Estimación del Índice de Gini mediante Diseños Muestrales Complejos

José Antonio Mayor Gallego

Departamento de Estadística e Investigación Operativa

Universidad de Sevilla. Facultad de Matemáticas

Septiembre de 2007

Problema Considerado: Estimación del Índice de Gini en poblaciones finitas

Elementos básicos

- Población finita.

$$U = \{1, 2, \ldots, N\}$$

- Variable de estudio.

$$Y = \{y_i \mid i \in U\}$$

- Función de distribución poblacional,

$$F(t) = \frac{1}{N} \sum_{i \in U} \Delta(t - Y_i) \qquad \Delta(t - Y_i) = \begin{cases} 1 & t \geq y_i \\ 0 & t < y_i \end{cases}$$

- Media y total poblacionales,

$$\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i = \frac{1}{N} T(Y)$$

Problema Considerado: Estimación del Índice de Gini en poblaciones finitas

Índice de Gini

$$IG = \frac{\int_{\mathbb{R}} \int_{\mathbb{R}} |t-u| dF(t) dF(u)}{\int_{\mathbb{R}} u\, dF(u)}$$

Características

- Medida de uniformidad en el reparto de la variable en estudio.
- Útil en estudios económicos y demográficos sobre distribución de bienes, salarios, población, etc.
- Habitualmente se estudia a partir de encuestas por muestreo..

Objetivos

- Desarrollar un estimador del índice de Gini en poblaciones finitas, fácilmente adaptable a diseños muestrales complejos.
- Estudiar sus propiedades en relación al sesgo y al error cuadrático medio.
- Realizar mediante simulación un estudio comparativo con otros estimadores de la bibliografía.

Problema Considerado: Estimación del Índice de Gini en poblaciones finitas

Índice de Gini

$$IG = \frac{\int_{\mathbb{R}} \int_{\mathbb{R}} |t-u| dF(t) dF(u)}{\int_{\mathbb{R}} u\, dF(u)}$$

Características

- Medida de uniformidad en el reparto de la variable en estudio.
- Útil en estudios económicos y demográficos sobre distribución de bienes, salarios, población, etc.
- Habitualmente se estudia a partir de encuestas por muestreo..

Objetivos

- Desarrollar un estimador del índice de Gini en poblaciones finitas, fácilmente adaptable a diseños muestrales complejos.
- Estudiar sus propiedades en relación al sesgo y al error cuadrático medio.
- Realizar mediante simulación un estudio comparativo con otros estimadores de la bibliografía.

Problema Considerado: Estimación del Índice de Gini en poblaciones finitas

Índice de Gini

$$IG = \frac{\int_{\mathbb{R}} \int_{\mathbb{R}} |t-u| dF(t) dF(u)}{\int_{\mathbb{R}} u\, dF(u)}$$

Características

- Medida de uniformidad en el reparto de la variable en estudio.
- Útil en estudios económicos y demográficos sobre distribución de bienes, salarios, población, etc.
- Habitualmente se estudia a partir de encuestas por muestreo..

Objetivos

- Desarrollar un estimador del índice de Gini en poblaciones finitas, fácilmente adaptable a diseños muestrales complejos.
- Estudiar sus propiedades en relación al sesgo y al error cuadrático medio.
- Realizar mediante simulación un estudio comparativo con otros estimadores de la bibliografía.

Parámetros de concentración

Familia de índices de Gini.

▸ Nygård and Sandström (1985a,1985b)

$$IG_J = \frac{1}{\overline{Y}} \int_0^{\infty} J[F(t)]tdF(t)$$

$J(\cdot)$ es una función de ponderación, continua.

Índice de Gini clásico. $J(p) = 2p - 1$

$$IG = \frac{1}{\overline{Y}} \int_{\mathbb{R}} [2F(t) - 1]tdF(t) = \frac{1}{\overline{Y}} \int_{\mathbb{R}} \int_{\mathbb{R}} |t - u| dF(t)dF(u)$$

$$= \frac{1}{2N^2\overline{Y}} \sum_{i,j \in U} \sum |y_i - y_j| \in [0, 1 - 1/N]$$

Parámetros de concentración

Familia de índices de Gini.

▸ Nygård and Sandström (1985a,1985b)

$$IG_J = \frac{1}{\overline{Y}} \int_0^{\infty} J[F(t)]tdF(t)$$

$J(\cdot)$ es una función de ponderación, continua.

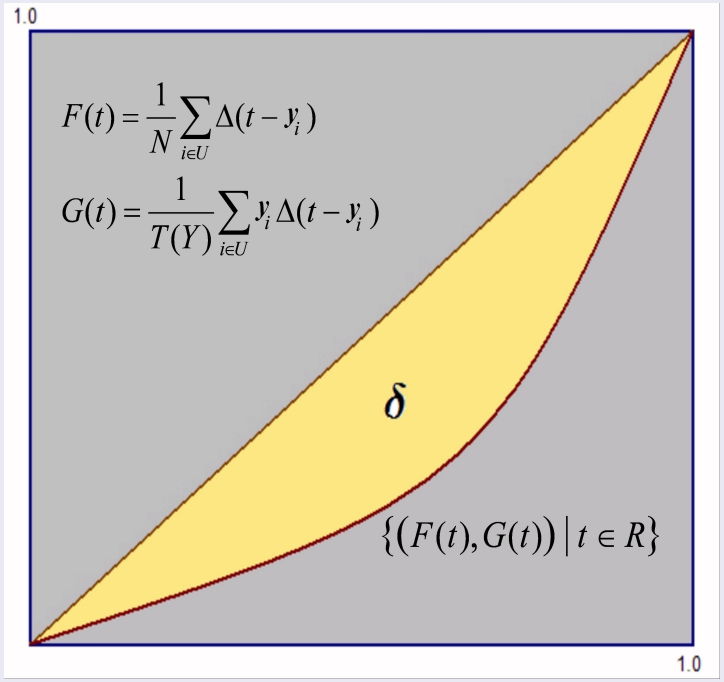
Índice de Gini clásico. $J(p) = 2p - 1$

$$IG = \frac{1}{\overline{Y}} \int_{\mathbb{R}} [2F(t) - 1]tdF(t) = \frac{1}{\overline{Y}} \int_{\mathbb{R}} \int_{\mathbb{R}} |t - u| dF(t)dF(u)$$

$$= \frac{\mathbf{1}}{\mathbf{2N^2\overline{Y}}} \sum_{\mathbf{i},\mathbf{j} \in \mathbf{U}} \sum |\mathbf{y_i} - \mathbf{y_j}| \in [0, 1 - 1/N]$$

Índice de Gini y curva de Lorenz

Curva de Lorenz



$$IG = \frac{1}{2N^2\overline{Y}}\sum_{i,j\in U}|y_i - y_j| = 2\,\delta$$

Estimación

- Diseño muestral $(M, p(\cdot)) \rightarrow m$, muestra.
- Probabilidades de inclusión,

$$\Pi = \{\pi_{ij} \mid i,j \in U\} > 0$$

- Estimaciones de $F(t)$ and $G(t)$,

$$\widehat{F}(t) = \frac{1}{\widehat{N}}\sum_{i\in m}\frac{\Delta(t-y_i)}{\pi_i}$$

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

- Estimación de la curva de Lorenz.

$$\{(\widehat{F}(t), \widehat{G}(t)) \mid t \in \mathbb{R}\}$$

Estimador de Nygård y Sandström

$$\widehat{IG} = 2\widehat{\delta} = \frac{1}{\widehat{N}^2\,\widehat{\overline{Y}}} \sum_{i=1}^{n} \left(2P_i + \frac{1}{\pi_{j_i}}\right) \frac{y_{j_i}}{\pi_{j_i}} \; - 1$$

- $\widehat{\overline{Y}} = \widehat{T}(Y)/\widehat{N}$. P_i dadas por,

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

- $j_1, j_2, \ldots j_n$ tales que, $y_{j_1} \leq y_{j_2} \leq \cdots \leq y_{j_n}$
- Para el diseño MAS(N, n), $\pi_i = n/N$,

$$\widehat{IG}_{\mathrm{NS,MAS}} = \frac{1}{2n^2\bar{y}} \sum_{i,j \in m} \sum |y_i - y_j|$$

▸ Nygård and Sandström (1985a,1985b)

Estimador alternativo

Expresión del índice de Gini

A partir de,

$$\sum\sum_{i\neq j\in U}(y_i + y_j) = 2N(N-1)\overline{Y}$$

obtenemos,

$$\begin{aligned} IG &= \frac{1}{2N^2}\frac{\sum\sum_{i,j\in U}|y_i - y_j|}{(\sum\sum_{i\neq j\in m}(y_i + y_j))/(2N(N-1)} \\ &= \frac{N-1}{N}\frac{\sum\sum_{i,j\in U}|y_i - y_j|}{\sum\sum_{i\neq j\in U}(y_i + y_j)} \end{aligned}$$

Estimador alternativo

- $\sum\sum_{i,j\in m} \frac{|y_i - y_j|}{\pi_{ij}}$ es un estimador insesgado de $\sum\sum_{i,j\in U} |y_i - y_j|$
- $\sum\sum_{i\neq j\in m} \frac{(y_i + y_j)}{\pi_{ij}}$ es un estimador insesgado de $\sum\sum_{i\neq j\in U} (y_i + y_j)$

$$\widehat{IG}_{\mathrm{M1}} = \frac{N-1}{N} \frac{\sum\sum_{i,j\in m} |y_i - y_j|/\pi_{ij}}{\sum\sum_{i\neq j\in m} (y_i + y_j)/\pi_{ij}}$$

$$= \frac{N-1}{N} \frac{\sum\sum_{i,j\in m} \omega_{ij} |y_i - y_j|}{\sum\sum_{i\neq j\in m} \omega_{ij} (y_i + y_j)}$$

Ponderaciones muestrales,

$$\omega_{ij} = \frac{1}{\pi_{ij}} \quad \forall i \neq j$$

Estimador $\widehat{IG}_{\mathrm{M1}}$. Observaciones

- Al ser, $|y_i - y_j| \leq (y_i + y_j)\ \forall i \neq j \in U$, para cualquier diseño muestral se tiene,

$$\widehat{IG}_{\mathrm{M1}} = \frac{N-1}{N} \frac{\sum\sum_{i,j\in m} |y_i - y_j|/\pi_{ij}}{\sum\sum_{i\neq j\in m}(y_i + y_j)/\pi_{ij}} \in [0, 1 - 1/N]$$

- Para el muestreo aleatorio simple, $\pi_{ij} = n(n-1)/N(N-1),\ i \neq j$, siendo,

$$\widehat{IG}_{\mathrm{M1,MAS}} = \frac{N-1}{2n(n-1)N\bar{y}} \sum_{i,j\in m}\sum |y_i - y_j|$$

- Podemos esperar una mayor precisión con respecto al estimador,

$$\widehat{IG}_{\mathrm{NS,MAS}} = \frac{1}{2n^2\bar{y}} \sum_{i,j\in m}\sum |y_i - y_j|$$

  al no ser este un cociente de estimadores insesgados.

- $\widehat{IG}_{\mathrm{M1}}$ puede ser adaptado fácilmente a diseños muestrales complejos basados en muestreo aleatorio simple, para los que las ponderaciones muestrales son fáciles de calcular

Variación del estimador alternativo

Estimando la media poblacional, $\overline{Y}$, del denominador del índice de Gini, directamente mediante el estimador de Horvitz-Thompson,

$$\widehat{\overline{Y}} = \frac{1}{N} \sum_{i \in m} \frac{y_i}{\pi_i}$$

obtenemos,

$$\widehat{IG}_{\mathrm{M2}} = \frac{1}{2N} \frac{\sum \sum_{i,j \in m} |y_i - y_j| / \pi_{ij}}{\sum_{i \in m} y_i / \pi_i}$$

que en general es distinto de $\widehat{IG}_{\mathrm{M1}}$, aunque para el caso de muestreo aleatorio simple ambos estimadores coinciden. También para este estimador podemos esperar una reducción del sesgo en relación al estimador clásico de Nygård and Sandström.

Estimador $\widehat{IG}_{M1}$. Aproximaciones

- $\theta_1 = (N-1)\sum\sum_{i,j\in U}|y_i - y_j|$
- $\widehat{\theta}_1 = (N-1)\sum\sum_{i,j\in m}|y_i - y_j|/\pi_{ij}$
- $\theta_2 = N\sum\sum_{i\neq j\in U}(y_i + y_j)$
- $\widehat{\theta}_2 = N\sum\sum_{i\neq j\in m}(y_i + y_j)/\pi_{ij}$

Aproximación lineal

$$\widehat{IG}_{M1} = \frac{\widehat{\theta}_1}{\widehat{\theta}_1} \approx IG + \frac{\widehat{\theta}_1 - IG\,\widehat{\theta}_2}{\theta_2}$$

Aproximación cuadrática

$$\widehat{IG}_{M1} = \frac{\widehat{\theta}_1}{\widehat{\theta}_1} \approx IG + \frac{\widehat{\theta}_1 - IG\,\widehat{\theta}_2}{\theta_2} - \frac{(\widehat{\theta}_1 - \theta_1)(\widehat{\theta}_2 - \theta_2)}{\theta_2^2} + \frac{\theta_1}{\theta_2^3}(\widehat{\theta}_2 - \theta_2)^2$$

Diseño muestral MAS(N, n). Resultados

- Termino principal del sesgo,

$$B_1[\widehat{IG}_{\mathrm{M1}}] = \frac{1}{\theta_2^2}\left(IG\, V[\widehat{\theta}_2] - \mathrm{Cov}[\widehat{\theta}_1, \widehat{\theta}_2]\right) = O(n^{-1})$$

- Varianza de la aproximación lineal,

$$V_1[\widehat{IG}_{\mathrm{M1}}] = \frac{1}{\theta_2^2}\left(V[\widehat{\theta}_1] + IG^2\, V[\widehat{\theta}_1] - 2\, IG\, \mathrm{Cov}[\widehat{\theta}_1, \widehat{\theta}_2]\right) = O(n^{-1})$$

- Error cuadrático medio,

$$\mathrm{ECM}_1[\widehat{IG}_{\mathrm{M1}}] = O(n^{-1})$$

Estimador $\widehat{IG}_{\mathrm{M1}}$. Estimador de la varianza

$$\begin{aligned}\widehat{V}[\widehat{IG}_{\mathrm{M1}}] &= \frac{1}{\widehat{\theta}_2^2}\Big(\sum\sum_{i\neq j\in m}\frac{z_{ij}^2}{\pi_{ij}^2}(1-\pi_{ij}) \\ &\quad +2\sum\sum_{\substack{i\neq j,\, i\neq k\in m\\ j\neq k}}\frac{z_{ij}}{\pi_{ij}}\frac{z_{ik}}{\pi_{ik}}\frac{\pi_{ijk}-\pi_{ij}\pi_{ik}}{\pi_{ijk}} \\ &\quad +\sum\sum_{\substack{i\neq j,\, k\neq l\in m\\ i\neq k,\, j\neq l}}\frac{z_{ij}}{\pi_{ij}}\frac{z_{kl}}{\pi_{kl}}\frac{\pi_{ijkl}-\pi_{ij}\pi_{kl}}{\pi_{ijkl}}\Big)\end{aligned}$$

- $z_{ij} = (N-1)|y_i - y_j| - N\,\widehat{IG}_{\mathrm{M1}}(y_i + y_j) \qquad i \neq j$
- $\widehat{\theta}_2 = N\sum\sum_{i\neq j\in m}(y_i + y_j)/\pi_{ij}$

Estudio comparativo. Simulación

Poblaciones.

▸ Chambers y Dunstan (1986), Särndal et al. (1992)

- SUGAR CANE. $N = 338$ plantaciones de caña de azucar. Chambers y Dunstan (1986). Y: PRODUCCIÓN.
- MU284. $N = 284$ municipios de Suecia. Särndal et al. (1992). Y: POBLACIÓN en 1985.

Simulación muestral

- Extraccion de $L = 1000$ muestras aleatorias simples. $n = 10, 15, 20, 25$ y 30.
- Se calculan,

$$\text{SESGO} = 10^4 \times \frac{1}{L \times IG} \sum_{1=1}^{L} (IG - \widehat{IG}_i)$$

$$\sqrt{\text{ECM}} = 10^4 \times \left(\frac{1}{L \times IG^2} \sum_{1=1}^{L} (IG - \widehat{IG}_i)^2 \right)^{1/2}$$

SUGAR CANE

POBLACIÓN SUGAR CANE. $N = 338$				
	$\widehat{IG}_{\mathrm{NS}}$		$\widehat{IG}_{\mathrm{M1-M2}}$	
n	SESGO	$\sqrt{\mathrm{ECM}}$	SESGO	$\sqrt{\mathrm{ECM}}$
10	2295	3660	387	3576
15	1168	2459	215	2407
20	572	1693	102	1676
25	484	1568	106	1554
30	375	1438	72	1433

Resultados comparativos para la población SUGAR CANE, siendo la variable de estudio Y: **PRODUCCIÓN**.

MU284

POBLACIÓN MU284. $N = 284$				
	$\widehat{IG}_{NS}$		$\widehat{IG}_{M1-M2}$	
n	SESGO	$\sqrt{ECM}$	SESGO	$\sqrt{ECM}$
10	1961	2998	1100	2741
15	1404	2578	823	2450
20	1182	2294	751	2195
25	850	2111	502	2068
30	831	2052	487	1948

Resultados comparativos para la población MU284, siendo la variable de estudio Y: **P85**.

Población.

▸ Särndal et al. (1992)

- MU284C. $N = 284$ municipios de Suecia clasificados en $M = 50$ conglomerados. Särndal et al. (1992). Y: POBLACIÓN en 1985.

Simulación muestral. Muestreo por conglomerados. Una etapa

- Extraccion de $L = 1000$ muestras aleatorias simples de conglomerados. $g = 4, 6, 8$ y 10 conglomerados.
- Se calculan,

$$\text{SESGO} = 10^4 \times \frac{1}{L \times IG} \sum_{1=1}^{L} (IG - \widehat{IG}_i)$$

$$\sqrt{\text{ECM}} = 10^4 \times \left(\frac{1}{L \times IG^2} \sum_{1=1}^{L} (IG - \widehat{IG}_i)^2 \right)^{1/2}$$

Resultados

MU284C

POBLACIÓN MU284C. $N = 284$. 50 CONGLOMERADOS						
	$\widehat{IG}_{\mathrm{NS}}$		$\widehat{IG}_{\mathrm{M1}}$		$\widehat{IG}_{\mathrm{M2}}$	
g	SESGO	$\sqrt{\mathrm{ECM}}$	SESGO	$\sqrt{\mathrm{ECM}}$	SESGO	$\sqrt{\mathrm{ECM}}$
4	1142	2352	511	2246	664	2285
6	732	1954	343	1900	416	1944
8	552	1679	257	1644	323	1673
10	420	1504	207	1498	235	1494

Resultados comparativos para la población MU284C, siendo la variable de estudio Y: **P85**

Conclusiones

- Los estimadores $\widehat{IG}_{\mathrm{M1}}$ y $\widehat{IG}_{\mathrm{M2}}$ presentan en general una notable reducción de sesgo en relación al estimador clásico $\widehat{IG}_{\mathrm{NS}}$. En el estudio realizado mediante un diseño muestral complejo como es el muestreo por conglomerados en una etapa, dicha reducción es más acentuada para $\widehat{IG}_{\mathrm{M1}}$.
- En relación al error cuadrático medio, todos los estimadores estudiados presentan resultados muy similares.

Líneas de trabajo

- Desarrollo de estimadores para otros parámetros de concentración, en el ámbito de los diseños muestrales complejos.
- Estudio de la viabilidad de las técnicas de “jackknife” para estimar el error de muestreo.

- **Nygård, F. and Sandström, A. (1985a)**. Income Inequality Measures Based on Sample Surveys. *Proceeding of the* 45^{th} *Session of the International Statistical Institute*. Amsterdam.
- **Nygård, F. and Sandström, A. (1985b)**. The Estimation of the Gini and the Entropy Inequality Parameters in Finite Populations. *Journal of Official Statistics.* **1**, 399-412.

- **Chambers, R.L. and Dunstan, R. (1986)**. Estimating distribution functions from survey data. *Biometrika.* **73**, 597-604.
- **Särndal, C., Swensson, B. and Wretman, J. (1992)**. *Model Assisted Survey Sampling.* Springer-Verlag. New York, Inc.