ESTADÍSTICA ESPAÑOLA
Vol. 39, Núm. 142, 1997, págs. 207 a 218

Una nota sobre el cálculo del índice de Gini

por
EVA FERREIRA(1)
Departamento.Economía Aplicada III
Universidad del País Vasco. Bilbao

ARACELI GARÍN(1)
Departamento Economía Aplicada III
Universidad del País Vasco. Bilbao

RESUMEN

En esta breve nota se establece la relación entre distintas fórmulas utilizadas para el cálculo del índice de Gini. Partiendo de la definición geométrica realizada en términos de la curva de Lorenz, proponemos obtener el valor exacto del área bajo dicha curva en lugar de las aproximaciones ofrecidas por gran parte de los libros básicos de estadística descriptiva. Para ello describiremos un método de cálculo sencillo que además permite una interpretación del índice en términos de la covarianza entre dos variables.

Palabras Clave: Concentración, Curva de Lorenz, Medidas de desigualdad.

Clasificación AMS: 6201

---

(1) Agradecemos los comentarios de dos evaluadores anónimos así como la financiación del proyecto PB95-0346 de la Dirección General de Enseñanza Superior del Ministerio de Educación y Ciencia junto con la Universidad del País Vasco.

1. INTRODUCCIÓN

El índice de Gini es muy utilizado en economía como medida del grado de concentración de variables como salarios o renta, entre otras. Existen formas alternativas de expresar y calcular este índice. De hecho, la formulación original de Gini no coincide con la definición que utilizaremos a lo largo del trabajo (ver Pyatt, 1976). Habitualmente, la definición del índice de Gini que se encuentra en los libros introductorios de estadística es geométrica, obteniéndose dicho índice como el cociente entre el área delimitada por la curva de Lorenz junto con la diagonal representativa de la equidad total y el área comprendida entre las representaciones respectivas de equidad y concentración total. A partir de esta fórmula se ofrecen aproximaciones supuestamente más simples que este cálculo geométrico (ver por ejemplo Escuder, 1982; Rodríguez y Arenales, 1988; Uriel y Muñiz, 1988; Martín Pliego, 1994).

En este artículo proponemos el uso del método exacto para el cálculo del área de concentración bajo la curva de Lorenz descriptiva, y a partir de ella, del índice de Gini definido en los términos anteriores. En la Sección 2 mostraremos que, reescribiendo dicha fórmula adecuadamente, el coste operativo de la obtención del valor exacto es igual o menor que el de cualquiera de las aproximaciones propuestas en los libros de texto. En la Sección 3 relacionaremos la definición del índice en colectivos discretos con la aproximación discreta del índice asociado a una variable continua. De esta forma demostraremos que el índice, en colectivos discretos, se puede calcular de forma exacta como una covarianza entre dos variables, al igual que sucede con su cálculo en forma continua. Esto permite su obtención directa usando los paquetes estadísticos habituales. En la Sección 4 mostraremos mediante algunos ejemplos las diferencias entre las aproximaciones y el valor exacto, así como una tabla de cálculo para facilitar el cómputo del índice. Finalmente extraeremos algunas conclusiones.

2. EL ÍNDICE DE GINI EN COLECTIVOS DISCRETOS

2.1. Definición y aproximaciones utilizadas

Consideremos un colectivo de tamaño N y una variable X objeto de análisis con la siguiente notación.

- $\{x_1 < x_2 < ... < x_k\}$: k valores distintos que toma la variable estadística X,
- $\{n_1, n_2, ..., n_k\}$: frecuencias absolutas asociadas,

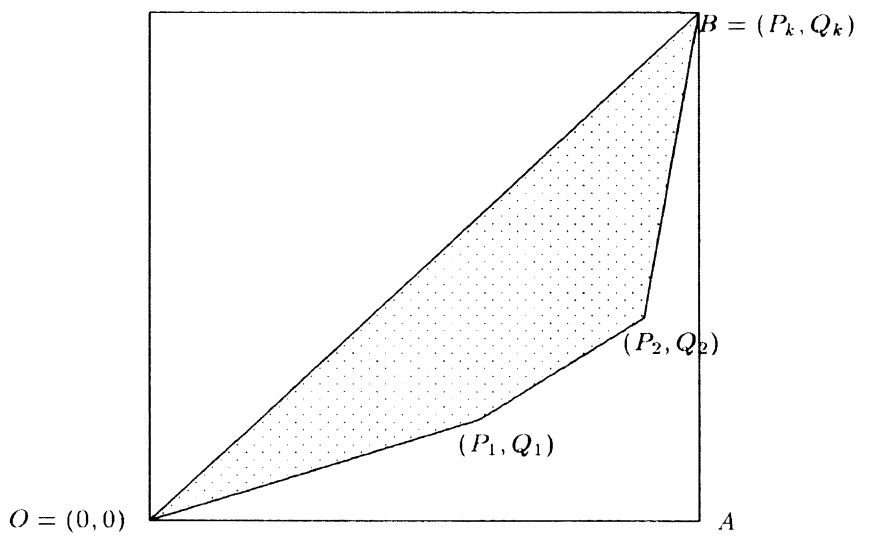
• $\{p_1, p_2, ..., p_k\}$: frecuencias relativas ($p_i = n_i / N$, con $N = \sum_{i=1}^{k} n_i$). Por tanto, $\sum_{i=1}^{k} x_i n_i$ resulta ser el total (masa salarial, renta, ...) que se reparte entre el colectivo,

• q_i: masa relativa repartida entre los miembros de la clase i-ésima; es decir, $q_i = (x_i n_i) / \sum_{j=1}^{k} (x_j n_j)$.

Una forma de visualizar el grado de uniformidad del reparto es mediante la representación de la curva de Lorenz. En ella, se representan los puntos (P_i, Q_i), donde $P_i = \sum_{j \le i} p_j$ y $Q_i = \sum_{j \le i} q_j$. Aunque no vamos a detallar la interpretación de esta curva, para ello remitimos a cualquier libro introductorio de estadística, si queremos apuntar que la diagonal principal del cuadrado ($[0,1] \times [0,1]$) representaría la equidad total, un reparto totalmente uniforme, mientras que una mayor cercanía a los ejes representa mayor concentración en el reparto.

FIGURA 1

Curva de Lorenz



Basado en la curva de Lorenz y en que la concentración se puede visualizar en términos del área que queda entre ella y la diagonal (área rayada), la definición habitual del índice de Gini en el campo de la Estadística descriptiva es la siguiente:

$$I_G = \frac{\text{área rayada}}{\text{área OAB}} \qquad [1]$$

donde OAB representa el triángulo bajo la diagonal principal. Sin embargo, son habituales las fórmulas aproximadas para la obtención del índice definido como (1).

Quizá una de las aproximaciones más utilizada es la que proporciona la siguiente fórmula:

$$I_{G1} = \frac{\sum_{i=1}^{k-1} (P_i - Q_i)}{\sum_{i=1}^{k-1} P_i} = 1 - \frac{\sum_{i=1}^{k-1} Q_i}{\sum_{i=1}^{k-1} P_i} \qquad [2]$$

que aparece en la mayoría de los libros de texto. (Escuder, 1982; Uriel y Muñiz, 1988; Escuder y Murgui, 1995; Martín Pliego, 1994; Montiel *et al.*, 1997, entre otros).

Otra fórmula utilizada, en general de mejor aproximación al valor (1), es:

$$I_{G2} = \frac{\sum_{i<j} (x_j - x_i) n_i n_j}{N(N-1)\bar{x}} \qquad [3]$$

(Ver Rodríguez y Arenales, 1988)(2) .

Ambas fórmulas se basan en una aproximación mediante rectángulos del área que se quiere calcular. Sin embargo, como veremos con más detalle a lo largo del trabajo, la fórmula (2) constituye una mala aproximación en colectivos con pocos valores diferentes (k pequeño) y la fórmula (3) depende del tamaño, N, del colectivo.

2.2. Método de cálculo exacto

A continuación detallaremos el método que permite obtener de forma general el valor exacto del valor del índice definido en (1).

Lema 1 Conservando la notación del apartado anterior, el valor exacto de la expresión (1) es:

$$I_G = 1 - 2\left(\frac{p_1 q_1}{2} + p_2\left(q_1 + \frac{q_2}{2}\right) + \ldots + p_k\left(q_1 + q_2 + \ldots + \frac{q_k}{2}\right)\right) =$$

(2) Esta expresión coincide con la definición original de Gini donde x_i son los valores de la v.a. X y $\bar{x}$ representa la media de la variable.

$$= 1 - \sum_{i=1}^{k} p_i \left[\sum_{j=1}^{i-1} 2q_j + q_i \right] \qquad [4]$$

Demostración:

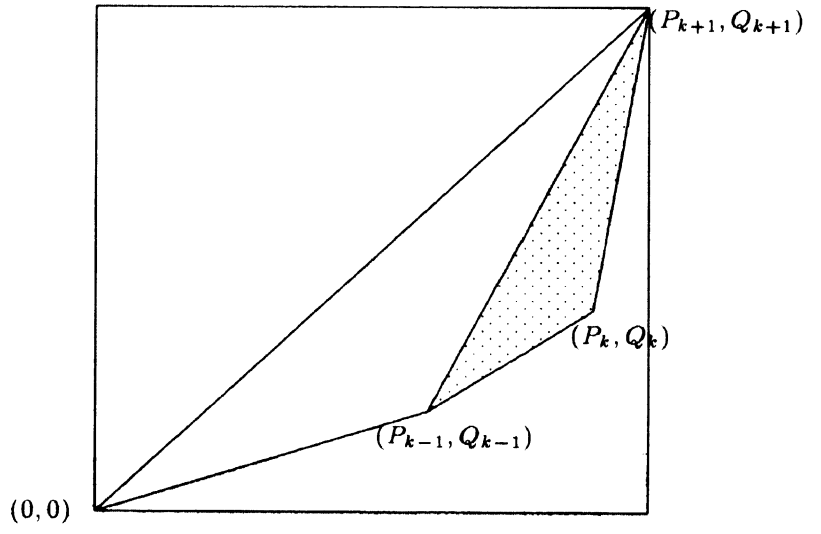
Usaremos el principio de inducción. Es claro de la representación gráfica (ver Figura 1) que para k=2 el valor exacto del índice es:

$$I_G = 1 - 2\left(\frac{p_1 q_1}{2} + p_2 \left(q_1 + \frac{q_2}{2} \right) \right) \qquad [5]$$

Supongamos ahora que el valor del índice para *k* valores distintos de X es:

$$I_G = 1 - 2\left(\frac{p_1 q_1}{2} + p_2 \left(q_1 + \frac{q_2}{2} \right) + \ldots + p_k \left(q_1 + q_2 + \ldots + \frac{q_k}{2} \right) \right) \qquad [6]$$

FIGURA 2



El hecho de dividir y añadir una nueva clase significa que en la representación gráfica el área aumenta en el trozo rayado, tal y como se indica en la Figura 2, y por

tanto el trozo añadido es igual al área del triángulo cuyos catetos miden $(p_k + p_{k+1})$ y $(q_k + q_{k+1})$ menos $(p_k\, q_k /2 + p_{k+1}\, q_k + p_{k+1}\, q_{k+1} /2)$ y en la fórmula final el área del rectángulo que queda por debajo se divide en dos bases, p_k y p_{k+1} con lo que el valor final del índice es:

$$I_G = 1 - 2\left(\frac{p_1 q_1}{2} + p_2 (q_1 + \frac{q_2}{2}) + \ldots + p_k (q_1 + q_2 + \ldots + q_{k-1}) + \right.$$

$$\left. + p_{k-1} (q_1 + q_2 + \ldots + q_{k-1}) + \frac{p_k q_k}{2} + p_{k-1} q_k + \frac{p_{k-1} q_{k+1}}{2} \right) \qquad [7]$$

que, reordenando términos resulta ser la fórmula expresada en (4).

Con el siguiente resultado veremos la similitud existente entre la fórmula de cálculo exacto y la expresión aproximada dada I_{G2}.

Lema 2 De la expresión obtenida en el Lema 1, y sustituyendo $p_i = n_i / N$ y $q_i = (n_i x_i)/(N \bar{x})$, se obtiene:

$$I_G = \frac{\sum_{j<i} (x_i - x_j) n_i n_j}{N^2 \bar{x}} = \frac{\sum_{j<i} (x_i - x_j) p_i p_j}{\bar{x}} \qquad [8]$$

Demostración:

Sin más que sustituir en (4) los valores p_i y q_i en función de los valores x_i y n_i, tenemos:

$$I_G = 1 - \sum_{i=1}^{k} \frac{n_i}{N} \left[\sum_{j=1}^{i-1} 2 \frac{n_j x_j}{N \bar{x}} + \frac{n_i x_i}{N \bar{x}} \right] =$$

$$= 1 - \frac{1}{N^2 \bar{x}} \left[2 \sum_{i=1}^{k} \sum_{j=1}^{i-1} n_i n_j x_j + \sum_{i=1}^{k} n_i^2 x_i \right] =$$

$$= \frac{1}{N^2 \bar{x}} \left(N^2 \bar{x} - \left[2 \sum_{i=1}^{k} \sum_{j=1}^{i-1} n_i n_j x_j + \sum_{i=1}^{k} n_i^2 x_i \right] \right)$$

Como además $N^2 \bar{x} = \sum_{i=1}^{k} n_i^2 x_i + \sum_{i=1}^{k} n_i \sum_{j \neq i} n_j x_j$ tenemos que:

$$I_G = \frac{1}{N^2\bar{x}} \sum_{i=1}^{K} \left[\sum_{j \neq i} n_i n_j x_j - 2 \sum_{j<i} n_i n_j x_j \right]$$

$$= \frac{1}{N^2\bar{x}} \left(\sum_{i=2}^{k} \sum_{j<i} n_j n_i x_i - \sum_{i=1}^{k} \sum_{j<i} n_i n_j x_j \right)$$

$$= \frac{1}{N^2\bar{x}} \sum_{j<i} n_j n_i (x_i - x_j) = \frac{1}{\bar{x}} \sum_{j<i} p_j p_i (x_i - x_j)$$

con lo que el lema queda demostrado.

Es decir, la expresión aproximada I_{G2} y la exacta I_G difieren únicamente en el denominador, que en la fórmula aproximada sustituye N por N-1. Es decir $I_G = I_{G2}$ (N-1)/N. Esta sensible diferencia hace obligatorio el conocimiento del tamaño del colectivo N para la obtención de I_{G2}, dato del que no depende la expresión exacta I_G (basta con conocer las frecuencias relativas). La fórmula (8) aparece propuesta en Lambert, 1996, [2.38], como una aproximación discreta del valor del índice en el caso continuo. Con este lema hemos probado que con dicha fórmula se obtiene el valor exacto en colectivos discretos.

3 INTERPRETACIÓN Y CÁLCULO DEL ÍNDICE COMO UNA COVARIANZA ENTRE DOS VARIABLES

En el caso en el que el índice de Gini se calcule para una distribución continua, se puede interpretar como la covarianza entre la variable X y la variable transformada mediante su función de distribución, F(X), (ver Lerman y Yitzhaki, 1984). Concretamente, el índice de Gini para el caso continuo se puede expresar como:

$$I_G = 2\frac{\mathrm{cov}(x, F(x))}{\mu} \qquad [9]$$

donde μ denota la media de *X*.

Un estimador de este valor, obtenido a partir de una muestra de tamaño *N*, en la que los datos estén agrupados en *k* valores diferentes $x_1 < x_2 < \ldots < x_k$ es:

$$\hat{I}_G = 2\frac{\sum_{i=1}^{k} p_i (x_i - \bar{x})(\hat{F}_i - \bar{F})}{\bar{x}} \qquad [10]$$

donde $\hat{F}_i = \sum_{j=1}^{i-1} p_j + \frac{p_i}{2}$ es la estimación de $F(\cdot)$ para los distintos valores muestrales, y tanto $\bar{x}$ como $\bar{F}$ representan las respectivas medias muestrales ponderadas por los pesos p_i. (Ver Lerman y Yitzhaki, 1989).

El siguiente lema justifica el uso de la expresión (10) no sólo en base al hecho de ser una aproximación discreta del caso continuo, si no porque se corresponde exactamente con la definición del índice de Gini en colectivos discretos.

Lema 3 Las expresiones del índice de Gini obtenidas mediante (4) y (10) coinciden.

Demostración:

Partiremos de la expresión exacta obtenida en (4).

$$(4) = 1 - \sum_{i=1}^{k} p_i \left[\sum_{j=1}^{i-1} 2q_j + q_i \right] = 1 - \frac{2}{\bar{x}} \sum_{i=1}^{k} x_i p_i \left[1 - \hat{F}_i \right] = \frac{2}{\bar{x}} \left(\sum_{i=1}^{k} x_i p_i \hat{F}_i - \frac{\bar{x}}{2} \right)$$

Como vemos, el resultado anterior se corresponderá exactamente con la expresión (10) si demostramos que la media ponderada de los valores $\hat{F}_i$, $\bar{F}$, es siempre 1/2.

Sin más que cambiar el orden de los sumatorios en sencillo ver que

$$\bar{F} = \sum_{i=1}^{k} \hat{F}_i p_i = \sum_{i=1}^{k} p_i \left(\sum_{j<i} p_j + p_i / 2 \right) = \sum_{i=1}^{k} p_i \left(\sum_{j>i} p_j + p_i / 2 \right)$$

Ahora bien,

$$\sum_{i=1}^{k} p_i \left(\sum_{j<i} p_j + p_i / 2 \right) + \sum_{i=1}^{k} p_i \left(\sum_{j>i} p_j + p_i / 2 \right) = \sum_{i=1}^{k} p_i \sum_{j=1}^{k} p_j = 1$$

con lo que la igualdad queda demostrada.

4 EJEMPLOS Y CONCLUSIONES

En esta sección mostramos tres ejemplos didácticos, ver Tabla 1, utilizados en distintos libros de texto para la ilustración del método de cálculo del índice de Gini. En la Tabla 2 obtendremos tanto los valores exactos del índice como los aproximados. También detallaremos mediante la Tabla 3 el método de cálculo en la práctica con objeto de que quede clara la facilidad de cómputo del valor exacto tanto por medio de paquetes estadísticos como por medio del cálculo manual.

Tabla 1
DESCRIPCIÓN DE LOS EJEMPLOS

A[1]		B[2]		C[3]	
x_i	n_i	x_i	n_i	x_i	n_i
800	20	15	1	80	10
7200	80	25	2	150	20
		40	4	200	15
		60	2	240	5
		155	1		

[1]Escuder (1982), pág. 91.

[2]Rodríguez y Arenales (1988), pág. 142

[3]Uriel y Muñiz (1988), pág. 45

Tabla 2
VALORES EXACTOS Y APROXIMADOS DEL ÍNDICE DE GINI

	I_{G1}	I_{G2}	I_G
A	0.865	0.175	0.173
B	0.350	0.373	0.336
C	0.162	0.171	0.167

Tabla 3

TABLA DE TRABAJO ASOCIADA AL EJEMPLO C

x_i	n_i	p_i	q_i	$\hat{F}_i = p_1 + \ldots + p_{i-1} + p_i / 2$
80	10	0.20	0.10	0.10
150	20	0.40	0.375	0.40
200	15	0.30	0.375	0.75
240	5	0.10	0.15	0.95

$$I_G = \frac{2}{\bar{x}} \sum_{i=1}^{4} x_i p_i \hat{F}_i - 1 = \frac{2}{\bar{x}} \text{cov}(x, \hat{F}) = 0.167$$

Mediante estos ejemplos apreciamos las diferencias existentes entre las dos aproximaciones y el valor exacto. En particular, en el ejemplo A la aproximación I_{G1} resulta claramente inadecuada mientras que el valor I_{G2} está más cerca del verdadero, dado el tamaño del colectivo. En el ejemplo B sin embargo, resulta mejor la aproximación obtenida por I_G, cuando en el libro referenciado con este preciso ejemplo se trata de ilustrar el cálculo mediante I_{G2}. Por último, en el ejemplo C los tres valores están relativamente próximos. Ahora bien, como se aprecia en la Tabla 3, el número de operaciones necesarias para obtener cualquiera de los tres valores es similar, lo que aconseja el uso de la fórmula exacta en cualquier caso.

5. CONCLUSIONES

1. En esta nota se analizan las relaciones existentes entre distintas expresiones del índice de Gini en colectivos con datos agrupados.

2. En contextos descriptivos, donde la definición del índice es (1), proponemos la obtención del valor exacto siguiendo el método que se muestra en la Tabla 3, en lugar de la expresión (2) propuesta habitualmente en los libros de texto y que resulta una mala aproximación.

3. De la tercera de las expresiones propuestas, (10), se deduce que el índice de Gini se puede computar usando los paquetes estadísticos habituales a partir de la covarianza entre dos columnas de datos, x_i y $\hat{F}_i$.

4. A la vista de (9) se puede interpretar la expresión I_{G2} como la resultante de estimar la covarianza entre la variable X y su función de distribución F(X) de forma

insesgada; es decir, tomando $p_i = n_i /(N-1)$. Así, mediante esta interpretación, el valor definido en (1) es coherente con la definición de covarianza en contextos descriptivos.

REFERENCIAS

ESCUDER, R., y MURGUI, J.S. (1995) «Estadística Aplicada. Economía y Ciencias Socíales». *Tirant lo Blanch*, Valencia.

ESCUDER, R. (1982) «Introducción a la Estadística Económica». *Tebar Flores*, Madrid,

LAMBERT, P.J. (1996) «La Distribución y Redístríbución de la Renta». *Instituto de Estudios Fiscales*, Madrid.

LERMAN, R.I., y YITZHAKI, S. (1984) «A note on the calculation and interpretation of the Gini coefficient», *Economics Letters, 15*, 363-368.

LERMAN, R.I., y YITZHAKI, S. (1989) «Improving the accuracy of estimases of the Gini coefficient», *Journal of Econometrics, 42*, 43-47.

MARTÍN PLIEGO, J. (1994) «Introducción a la Estadística Económica y Empresarial». *AC*, Madrid.

MONTIEL, A.M.; RIUS, F. y BARÓN, F.J. (1997) «Elementos básicos de Estadística Económica y Empresarial». *Prentice Hall*, Madrid.

PYATT, G. (1976) «On the interpretation and disaggregation of the Gini coefficient». *The Economic Journal*, June, 243-254.

RODRÍGUEZ, J., y ARENALES, C. (1988) «Problemas de Estadísticas Económicas». *Pirámide*, Madrid.

URIEL, E. y MUÑIZ M. (1988) «Estadística Económica y Empresarial». *AC*, Madrid.

A NOTE ON THE CALCULATION OF THE GINI COEFFICIENT

SUMMARY

In this short note, we establish the relation between different formulations to calculate the Gini index. From the geometric definition, based on the Lorenz curve, we propose to obtain the exact value of the area below this curve, instead of the aproximations used in basic books of descriptive statistics. For this purpose, we will describe a simple method for the computation that also allows us to interpret the index as a covariance between two variables.

Keywords: Concentration, Lorenz curve, Inequality measures

Clasification AMS: 6201