Statistical Society, Vol. 149, Part 1, 65-82.

[7] GRANOVETTER, M. (1977) *Network Sampling: Some First Steps*, American Journal of Sociology, Vol. 81, 1287-1303.

[8] GOODMAN, L. A. (1961) *Snowball Sampling*, Annals of Mathematical Statisitics 32, p. 148-170.

[9] LEVY, S. P. (1972) *Quality Control of Statistical Reports*, Proceedings of the American Statistical Association, Social Statistics Section, 356-359.

[10] LEVY, S. P. (1977a) *Estimation of Rare Events by Simple Cluster Sampling With Multiplicity*, Proceedings of the American Statistical Association, Social Statistics Section, 963-966.

[11] LEVY, S. P. (1977b) *Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations*, Journal of the American Statistical Association, Vol. 72, 758-763.

[12] NATHAN, G. (1976) *The Evaluation of Different Counting Rules and Weighting Procedures for Surveys with Multiplicity*, Proceedings of the American Statistical Association, Social Statistics Section, 639-644.

[13] NCHS (1999) *National Health Interview Survey: Research for the 1995-2004 Redesign*, Vital and Health Statistics, Series 2, N° 126.

[14] SHIMIZU, I., SIRKEN, M. (1998) *More on Population Based Establishment Surveys*, Proceedings of the American Statistical Association, Survey Research Methods Section, 7-12.

[15] SIRKEN, M. G. (1970) *Household Surveys with Multiplicity*, Journal of the American statistical Association, Vol. 65, 257-266.

[16] SIRKEN, M. (1972a) *Variance Components of Multiplicity Estimators*, Biometrics, Vol. 28, N° 3, 869-873

[17] SIRKEN, M. (1972b) *Stratified Sample Surveys with Multiplicity*, Journal of the American statistical Association, Vol. 67, 224-227.

[18] SIRKEN, M. (1974) *Multiplicity Estimation of Proportions Based on Ratios of Random Variables*, Journal of the American Statistical Association, Vol. 69, 68-73.

[19] SIRKEN, M. (1975) *Network Surveys*, Proceedings of the International Statistical Institute, Section 40th, Vol. 76, Tomo 4, 332-342.

[20] SIRKEN, M. (1979) *A Dual System Network Estimator, Stratified Sample Surveys with Multiplicity*, Proceedings of the American Statistical Association, Survey Research Methods Section, 340-342.

[21] SIRKEN, M. (1983) *Handling Missing Data by Network Sampling, Incomplete Data in Sample Surveys*, Vol. 2, Part II, Chapter 8. Great Britain: Academic Press, Inc.

[22] SIRKEN, M. (1998) *A Short History of Network Sampling*, Proceedings of the American Statistical Association, Survey Research Methods Section.

[23] SIRKEN, M. (1999) *Population Based Establishment Sample Surveys: The Horvitz-Thompson Estimator*, Survey Methodology, Vol. 25, N°2, 187-191.

[24] SPREEN. (1999) *Sampling Personal Network Structures: Statistical Inference in Ego-graphs, Disertación (Sociología)*, Department Statistics & Measurement Theory. University of Maastricht, Interuniversity Center for Social Science Theory and Methodology. Groningen (Suecia).

[25] SUDMAN, S., SIRKEN, M. and COWAN, Ch. (1988) *Sampling Rare and Elusive Populations*, Science, Vol. 240, 991-996.

[26] TORRES S., PEDRO A. (2001) *Muestreo de Redes: Una alternativa para estimar el total poblacional en poblaciones raras mediante encuentas de hogares*, Trabajo de grado. Estadístico. Universidad Nacional de Colombia, Bogotá.

Agradecimientos

Queremos expresar nuestros agradecimientos al Doctor Monroe Sirken, Director de la oficina de Métodos Estadísticos del National Center for Health Statistics (NCHS), a los doctores Ronald F. Czaja (Profesor de Sociología y Antropología de North Carolina State University), Mark Granovetter (Profesor de Sociología de Stanford University), Paul Levy (Profesor del programa de Biometría de University of Illinois), Gad Nathan (Profesor de Estadística de The Hebrew University of Jerusalem) y Marinus Spreen (Profesor del Department of Methodology and Statistics, University of Maastricht), por sus importantes aportes bibliográficos. De igual manera, agradecemos al colega Carlos Cáceres por la colaboración en la elaboración del programa para la simulación.