Metodología de Encuestas
Vol 3, Núm 1, 2001, 97-111

DISEÑOS MUESTRALES UNIFORMES APROPIADOS PARA ESTIMAR LA FUNCIÓN DE DISTRIBUCIÓN POBLACIONAL EMPLEANDO MUESTREO POR CONGLOMERADOS EN UNA ETAPA

José A. Mayor Gallego
Manuel Martínez Blanes
Universidad de Sevilla

RESUMEN

Para estimar la función de distribución de una variable sobre una población finita, se propone el estimador de Horvitz-Thompson, lo que proporciona una estrategia muestral insesgada. La varianza de dicho estimador es una función real de variable real, cuya minimización permite obtener diseños muestrales óptimos bajo diferentes criterios. En este trabajo empleamos la norma $\|\cdot\|_1$ como criterio de optimización, minimizando la norma de la función varianza. De esta forma, considerando como dominio de búsqueda el conjunto de los diseños muestrales uniformes, en el sentido de ser iguales las probabilidades de inclusión de primer orden, se estudia la obtención de diseños muestrales adecuados, bajo muestreo por conglomerados de una etapa.

Palabras clave: muestreo, poblaciones finitas, diseño muestral, función de distribución, conglomerados.

Introducción

Aunque la teoría del muestreo en poblaciones finitas se ha centrado fundamentalmente en la estimación de parámetros poblacionales de tipo puntual, como totales, medias, proporciones y varianzas, existe una serie de parámetros de tipo funcional que pueden proporcionar información relevante acerca del comportamiento global de la población.

En este trabajo se considera el problema de la estimación de un parámetro de este tipo, en concreto, la función de distribución poblacional asociada a una variable numérica definida sobre la población. Este problema es importante por el interés intrínseco del parámetro funcional mencionado y también por su relación con otros parámetros de tipo no funcional, como la mediana, los cuantiles o el índice de Gini.

En efecto, los parámetros de tipo funcional permiten conocer de forma muy completa el comportamiento global de una variable con respecto a diferentes criterios. Consideraremos por ejemplo una variable, Y, definida sobre una población, y la función de distribución asociada

$$F(t) = \frac{1}{N} \mathrm{CARD}\left(\left\{i \in U | Y_i \leq t\right\}\right)$$

donde CARD denota cardinal de conjunto, es decir, $F(t)$ representa la proporción de elementos o unidades poblacionales que presentan un valor de la variable menor o igual que t. Otra función poblacional importante es la función de concentración

$$G(t) = \frac{1}{T(Y)} \sum_{i \in U, Y_i \leq t} Y_i$$

siendo $T(Y)$ el total de la variable. Esta función expresa la proporción, con respecto al total poblacional de la variable, que representan los valores de la variable menores o iguales que t.

Como ejemplo, mostramos en la figura 1 estas dos funciones para una población, denominada DCA757, consistente en 757 localidades de la Comunidad Autónoma Andaluza, con exclusión de las capitales de provincia, siendo Y una variable numérica que indica el número de habitantes de cada localidad.

Como puede verse, la función de concentración se mantiene marcadamente separada de la función de distribución. Por ejemplo, para $F(t) = 0{,}5$, el valor de t, es decir, la mediana, es 2500, ello significa que las localidades con 2500 habitantes o menos representan el 50% de todas las localidades, pero los habitantes que acumulan ese 50% representan menos del 10% del total de habitantes. De hecho, son las

poblaciones con un número de habitantes menor o igual que 15 millas que concentran el 50% del total de habitantes.

La estimación de las funciones de distribución o de concentración tienen alicientes especiales. No sólo cuentan con un interés de por sí, que comparten con la estimación de cualesquiera características poblacionales. También constituyen un resultado intermedio, un instrumento para llegar hasta la estimación de multitud de parámetros.

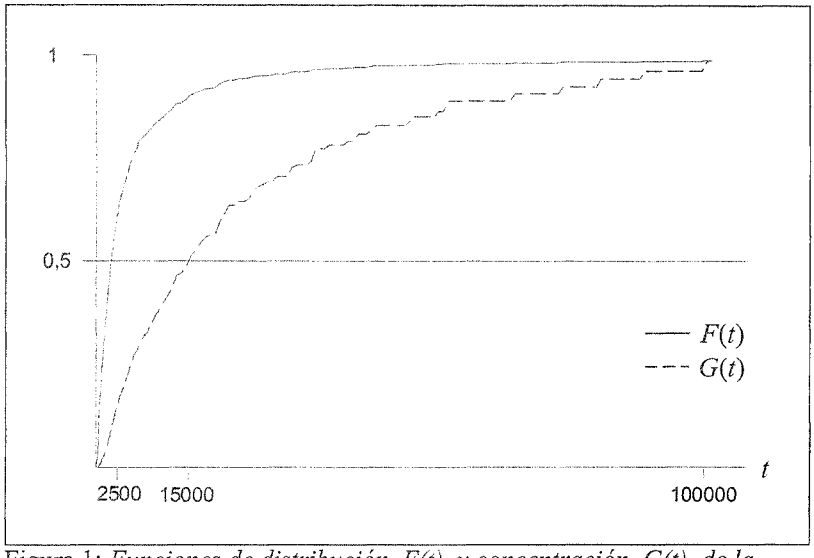

Figura 1: *Funciones de distribución, F(t), y concentración, G(t), de la variable HABITANTES en la población DCA757.*

En la bibliografía especializada es posible encontrar diferentes enfoques para abordar tanto el problema de la estimación de la función de distribución como los parámetros relacionados con la misma. Así, en relación a la estimación de la mediana y los cuantiles, es obligado citar el trabajo inicial de Wooddruff (1952), en el que se construye un intervalo de confianza para la estimación de la mediana poblacional y otras medidas de posición, empleando el muestreo aleatorio simple. Por otra parte, Sendransk y Meyer (1978) estudian este problema bajo un enfoque puramente probabilístico de distribución de estadísticos ordenados, para muestreo aleatorio simple y estratificado. Hill (1968) emplea un enfoque bayesiano; y Kuk y Mak (1989), técnicas de información auxiliar proporcionada por otras variables.

Para el problema de la estimación de la función de distribución poblacional propiamente dicha, también encontramos diferentes enfoques en la bibliografía. Por ejemplo, Chambers y Dunstan (1986) emplean un modelo de superpoblación para desarrollar un procedimiento de estimación. Kuk (1988) estudia y compara varios estimadores de la función de distribución poblacional, empleando muestreo con

probabilidades variables; y Rao, Kovar y Mantel (1990) desarrollan estimadores que emplean información auxiliar. Citemos también los trabajos de Chambers, Dorffman y Hall (1992) y Rao (1994).

Aquí consideraremos este problema con un enfoque distinto de los anteriores, y que se basa en el estudio de la función de varianza del estimador de Horvitz-Thompson, con el fin de buscar diseños muestrales óptimos, en una clase especial de diseños muestrales. para ello, vamos a considerar una población finita, $U = \{1, 2 \ldots N\}$ y una variable de estudio numérica, Y, cuyos valores sobre U son $(Y_1, Y_2 \ldots Y_N\}$ que, sin pérdida de generalidad, supondremos ordenados de menor a mayor, esto es, $Y_1 \leq Y_2 \leq \ldots \leq Y_N$. Nuestro objetivo es la estimación de la función de distribución de la variable Y,

$$F(t) = \frac{1}{N} \mathrm{CARD}(\{i \in U \mid Y_i \leq t\})$$

Si empleamos las funciones indicadoras de los intervalos de la forma $[Yi, +\infty)$, denotándolas $I_{[Yi, +\infty)}(t)$, podemos expresar $F(t)$ como

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i, +\infty)}(t)$$

es decir, una expresión lineal, que puede ser estimada mediante el estimador de Horvitz-Thompson. Sea pues m una muestra obtenida de U mediante un diseño muestral $d=(M,P(\cdot))$, sin reemplazamiento y con matriz de diseño que denotamos $\Pi=\{\pi_{ij}\}_{1 \leq i,j \leq N}$, con $\pi_{ii}=\pi_i$. El mencionado estimador de $F(t)$ resulta ser

$$\hat{F}(t) = \frac{1}{N} \sum_{i \in m} \frac{I_{[Y_i, +\infty)}(t)}{\pi_i}$$

Como sabemos, este estimador es insesgado y su varianza puede ser expresada mediante la clásica fórmula

$$V\left[\hat{F}(t)\right] = \frac{1}{N^2} \sum_{i,j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{I_{[Y_i, +\infty)}(t)}{\pi_i} \frac{I_{[Y_j, +\infty)}(t)}{\pi_j}$$

y, así, para cada valor de $t \in \mathbb{R}$, podemos emplear dicha expresión como medida de la bondad de la estimación.

Muestreo por conglomerados en una etapa

Al ser la varianza de una función, no tiene sentido hablar de varianza mínima en el sentido usual. Por ello, vamos a definir un criterio apropiado que nos permita obtener propiedades de los diseños muestrales más adecuados para la estimación que estamos realizando. El criterio que definimos viene determinado por la siguiente distancia, basada en la norma funcional $\|\cdot\|_1$,

$$d\left(V\left[\hat{F}(t)\right],0\right)=\left\|V\left[\hat{F}(t)\right]\right\|_1=\int_{S(Y_1,Y_N)}\left|V\left[\hat{F}(t)\right]\right|\mathrm{dt}=\int_{S(Y_1,Y_N)}V\left[\hat{F}(t)\right]\mathrm{dt}$$

siendo $S(Y_1,Y_N)$ el soporte de la variable Y, es decir, $S(Y_1,Y_N)=[Y_1,Y_N]$. Notemos que

$$V\left[\hat{F}(t)\right]$$

cumple los requerimientos analíticos para que esta distancia esté bien definida.

Supondremos que se realiza un muestreo por conglomerados en una etapa. Así, la población sobre la que se realiza el muestreo es una población de conglomerados,

$$U_c=\{C_1,\ldots,C_i,\ldots,C_M\},$$

cada uno con tamaños respectivos $N_1,\ldots,N_i,\ldots,N_M$.

Sobre esta población se emplea un diseño muestral $d_c=(M_c,P_c)$, con matiz de diseño $\{\pi_{ij}^c\}$, siendo entonces las probabilidades de inclusión de las unidades finales:

$$\pi_k = \pi_i^c \quad \text{si } k\in C_i \qquad \forall k\in U$$

$$\pi_{kl} = \begin{cases}\pi_{ij}^c & \text{si } k\in C_i, l\in C_j \\ \pi_i^c & \text{si } k,l\in C_i\end{cases} \qquad \forall k,l\in U$$

De esta forma, denotando por m_c la muestra obtenida de n conglomerados, y por m a la muestra de unidades finales provenientes de dichos conglomerados, el estimador de Horvitz-Thompson será:

$$\hat{F}(t)=\sum_{k\in m}\frac{1}{N}\frac{I_{[Y_k,+\infty)}(t)}{\pi_k}=\sum_{i\in m_c}\sum_{k\in C_i}\frac{I_{[Y_k,+\infty)}(t)}{\pi_k}=$$

$$=\sum_{i\in m_c}\frac{1}{\pi_i^c}\sum_{k\in C_i}\frac{1}{N}I_{[Y_k,+\infty)}(t)=\sum_{i\in m_c}\frac{F_i(t)}{\pi_i^c}$$

donde $F_i(t)$ denota el parámetro funcional que se está estimando, valorado sobre todos los elementos del conglomerado C_i, es decir,

$$F_i(t) = \sum_{k \in C_i} \frac{1}{N} I_{[Y_k, +\infty)}(t)$$

La varianza de dicho estimador vendrá dada por la expresión clásica, que adaptada a la estructura poblacional considerada, será

$$V\left[\hat{F}(t)\right] = \sum_{i \in U_c} \sum_{j \in U_c} \left(\pi_{ij}^c - \pi_i^c \pi_j^c\right) \frac{F_i(t)}{\pi_i^c} \frac{F_j(t)}{\pi_j^c}$$

Para aplicar nuestra metodología, vamos a calcular la norma $\|\cdot\|_1$ de esta varianza. En primer lugar, observemos que

$$F_i(t) F_j(t) = \frac{1}{N^2} \sum_{k \in C_i} \sum_{l \in C_j} I_{[Y_k, +\infty)}(t) I_{[Y_l, +\infty)}(t)$$

y que

$$\int_{S(Y_1, Y_N)} I_{[Y_k, +\infty)}(t) I_{[Y_l, +\infty)}(t)\, dt = \int_{S(Y_1, Y_N)} I_{[\max\{Y_k, Y_l\}, +\infty)}(t)\, dt = Y_N - \max\{Y_k, Y_l\}$$

siendo, pues,

$$\int_{S(Y_1, Y_N)} F_i(t) F_j(t)\, dt = \frac{1}{N^2} \sum_{k \in C_i} \sum_{l \in C_j} \int_{S(Y_1, Y_N)} I_{[Y_k, +\infty)}(t) I_{[Y_l, +\infty)}(t)\, dt =$$

$$= \frac{1}{N^2} \sum_{k \in C_i} \sum_{l \in C_j} \left(Y_N - \max\{Y_k, Y_l\}\right)$$

Por consiguiente,

$$\left\| V\left[\hat{F}(t)\right] \right\|_1 = \int_{S(Y_1, Y_N)} \sum_{i \in U_c} \sum_{j \in U_c} \left(\pi_{ij}^c - \pi_j^c \pi_j^c\right) \frac{F_i(t)}{\pi_i^c} \frac{F_j(t)}{\pi_j^c}\, dt =$$

$$= \sum_{i\in U_c}\sum_{j\in U_c}\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\int_{S(Y_1,Y_N)} F_i(t)\,F_j(t)\,dt - \sum_{i\in U_c}\sum_{j\in U_c}\int_{S(Y_1,Y_N)} F_i(t)\,F_j(t)\,dt$$

Y se tiene

$$\sum_{i\in U_c}\sum_{j\in U_c}\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\int_{S(Y_1,Y_N)} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{i\in U_c}\sum_{j\in U_c}\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\sum_{k\in C_i}\sum_{l\in C_j}\left(Y_N - \max\{Y_k, Y_l\}\right)$$

y

$$\sum_{i\in U_c}\sum_{j\in U_c}\int_{S(Y_1,Y_N)} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{i\in U_c}\sum_{j\in U_c}\sum_{k\in C_i}\sum_{l\in C_j}\left(Y_N - \max\{Y_k, Y_l\}\right)$$

Así pues, podemos afirmar que los mejores diseños muestrales, bajo el criterio que hemos introducido, son los que tienden a hacer mínimo el término

$$V_1 = \sum_{i\in U_c}\sum_{j\in U_c}\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\sum_{k\in C_i}\sum_{l\in C_j}\left(Y_N - \max\{Y_k, Y_l\}\right)$$

puesto que el segundo término no depende del diseño muestral.

Observemos que la variable Y es desconocida, por lo que el problema de minimización obtenido no puede ser considerado directamente. Por esta razón, realizaremos algunas hipótesis sobre la estructura de la población. Así, vamos a estudiar la expresión a minimizar para dos estructuras poblacionales muy comunes, y que formalizaremos mediante modelos de superpoblación.

Modelo de superpoblación completamente aleatorio

Este modelo viene dado par las siguientes especificaciones

$$Y_k = \alpha + \varepsilon_k, \qquad E_s[\varepsilon_i] = 0 \qquad \forall k \in U$$

y formaliza la situación real en la que la variable de estudio fluctúa globalmente en torno a un valor medio o central, no estando en concomitancia con ninguna otra variable inherente al problema.

Con este modelo, sustituimos el problema de minimización anteriormente considerado, por el de minimizar la esperanza, en dicho modelo, de la función objetivo; es decir, minimizar $E_s[V_1]$. A continuación, calculamos dicha esperanza. Observemos, en primer lugar, que

$$Y_k \le \max\{Y_k, Y_l\}, \qquad Y_l \le \max\{Y_k, Y_l\}, \qquad \forall k,l \in U$$

por consiguiente,

$$E_s[Y_k] \le E_s[\max\{Y_k, Y_l\}], \qquad E_s[Y_l] \le E_s[\max\{Y_k, Y_l\}], \qquad \forall k,l \in U$$

de donde se deduce que

$$\max\{E_s[Y_k], E_s[Y_l]\} \le E_s[\max\{Y_k, Y_l\}], \qquad \forall k,l \in U$$

y se tiene, pues,

$$E_s[(Y_N - \max\{Y_k, Y_l\})] = E_s[Y_N] - E_s[\max\{Y_k, Y_l\}] \le \alpha - \max\{E_s[Y_k], E_s[Y_l]\} = \alpha - \alpha = 0$$

Así pues, tendremos que $E_s[V_1]$=0, es decir, bajo el criterio empleado, cualquier diseño muestral proporcionará análogos resultados, pudiéndose pues aplicar el muestreo aleatorio simple por razones de facilidad. Este resultado es interesante pues contrasta con el hecho, bien conocido, de la influencia de los tamaños de los conglomerados en la estimación de parámetros como la media y el total poblacionales, cuando se emplea muestreo aleatorio simple de conglomerados (véase, Fernández y Mayor, 1995).

Modelo de superpoblación de regresión lineal por conglomerados

Sus especificaciones son

$$Y_k = \alpha + \beta X_i + \varepsilon_k, \qquad \beta > 0, \qquad E_s[\varepsilon_k] = 0 \qquad \forall k \in C_i, \forall i \in U$$

Dicho modelo formaliza aquellas situaciones en las que existe una relación aproximada, de tipo lineal, entre la variable de estudio y una variable auxiliar, completamente conocida, X, con el mismo valor en cada uno de los conglomerados.

Este tipo de relaciones suelen darse en numerosas situaciones reales en las que la variable de estudio presenta cierta homogeneidad en cada conglomerado, pero éstos tienen comportamientos distintos, aunque con un patrón de tipo lineal en relación a una variable conocida sobre U_c.

Como en el caso anterior, sustituiremos el problema de minimización inicial por el de minimizar $E_s[V_1]$. En este caso, denotando con X_{MAX} el valor máximo de la variable X, y por v el índice del conglomerado al que pertenece la unidad poblacional N-ésima, y suponiendo que $\forall k \in C_i$ y $\forall l \in C_j$, se tiene

$$E_s[(Y_N - \max\{Y_k, Y_l\})] = E_s[Y_N] - E_s[\max\{Y_k, Y_l\}] \leq \alpha + \beta X_v - \max\{E_s[Y_k], E_s[Y_l]\} =$$

$$= \alpha + \beta X_v - \max\{\alpha + \beta X_i, \alpha + \beta X_j\} \leq \beta(X_{MAX} - \max\{X_i, X_j\})$$

con lo que los diseños muestrales apropiados para la estimación que estamos realizando son aquéllos que minimizan la expresión

$$\sum_{i \in U_c} \sum_{j \in U_c} \frac{\pi_{ij}^c}{\pi_i^c \pi_j^c} N_i N_j (X_{MAX} - \max\{X_i, X_j\})$$

Implementación bajo el modelo de regresión lineal

A continuación, vamos a estudiar la viabilidad del método, previamente considerado, en el sentido de su aplicabilidad a situaciones reales, en el contexto del segundo modelo de superpoblación considerado, pues para el primero ya hemos visto que el muestreo aleatorio simple es la elección más apropiada.

En primer lugar, consideremos el estimador empleado en este trabajo,

$$\hat{F}(t) = \sum_{i \in m_c} \frac{F_i(t)}{\pi_i^c}$$

con

$$F_i(t) = \sum_{k \in C_i} \frac{1}{N} I_{[Y_k, +\infty)}(t)$$

Podemos observar que no tiene sentido intentar reducir la varianza de la estimación empleando probabilidades de inclusión de primer orden proporcionales al tamaño, como se hace cuando se estiman parámetros puntuales como medias o totales, ya que aquí dichas probabilidades afectan a funciones con una variabilidad inherente. Pensemos que si escogemos unas probabilidades de tal tipo para un determinado valor de la variable t, dichas probabilidades pueden no ser adecuadas para otro valor diferente. Esto justifica que, buscando simplificar la implementación de la metodología introducida, restrinjamos la búsqueda a diseños muestrales de tipo uniforme, en el sentido de ser constantes las probabilidades de inclusión de primer orden.

Con esta elección, tendremos necesariamente $\pi_i^c = n/M, \forall i \in U_c$ y, con esta simplificación, la expresión a minimizar queda como sigue

$$\sum_{i \in U_c} \sum_{\{j \neq i\} \in U_c} \pi_{ij}^c N_i N_j (X_{MAX} - \max\{X_i, X_j\})$$

Es decir, se trata de una función objetivo lineal de las variables $\pi_{ij}^{c}, i \neq j$. Observemos que, por la simetría de dichas variables, tendremos en total $M(M-1)/2$ variables y, teniendo en cuenta que el número de conglomerados no suele ser elevado en comparación con el tamaño de la población, el problema resultante posee un tamaño razonable para abordar su resolución con los programas usuales. Por ejemplo, para un problema con M=100 conglomerados, tendríamos un problema de programación lineal con 4950 variables, resoluble por las rutinas de uso común para este tipo de problemas.

En cuanto a las restricciones del problema, tendremos en primer lugar $\pi_{ij}^{c} > 0$, con objeto de que el diseño permita estimar la varianza de la estimación. Por otra parte, de la relación entre probabilidades de inclusión de primer y segundo orden, obtendremos las restricciones

$$\sum_{j \in U_c, j \neq i} \pi_{ij}^{c} = (n-1)\pi_i^{c} = \frac{(n-1)n}{M} \qquad \forall i \in U_c$$

y, finalmente, con objeto de poder obtener estimaciones no negativas de la varianza de la estimación, tendremos las restricciones

$$\pi_{ij}^{c} \leq \pi_i^{c}\pi_j^{c} = \frac{n^2}{M^2} \qquad \forall i, j \in U_c, i \neq j$$

En el apéndice de este trabajo se exponen algunos aspectos computacionales, incluyendo una sencilla rutina de implementación en el lenguaje de descripción de problemas AMPL (Fourier y col. ,1993), así como un ejemplo numérico sobre tres poblaciones de conglomerados.

Conclusiones

En primer lugar, es interesante observar que los métodos clásicos estudiados en el muestreo en poblaciones finitas, para estimar los parámetros usuales como medias y totales, no son los más apropiados para estimar parámetros de tipo funcional, como lo es la función de distribución poblacional.

El método introducido en este trabajo, basado en minimizar una determinada norma de la función $V\left[\hat{F}(t)\right]$, en este caso, la norma $\|\cdot\|_1$, se muestra como una alternativa prometedora, pues minimizando dicha norma también se minimiza globalmente dicha función de varianza.

Para el caso de una estructura poblacional completamente aleatoria, formalizada por el correspondiente modelo de superpoblación, la expresión a minimizar es constante, con lo que concluimos que todos los diseños muestrales proporcionan

similares resultados, pudiéndose emplear el muestreo aleatorio simple por razones de facilidad.

Finalmente, para el caso de que exista relación entre la variable de estudio y una variable auxiliar, formalizada mediante un modelo de superpoblación más general, de tipo regresión lineal, la reducción de varianza se realiza a partir de la resolución de un problema de programación lineal, que proporciona las probabilidades de inclusión de segundo orden. Los resultados numéricos obtenidos en el estudio empírico que se muestra en el apéndice, indican que este método puede producir de forma efectiva una reducción del error de la estimación.

Referencias

Chambers, R.L. y Dunstan, R. (1986) Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chambers, R.L.; Dorfman, A.H. y Hall, P. (1992) Properties of stimators of the finite population distribution function. *Biometrika*, 79, 577-582.

Fernández, F.R. y Mayor, J.A. (1995) *Muestreo en poblaciones finitas: curso básico.* Barcelona: PPU.

Fourier, R.; Gay, D.M. y Kernighan, B.W. (1993) *AMPL. A modeling language for mathematical programming.* Danvers, Massachusetts: Boyd & Fraser Publishing Company.

Hill, B.M. (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63, 677-691.

Kuk,, A.Y.C. y Mak, T.K. (1989) Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.

Rao, J.N.K.; Kovar, J.G. y Mantel, H.J. (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J.N.K. (1994) Estimating totals and distributions functions using auxiliary information in the estimation stage. *Journal of Official Statistics*, 10, 153-166.

Sedransk, J. y Meyer, J. (1978) Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling. *Journal of the Royal Statistical Society, Series B*, 40, 239-252.

Woodruff, R.S. (1952) Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Apéndice

En el cuadro 1, se expone de forma esquemática una rutina en el lenguaje de descripción de problemas AMPL, para el tratamiento del problema considerado.

Obsérvese que la restricción

$$\pi_{ij}^{c} > 0 ,$$

intratable computacionalmente, ha sido sustituida por

$$\pi_{ij}^{c} > \varepsilon ,$$

siendo $\varepsilon > 0$ un número real muy próximo a cero.

Realizamos, seguidamente, un estudio empírico empleando tres poblaciones, P_1, P_2 y P_3, cada una con M=20 conglomerados. Para simplificar la exposición, supondremos que los tamaños de los conglomerados son todos iguales.

Para la primera población, la variable auxiliar X, ya ordenada, toma los valores:

$$\{55, 57, 59, 60, 61, 63, 63, 64, 65, 65, 65, 66, 66, 66, 67, 67, 67, 68, 68, 68\}$$

es decir, datos con cierta homogeneidad. Para P_2, la variable X toma los valores

$$\{55, 57, 59, 60, 100, 120, 150, 160, 180, 200, 500, 550, 600, 700, 750, 800, 2000, 3000, 5000, 10000\}$$

presentando gran heterogeneidad y con algunos valores muy desviados de la masa principal de datos.

```
param M;
set U:= 1..M;
set PAIRS:={i in U,j in U: i < j};
param X{U} >= 0;
var PI{PAIRS} >= epsilon;
minimize FUNC: sum{(i,j) in PAIRS} (PI[i,j]*Ni*Nj
        *(XMAX-max(X[i],X[j])));
subject to RELATION {i in U}: sum {j in U: j>i} PI[i,j]
        + sum{j in U : j < i} PI[j,i] = (n-1)*n(M;
subject to BOUND {(i,j) in PAIRS}: PI[I,J] <= n*n(M*M;
```

Cuadro 1: *rutina AMPL para la simulación de distribuciones.*

El método considerado en este trabajo será comparado con el muestreo aleatorio simple de conglomerados, para lo que utilizaremos como medida de eficiencia relativa la siguiente cantidad, expresada como porcentaje:

$$C = 100 \text{ x } \frac{V_{\text{MAS}} - V_{\text{MET}}}{V_{\text{MAS}}} \%$$

siendo

$$V_{\text{MET}} = \sum_{i \in U_c} \sum_{j \in U_c} \left[\frac{\pi_{ij,\text{MET}}^{c}}{(n/M)^2} - 1 \right] \left(X_{\text{MAX}} - \max\{X_i, X_j\} \right)$$

y

$$V_{\mathrm{MAS}} = \sum_{i\in U_c}\sum_{j\in U_c}\left[\frac{\pi^{c}_{ij,\mathrm{MAS}}}{(n/M)^2} - 1\right]\left(X_{\mathrm{MAX}} - \max\{X_i, X_j\}\right)$$

El empleo de estas cantidades se justifica en la teoría previamente desarrollada y, en concreto, en la norma de la varianza obtenida en el segundo apartado. Observemos que en V_{MET} aparecen las probabilidades de inclusión de segundo orden obtenidas por minimización, mientras que en V_{MAS} aparecen las correspondientes al muestreo aleatorio simple de conglomerados, es decir,

$$\pi^{c}_{ij,\mathrm{MAS}} = n(n-1)/[M(M-1)]\; i \neq j$$

Un coeficiente mayor que cero indicará un aumento de eficiencia con respecto al muestreo aleatorio simple, bajo el criterio considerado, y el correspondiente porcentaje indicará la cuantía de dicho aumento. Los resultados comparativos obtenidos se exponen en la tabla 1, para cada una de las tres poblaciones y para tamaños muestrales de valor n=2, 3, 4 y 5.

Tabla 1: *Valor del coeficiente C.*

n	P_1	P_2	P_3
2	37,4	29,1	21,2
3	59,1	43,8	30,1
4	67,8	53,5	41,6
5	75,3	59,3	52,3

El coeficiente se expresa en %.

Como puede verse, en todos los casos se ha obtenido una evidente reducción de la varianza, lo que manifiesta que el método estudiado en este trabajo se presenta como una alternativa muy prometedora.

Realizamos, seguidamente, una simulación, empleando los datos de la población P_2 y generando los valores a partir del modelo de superpoblación

$$\boxed{Y_k = 30 + 2X_i + \varepsilon_k, \quad \forall k \in C_i, \forall i \in U}$$

siendo $\varepsilon \sim N(0,10^2)$, $\forall k$, es decir, variables aleatorias normales con μ=0 y σ^2=10^2.

Para cada conglomerado, se han generado N_i=100 valores de la variable Y, y se ha empleado un tamaño de muestra n=5, seleccionando una muestra aleatoria simple, y una muestra según el método expuesto en este trabajo.

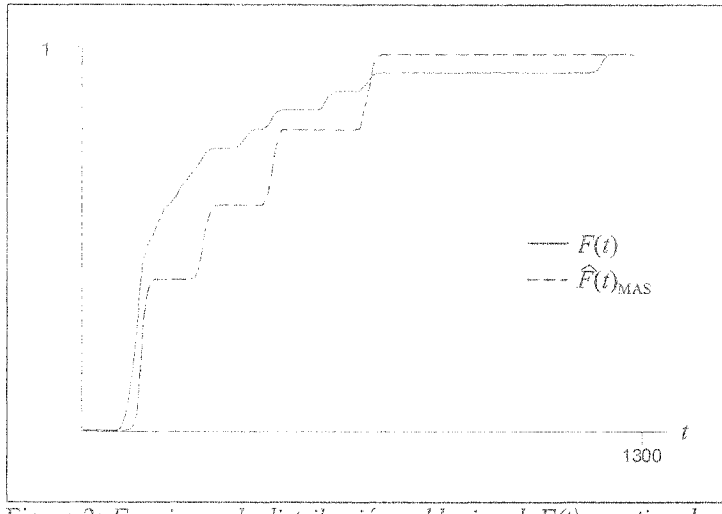

Figura 2: *Funciones de distribución poblacional, F(t), y estimada mediante un muestreo aleatorio simple, $\hat{F}(t)_{MAS}$, para la población P_2 descrita en el texto.*

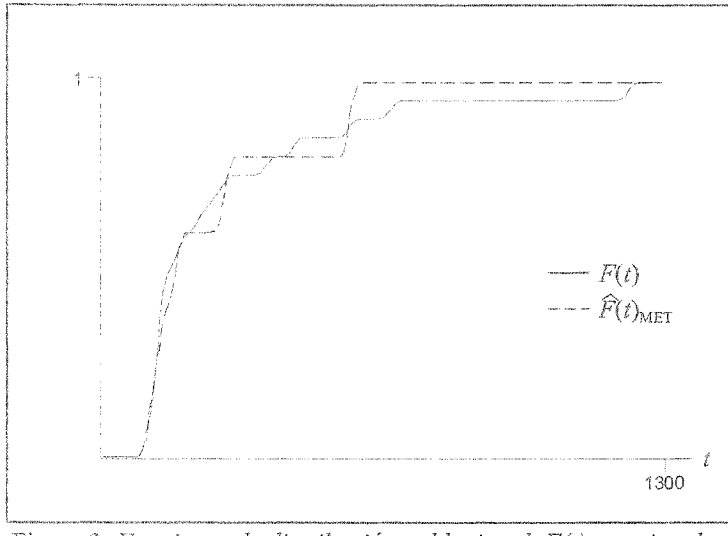

Figura 3: *Funciones de distribución poblacional, F(t), y estimada mediante un muestreo basado en el método propuesto, $\hat{F}(t)_{MET}$, para la población P_2 descrita en el texto.*

En las figuras 2 y 3 se compara la función de distribución poblacional con las estimadas mediante cada uno de los métodos. Nótese que, a pesar de la natural aleatoriedad inherente a este tipo de comparaciones, la estimación basada en nuestro método resulta más acurada que la obtenida a partir del muestreo aleatorio simple.

Para concluir, exponemos en la tabla 2 los cuartiles Q_1, Q_2 (mediana) y Q_3, tanto los estimados a partir de las estimaciones de la función de distribución, como los valores poblacionales exactos. Como puede verse, la mejora en la estimación de la función de distribución se traduce también en mejores estimaciones de los cuartiles.

Tabla 2: *Valores reales y estimados para Q_i.*

Procedimiento	Q_1	Q_2	Q_3
Estimado – MAS	230	360	520
Estimado – MET	215	255	370
Poblacional	215	245	395