ESTADÍSTICA ESPAÑOLA
Vol. 46, Núm. 156 2004, págs. 189 a 204

Estimación predictiva del índice de Gini en poblaciones finitas bajo muestreo aleatorio simple

por
JOSÉ A. MAYOR GALLEGO
Dpto. de Estadística e Investigación Operativa. Facultad de Matemáticas
Universidad de Sevilla

RESUMEN

El estudio de características de concentración, en el ámbito del muestreo en poblaciones finitas, permite cuantificar el grado de uniformidad o equidad en el reparto de una variable sobre los elementos de la población. Ello resulta especialmente interesante en investigaciones de carácter demográfico o económico, al estudiar variable como número de habitantes, salario o renta por persona. En este trabajo construimos estimadores del índice de Gini, empleando un enfoque predictivo basado en un modelo de superpoblación muy general y común, bajo diseño muestral aleatorio simple. Para estos estimadores realizamos mediante simulación un estudio comparativo con otros estimadores basados en el diseño. Dicho estudio muestra mejoras sustanciales en relación al sesgo y al error cuadrático medio.

Palabras clave: Muestreo, Poblaciones Finitas, Concentración, Gini.

Clasificación AMS: 62D05

1. INTRODUCCIÓN

Un aspecto de gran interés cuando se estudia una variable cuantitativa sobre una población, es la evaluación del grado de equidad en la distribución o reparto del total de la misma sobre los elementos poblacionales. El conocido índice de Gini es un parámetro, catalogado como de concentración, muy empleado en este tipo de investigaciones, siendo su estimación en el contexto de poblaciones finitas un problema relevante, tanto en sus aspectos teóricos como por sus aplicaciones prácticas.

Dicho problema ha sido tratado por varios autores, fundamentalmente bajo un enfoque de población fija, basado en el diseño muestral. Entre ellos citaremos a Brewer (1981) que considera la estimación del índice de Gini en un estudio de caso real sobre un problema de recursos escolares. Nygård y Sandström (1985a,1985b) proponen estimadores para este y otros parámetros de concentración, y Sandström, Wretman, y Waldén (1985,1988) estudian varios métodos para estimar la varianza de las estimaciones.

En el presente trabajo, consideramos este problema de estimación bajo un enfoque predictivo. Genéricamente, este enfoque se basa en la estimación de un parámetro poblacional que depende de una variable de estudio, a partir de una estimación o predicción de dicha variable.

Esta predicción se realiza, usualmente, a partir de un modelo que relaciona dicha variable de estudio con una o varias variables auxiliares sobre las que se tiene información previa, lo que se conoce como modelo de superpoblación.

Así, en primer lugar, en la sección 2, estudiamos este parámetro en relación a las funciones de distribución y concentración asociadas a la variable objeto de nuestro estudio, lo que nos permite construir un estimador básico de tipo predictivo, bajo la hipótesis de un modelo de superpoblación muy común y general.

En la sección 3, partiendo de la estimación de los parámetros del modelo mediante un ajuste mínimo cuadrático, realizamos un estudio del sesgo del estimador anteriormente obtenido, lo que nos permite realizar una corrección del mismo obteniendo un nuevo estimador predictivo alternativo.

En la sección 4 realizamos, mediante simulación, una comparación de los estimadores obtenidos con los estimadores tradicionales, basados en el diseño muestral, en lo que respecta al sesgo y al error cuadrático medio. Emplearemos para ello dos poblaciones reales y clásicas en la bibliografía de la teoría del muestreo.

Finalmente, en la sección 5, proponemos una estimación de la varianza basada en un método de exploración intensiva de la muestra, en concreto el jackknife. Mediante simulación, estudiamos el nivel de cubrimiento de los intervalos de confianza basados en dicha estimación.

2. ÍNDICE DE GINI. ESTIMACIÓN PREDICTIVA

Consideremos una población finita, U = {1,2,…,N}, y una variable de estudio, Y, con valores sobre la población { Y_i | i ∈ U}. Nuestro objetivo es la estimación, a partir de una muestra, del parámetro poblacional índice de Gini, correspondiente a dicha variable,

$$I_G = \frac{1}{2N^2 \overline{Y}} \sum_{i,j \in U} \sum \left| Y_i - Y_j \right|$$

donde $\overline{Y}$ = T(Y)/N denota la media poblacional de Y, es decir, T(Y)= $\sum_{i \in U} Y_i$

Sabe que este parámetro verifica $I_G \in [0,1-1/N]$ y está fuertemente relacionado con la función de distribución de la variable Y,

$$F(t) = \frac{1}{N} \sum_{i \in U} \Delta(t - Y_i)$$

y con la que denominaremos función de concentración,

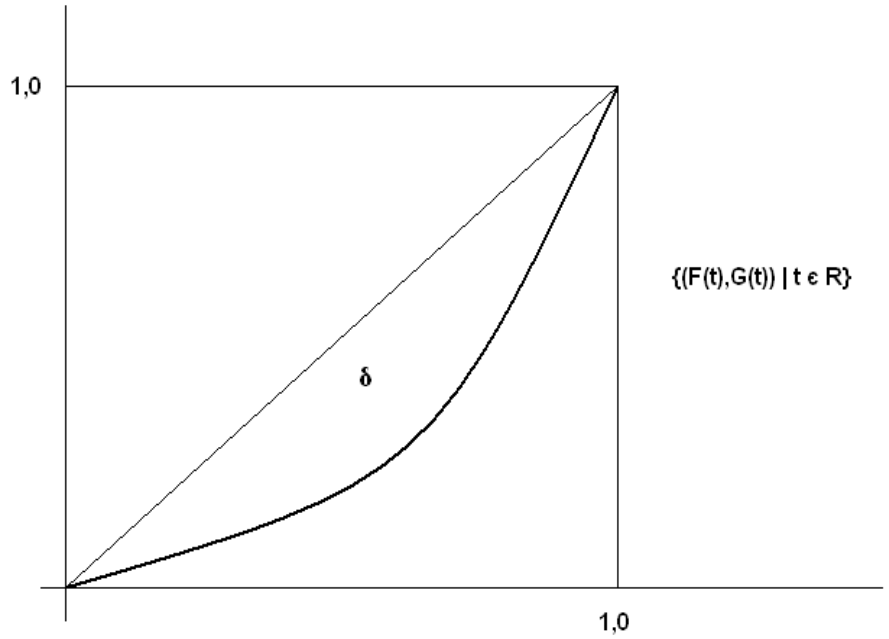
$$G(t) = \frac{1}{T(Y)} \sum_{i \in U} Y_i \Delta(t - Y_i)$$

siendo Δ $(t-Y_i)$ la función indicadora del intervalo $[Y_i, +\infty)$, es decir,

$$\Delta(t - Y_i) = \begin{cases} 1 & si \quad t \geq Y_i \\ 0 & si \quad t < Y_i \end{cases}$$

De hecho, como se indica en la Figura 1., se verifica IG = 2δ, siendo δ el área limitada, dentro del cuadrado unidad, por la bisectriz del primer cuadrante y la curva de Lorenz, definida como la poligonal que conecta los puntos {(F (t),G (t)) | t ∈ IR },

Figura 1
CURVA DE LORENZ. RELACIÓN CON EL ÍNDICE DE GINI



Con objeto de estimar el parámetro IG, se extrae de la población U una muestra m, de tamaño n, empleando un diseño muestral con probabilidades de inclusión de primer y segundo orden asociadas $\Pi = \{\pi_{ij} \setminus i, j \in U\}$, $\pi_{ij} > 0, \quad \forall i, j \in U$

Como ya hemos mencionado anteriormente, podemos encontrar en la bibliografía especializada distintas formas de realizar dicha estimación, que se basan fundamentalmente en el diseño muestral. Así, podemos considerar las estimaciones tipo Hayeck de las funciones de distribución y concentración, F(t) y G(t),

$$\hat{F}(t) = \frac{1}{\hat{N}} \sum_{i \in m} \Delta(t - Y_i) / \pi_i$$

$$\hat{G}(t) = \frac{1}{\hat{T}(Y)} \sum_{i \in m} Y_i \Delta(t - Y_i) / \pi_i$$

siendo $\hat{N}$ y $\hat{T}(Y)$ y $\hat{T}(Y)$ las correspondientes estimaciones de Horvitz-Thompson de N y T(Y), es decir, $\hat{N} = \sum_{i \in m} 1/\pi_i$ y $\hat{T}(Y) = \sum_{i \in m} Y_i / \pi_i$.

Podemos entonces construir la curva de Lorenz estimada, es decir, la poligonal que conecta los puntos $\{(F(t), G(t)) \mid t \in R\}$. Así, si aplicamos la interpretación geométrica del índice de Gini, como el doble de la superficie limitada por la poligonal y la bisectriz, y expresamos la muestra como m = $\{j_1, j_2, \ldots, j_n\}$, siendo j_1, $j_2, \ldots, j_n \in m$, los índices de las unidades muestrales según el orden no decreciente de los valores de la variable de estudio, es decir,

$$Y_{j_1} \leq Y_{j_2} \leq \cdots \leq Y_{j_n}$$

se obtiene mediante un cálculo directo el siguiente estimador,

$$\hat{I}_G = \frac{1}{\hat{N}^2 \hat{\bar{Y}}} \sum_{i=1}^{n} \left(2P_i + \frac{1}{\pi_{j_i}} \right) \frac{Y_{j_i}}{\pi_{j_i}} - 1$$

donde $\hat{\bar{Y}}$ es el estimador tipo Hayeck de $\bar{Y}$, es decir, $\hat{\bar{Y}} = \hat{T}(Y) / \hat{N}$, siendo,

$$P_1 = 0, \quad P_i = \sum_{k=1}^{i-1} 1/\pi_{j_k} \quad i = 2 \ldots n$$

Nygård y Sandström (1985a,1985b), proponen este mismo estimador del índice de Gini, aunque obteniéndolo por un procedimiento diferente, y lo estudian básicamente para el caso de muestreo aleatorio simple. Si lo particularizamos para dicho diseño muestral, se tiene $\pi_i = n/N$, y obtendremos mediante un cálculo simple,

$$\hat{I}_{G,d1} = \frac{1}{2n^2 \bar{y}} \sum_{i,j \in m} \left| Y_i - Y_j \right|$$

donde $\bar{y}$ denota la media muestral.

Los autores citados estudian el comportamiento de este estimador mediante simulación. Dichos autores destacan la presencia de sesgo así como la aplicabili-

dad del jackknife para estimar el error de muestreo y construir intervalos de confianza.

Podemos obtener un estimador alternativo, también exclusivamente basado en el diseño, observando que,

$$\sum_{i,j\in m}\frac{\left|Y_i - Y_j\right|}{\pi_{ij}}$$

es un estimador insesgado de,

$$\sum_{i,j\in U}\left|Y_i - Y_j\right|$$

por lo que podemos construir el estimador alternativo,

$$\hat{I}_{G,d2} = \frac{1}{2N^2\hat{\bar{Y}}}\sum_{i,j\in m}\frac{\left|Y_i - Y_j\right|}{\pi_{ij}}$$

que para el diseño muestral aleatorio simple, $\pi_{ij} = n(n-1)/N(N-1)$, resulta ser,

$$\hat{I}_{G,d2} = \frac{N-1}{2n(n-1)N\bar{y}}\sum_{i,j\in m}\left|Y_i - Y_j\right|$$

Seguidamente construiremos un estimador de tipo predictivo para el índice de Gini, a partir de su relación con la función de distribución poblacional. Para ello, expresamos dicho parámetro en la forma,

$$I_G = \frac{D}{2\bar{\bar{Y}}}$$

siendo,

$$D = \frac{1}{N^2}\sum_{i,j\in U}\left|Y_i - Y_j\right|$$

Si empleamos la función de distribución poblacional de la variable *Y* como integrador, podemos expresar la cantidad *D* como,

$$D = \int_R \int_R |u - w| dF(u) dF(v)$$

siendo además,

$$\overline{Y} = \int_R u \, dF(u)$$

La relación entre I_G y F(t) nos sugiere estimar el índice de Gini a partir de una estimación apropiada de F(t). Para ello, emplearemos un enfoque predictivo basado en un modelo de superpoblación muy usual, en concreto,

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \alpha, \beta \in R, \beta \neq 0$$

siendo $\varepsilon_i, i \in U$ variables aleatorias independientes y verificándose para la esperanza y varianza en el modelo de superpoblación $E_s[\varepsilon_i] = 0, V_s[\varepsilon_i] = \sigma_i^2$. En este trabajo, supondremos que las cantidades $\sigma_i, i \in U$ son desconocidas.

A partir de la muestra m, y siguiendo el enfoque predictivo usual, se estiman los parámetros del modelo, α y β, obteniéndose respectivamente $\hat{\alpha}_n$ *y* β_n Si ahora definimos la variable Y estimada como,

$$\hat{Y}_i = \begin{cases} Y_i & si \quad i \in m \\ \hat{\alpha}_n + \hat{\beta}_n X_i & si \quad i \in U - m \end{cases}$$

podemos estimar F(t) mediante el siguiente predictor,

$$\hat{F}(t) = \frac{1}{N} \sum_{i \in U} \Delta(t - \hat{Y}_i)$$

Dicho predictor es una función escalonada con los puntos de salto en el conjunto $\{Y_i \mid i \in U\}$, y además conserva las propiedades básicas del parámetro funcional que estima, es decir, monotonía, $F(-\infty) = 0$ y $F(+\infty) = 1$. Podemos pues estimar,

$$\hat{D} = \int_R \int_R |u - w| d\hat{F}(u) d\hat{F}(v) = \frac{1}{N^2} \sum_{i,j \in U} \left| \hat{Y}_i - \hat{Y}_j \right|$$

y

$$\hat{\bar{Y}} = \int_R u \, d\hat{F}(u) = \frac{1}{N} \sum_{i \in U} \hat{Y}_i$$

lo que nos permite construir el siguiente estimador predictivo del índice de Gini,

$$\hat{I}_{G,p} = \frac{\hat{D}}{2\hat{\bar{Y}}} = \frac{1}{2NT(\hat{Y})} \sum_{i,j \in U} \left| \hat{Y}_i - \hat{Y}_j \right|$$

donde hemos denotado $T(\hat{Y}) = \sum_{i \in U} \hat{Y}_i$.

Observemos que las propiedades y eficiencia de este estimador dependerán básicamente del diseño muestral empleado y de la metodología que se aplique para estimar los parámetros del modelo. En este trabajo, emplearemos la metodología de mínimos cuadrados con objeto de estimar dichos parámetros α y β .

Supondremos que la muestra *m* se obtiene mediante un diseño muestral aleatorio simple, con tamaño de muestra *n*. Se tiene pues para las probabilidades de inclusión de primer y segundo orden, $\pi_i = n/N, \forall i \in U$, y $\pi_{ij} = n(n-1)/N(N-1), \forall i \neq j \in U$.

Para construir pues las estimaciones de α y β se plantea el problema de minimización,

$$\min_{\alpha,\beta} \sum_{i \in U} (Y_i - \alpha - \beta X_i)^2$$

cuya solución teórica, que denotaremos α_N y β_N , es inviable al ser dicho problema poblacional. Siguiendo la metodología usual en la teoría del muestreo, consideraremos en su lugar el "problema estimado",

$$\min_{\alpha,\beta} \sum_{i \in U} (Y_i - \alpha - \beta X_i)^2 / \pi_i$$

Teniendo en cuenta la uniformidad de las probabilidades de inclusión de primer orden, dicho problema es equivalente a,

$$\min_{\alpha,\beta} \sum_{i \in m} (Y_i - \alpha - \beta X_i)^2$$

cuya solución es como sabemos,

$$\hat{\beta}_n = \frac{\sum_{i\in m}(Y_i - \overline{y})(X_i - \overline{x})}{\sum_{i\in m}(X_i - \overline{x})^2}$$

$$\hat{\alpha}_n = \overline{y} - \hat{\beta}_n \overline{x}$$

siendo $\overline{y}$ y $\overline{x}$ las medias muestrales respectivas de las variables Y y X .

3. ESTUDIO Y CORRECCIÓN DEL SESGO

Con objeto de estudiar el sesgo del estimador $\hat{I}_{G,p}$, obtenido en la sección anterior,

$$\hat{I}_{G,p} = \frac{\hat{D}}{2\hat{\overline{Y}}}$$

y teniendo en cuenta que $\hat{\overline{Y}}$ es realmente el estimador de regresión común, y que por consiguiente es asintóticamente insesgado, Sukhatme et al. (1984), utilizaremos la siguiente aproximación del sesgo de $I_{G,p}$,

$$B\left[\hat{I}_{G,p}\right] \approx \frac{E\left[\hat{D} - D\right]}{2\overline{Y}} = \frac{B\left[\hat{D}\right]}{2\overline{Y}}$$

Con objeto de hallar $E\left[\hat{D} - D\right]$, emplearemos las variables aleatorias indicadoras, $I_i = 1$ si $i \in m$, y cero en caso contrario. Dada entonces una diferencia genérica con $i \neq j$, podemos expresarla en la forma,

$$\left|\hat{Y}_i - \hat{Y}_j\right| = \left|Y_i - Y_j\right| I_i I_j + \left|Y_i - \hat{Y}_j\right| I_i (1 - I_j) + \left|\hat{Y}_i - Y_j\right| (1 - I_i) I_j + \left|\hat{Y}_i - \hat{Y}_j\right| (1 - I_i)(1 - I_j)$$

$$\approx \left|Y_i - Y_j\right| I_i I_j + \left|Y_i - \alpha_N - \beta_N X_j\right| I_i (1 - I_j)$$

$$+ \left|\alpha_N + \beta_N X_i - Y_j\right| (1 - I_i) I_j + \left|\beta_N X_i - \beta_N X_j\right| (1 - I_i)(1 - I_j)$$

Donde hemos aproximado $\hat{\alpha}_n$ y $\hat{\beta}_n$ por sus correspondientes versiones poblacionales, α_N y β_N respectivamente. Se tiene pues,

$$E\left[\hat{D}\right]=\frac{1}{N^2}E\left[\sum_{i,j\in U}\left|\hat{Y}_i-\hat{Y}_j\right|\right]$$

$$\approx\frac{1}{N^2}\left[\sum_{i,j\in U}\left|Y_i-Y_j\right|\pi_{ij}+\sum_{i,j\in U}\left|Y_i-\alpha_N-\beta_N X_j\right|(\pi_i-\pi_{ij})\right.$$

$$\left.+\sum_{i,j\in U}\left|\alpha_N+\beta_N X_i-Y_j\right|(\pi_j-\pi_{ij})+\beta_N\sum_{i,j\in U}\left|X_i-X_j\right|(1-\pi_i-\pi_j+\pi_{ij})\right]$$

y sustituyendo en esta expresión las cantidades poblacionales por estimaciones muestrales, podemos dar la siguiente estimación del sesgo de $\hat{I}_{G,\,p}$,

$$\hat{B}\left[\hat{I}_{G,p}\right]=\frac{1}{2\bar{\hat{Y}}N^2}\left[\sum_{i,j\in m}\left|Y_i-Y_j\right|\ \frac{\pi_{ij}-1}{\pi_{ij}}+\sum_{i,j\in m}\left|Y_i-\hat{\alpha}_N-\hat{\beta}_N X_j\right|\ \frac{\pi_i-\pi_{ij}}{\pi_{ij}}\right.$$

$$\left.+\sum_{i,j\in m}\left|\hat{\alpha}_N+\hat{\beta}_N X_i-Y_j\right|\ \frac{\pi_j-\pi_{ij}}{\pi_{ij}}+\hat{\beta}_N\sum_{i,j\in m}\left|X_i-X_j\right|\ \frac{1-\pi_i-\pi_j+\pi_{ij}}{\pi_{ij}}\right]$$

a partir de la cual, y bajo muestreo aleatorio simple, podemos construir el siguiente estimador predictivo con corrección del sesgo,

$$\hat{I}_{G,pc}=\hat{I}_{G,p}-\hat{B}\left[\hat{I}_{G,p}\right]$$

donde π_i y π_{ij} se sustituyen por las correspondientes probabilidades de inclusión asociadas a este tipo de muestreo.

4. EVALUACIÓN MEDIANTE SIMULACIÓN

Se ha realizado una comparación mediante simulación numérica, de los estimadores predictivos básico y corregido, $\hat{I}_{G,p}$ y $\hat{I}_{G,pc}$, con los estimadores $\hat{I}_{G,d1}$ y $\hat{I}_{G,d2}$, basados exclusivamente en el diseño muestral. Para ello se ha empleado muestreo aleatorio simple, simulando la extracción de L=1000 muestras para cada uno de los tamaños muestrales n = 10, 15, 20 y 30.

La comparación se ha realizado en relación al sesgo y al error cuadrático medio de los estimadores. Así, denotando genéricamente por $\hat{I}_G^i$ la estimación obtenida en la simulación *i*-ésima, hemos calculado las cantidades,

$$SESGO = 10^4 \times \frac{1}{L \times I_G} \sum_{i=1}^{L} (I_G - \hat{I}_G^i)$$

$$\sqrt{ECM} = 10^4 \times \left(\frac{1}{L \times I_G^2} \sum_{i=1}^{L} (I_G - \hat{I}_G^i)^2 \right)^{1/2}$$

Obsérvese que dichas cantidades están definidas en términos relativos, y además se han multiplicado por 10^4 para facilitar la comparación.

Se han empleado dos poblaciones reales. Por una parte, MU284, descrita por Särndal et al. (1992), y que consta de 284 municipios de Suecia, con variables de tipo demográfico. En concreto se ha tomado la variable P85, que indica el número de habitantes, en miles, de cada municipio en 1985, como variable de estudio Y, y el número de habitantes en 1975, P75, como variable auxiliar X.

Para estas variables, los parámetros del modelo $Y_i = \alpha + \beta X_i + \varepsilon_i$, ajustados mediante mínimos cuadrados, se exponen a continuación, siendo r^2 el coeficiente de determinación,

α_N	β_N	r^2
1.315	0.974	0.997

La segunda población utilizada para nuestra comparación, clásica en la bibliografía de la teoría del muestreo sobre estimación de la función de distribución, se denomina SUGAR CANE y es descrita por Chambers y Dunstan (1986). Consta de 338 plantaciones de caña de azúcar para cada una de las cuales se recogen diversas variables relacionadas con la producción. Hemos empleado la variable PRODUCCIÓN como variable de estudio, y la variable ÁREA como auxiliar. Para estas variables, los parámetros del modelo $Y_i = \alpha + \beta X_i + \varepsilon_i$, ajustados mediante mínimos cuadrados, son,

α_N	β_N	r^2
3439.821	1535.732	0.787

Los resultados obtenidos a partir de la simulación son expuestos en las tablas 1. y 2. que aparecen a continuación.

Tabla 1

RESULTADOS COMPARATIVOS PARA LA POBLACION MU284, SIENDO LAS VARIABLES DE ESTUDIO Y AUXILIAR RESPECTIVAMENTE Y: **P85**, X: **P75**

POBLACIÓN MU284. N=284								
	$\hat{I}_{G,p}$		$\hat{I}_{G,pc}$		$\hat{I}_{G,d1}$		$\hat{I}_{G,d2}$	
n	*SESGO*	$\sqrt{ECM}$	*SESGO*	$\sqrt{ECM}$	*SESGO*	$\sqrt{ECM}$	*SESGO*	$\sqrt{ECM}$
10	14	435	-68	359	2002	3021	1145	2785
15	12	364	-63	276	1480	2648	904	2513
20	33	348	-49	243	1159	2366	727	2212
30	47	343	-40	200	842	1913	560	1857

Tabla 2

RESULTADOS COMPARATIVOS PARA LA POBLACION SUGAR CANE, SIENDO LAS VARIABLES DE ESTUDIO Y AUXILIAR RESPECTIVAMENTE Y: **PRODUCCION**,, X: **ÁREA**

POBLACIÓN SUGAR CANE. N=338								
	$\hat{I}_{G,p}$		$\hat{I}_{G,pc}$		$\hat{I}_{G,d1}$		$\hat{I}_{G,d2}$	
N	SESGO	$\sqrt{ECM}$	SESGO	$\sqrt{ECM}$	SESGO	$\sqrt{ECM}$	SESGO	$\sqrt{ECM}$
10	669	670	-93	1296	1247	2545	348	2521
15	660	662	-79	1001	787	2064	145	2055
20	648	651	-61	834	634	1772	171	1772
30	628	632	-52	660	457	1414	158	1404

Como puede verse, para ambas poblaciones los estimadores predictivos, tanto básico como corregido, mejoran a los estimadores exclusivamente basados en el diseño, en lo que se refiere al error cuadrático medio.

Para la población MU284, el alto grado de correlación entre las variables también produce una mejora sustancial del estimador predictivo básico, en términos

de sesgo relativo, prácticamente despreciable, lo que no sucede para el caso de la población SUGAR CANE, con menor grado de correlación entre las variables, y para la cual el sesgo del estimador predictivo básico es considerable.

No obstante, este sesgo se reduce drásticamente para el estimador predictivo corregido, pudiéndose considerar despreciable. Por contra, el error cuadrático medio se incrementa para muestras pequeñas, pero este incremento se reduce aumentando el tamaño muestral. En cualquier caso, este error cuadrático medio se mantiene muy por debajo del que presentan los estimadores $\hat{I}_{G,d1}$ y $\hat{I}_{G,d2}$.

En resumen, los resultados de la simulación numérica muestran una mejora sustancial en términos de sesgo y error cuadrático medio del estimador predictivo corregido, $\hat{I}_{G,pc}$, incluso en situaciones para las cuales el grado de correlación entre las variables no es muy elevado.

5. ESTIMACIÓN DE LA VARIANZA

Con objeto de cuantificar el error de muestreo del estimador predictivo, vamos a aplicar un método de exploración intensiva de la muestra, en concreto, el procedimiento basado en la técnica del jackknife. Como es conocido, esta técnica se introdujo inicialmente en el campo de la estadística matemática como una forma de reducir el sesgo de las estimaciones. Posteriormente su ámbito de aplicación se ha ampliado también al muestreo en poblaciones finitas, como método para estimar la varianza de los estimadores. Véanse Wolter (1985) y Fernández y Mayor (1995).

Para el problema que estudiamos, Brewer (1981), Nygård y Sandström (1985a) y Sandström et al. (1985,1988), sugieren emplear la metodología del jackknife para estimar la varianza. Esta opción se justifica por la enorme complejidad que presenta el estudio del error por técnicas convencionales, debido a la estructura de los estimadores predictivos que hemos construido.

Siguiendo la descripción de Wolter (1985), dado un estimador genérico $\hat{I}_G$, consideremos la muestra obtenida,

$$m = \{j_1, j_2, \ldots, j_n\}$$

Denotemos por $\hat{I}_G(m)$ la estimación a partir de la muestra *m*, y por $\hat{I}_G(m - \{j_i\})$ la estimación a partir de la muestra *m* en la que se omite la unidad muestral *i*-ésima. Se definen los pseudovalores,

$$\hat{I}_G^{(i)} = n\hat{I}_G(m) - (n-1)\hat{I}_G(m - \{j_i\}) \quad i = 1 \ldots n$$

a partir de los cuales se tiene el siguiente estimador de la varianza,

$$\hat{V}\left[\hat{I}_G\right] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{I}_G(m) - \hat{I}_G^{(i)}\right)^2$$

Para evaluar empíricamente el comportamiento de este estimador de la varianza, lo hemos aplicado al estimador predictivo del índice de Gini con corrección del sesgo, $\hat{I}_{G,pc}$, realizando 1000 simulaciones sobre la población SUGAR CANE, construyendo para cada una el intervalo de confianza,

$$\hat{I}_{G,pc} \pm \delta \qquad \text{siendo} \quad \delta = 1.96 \times \sqrt{\hat{V}\left[\hat{I}_{G,pc}\right]}$$

Para diferentes tamaños de muestra, se ha calculado la proporción de CUBRIMIENTO, es decir, la proporción de aquellos intervalos construidos que contienen al valor exacto del parámetro. También se ha calculado el promedio de las cantidades,

$$\delta = 1.96 \times \sqrt{\hat{V}\left[\hat{I}_{G,pc}\right]}$$

esto es, el RADIO MEDIO de los intervalos. Los resultados obtenidos se muestran a continuación en la tabla 3.

Tabla 3

PROPORCION DE CUBRIMIENTO Y AMPLITUD DE INTERVALOS DE CONFIANZA CONSTRUÍDOS MEDIANTE JACKKNIFE. POBLACIÓN SUGAR CANE, SIENDO LAS VARIABLES DE ESTUDIO Y AUXILIAR RESPECTIVAMENTE Y: **PRODUCCIÓN, X: ÁREA**

POBLACIÓN SUGAR CANE. N=338		
n	*CUBRIMIENTO*	*RADIO MEDIO*
10	0.930000	0.086168
15	0.940000	0.065887
20	0.942000	0.054099
30	0.956000	0.043118

La magnitud y evolución del radio medio que aparecen en estos resultados indican un adecuado grado de precisión de los intervalos. Por otra parte se mani-

fiesta una elevada concordancia entre la proporción de cubrimiento obtenida y el nivel de confianza teórico del *95 %* empleado para construir los intervalos.

REFERENCIAS

BREWER, K.R.W. (1981). «The Analytical Use of Unequal Probability Samples: A case Study». Proceeding of the 43th Session of the International Statistical Institute. Buenos Aires.

CHAMBERS, R.L. Y DUNSTAN, R. (1986). «Estimating distribution functions from survey data». *Biometrika*. 73, 597-604.

FERNÁNDEZ, F.R. Y MAYOR GALLEGO, J.A. (1995). «Muestreo en Poblaciones Finitas*»:* Curso Básico. Ediciones Universitarias de Barcelona (E.U.B.). Barcelona.

NYGÅRD, F. Y SANDSTRÖM, A. (1985A). «Income Inequality Measures Based on Sample Surveys». Proceeding of the 45th Session of the International Statistical Institute. Amsterdam.

NYGÅRD, F. Y SANDSTRÖM, A. (1985b). «The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations» Journal of Official Statistics. **1**, 399-412.

SÄRNDAL, C., SWENSSON, B. Y WRETMAN, J. (1992). «Model Assisted Survey Sampling». Springer-Verlag. New York, Inc.

SANDSTRÖM, A., WRETMAN, J.H. Y WALDÉN, B. (1985). «Variance Estimators of the Gini Coefficient, Simple Random Sampling». Metron. 43, 41-70.

SANDSTRÖM, A., WRETMAN, J.H. Y WALDÉN, B. (1988). «Variance Estimators of the Gini Coefficient-Probability Sampling». Journal of Business & Economic Statistics. 6-1, 113-119.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. Y ASOK, C. (1984). «Sampling Theory of Surveys Applications». Tercera edición. Iowa State University Press. Ames. Iowa.

WOLTER, K.M. (1985). «Introduction to Variance Estimation. Springer-Verlag». New York, Inc.

ESTIMATING THE GINI COEFFICIENT OVER A FINITE POPULATION:A PREDICTIVE APPROACH UNDER SIMPLE RANDOM SAMPLING

ABSTRACT

In this paper, a predictive estimator of the Gini coefficient (the well-known income inequality measure) of a finite population is defined for an arbitrary probability sampling design, taking a very common superpopulation model in consideration. Under simple random sampling, and by means of a Monte Carlo study, the sampling performance of this predictive estimator is compared with other classical design based estimators in literature.

Key words: Survey Sampling, Finite Populations, Gini Coefficient.

AMS Classification: 62D05