Sull'interpretazione e scomposizione dell'indice di concentrazione

Laura Grassini, Università di Firenze.

In questo lavoro viene introdotta una definizione alternativa dell'indice di concentrazione di una funzione non negativa $g(x)$ della variabile statistica $x \geq 0$, riferito a N osservazioni. La formula proposta consente una immediata interpretazione economica e statistica dell'indice e lo studio della sua scomponibilità qualora le N osservazioni originarie siano raggruppate in classi. La proprietà di scomponibilità dell'indice di concentrazione viene quindi discussa in vista della sua utilità applicativa.

1. Introduzione

Il concetto di curva di Lorenz, tradizionalmente associato allo studio della distribuzione personale del reddito, è stato esteso e generalizzato trovando utile applicazione nell'analisi di numerosi altri fenomeni economici (cfr. Mahalanobis, 1960; Iyengar, 1960; Kakwani, 1977; Grassini, 1982). La curva di Lorenz e il noto indice (o rapporto) di concentrazione Gx di una variabile statistica non negativa x con funzione di ripartizione $F(x)$, introdotto da Gini (Gini, 1914) e definito, nella letteratura anglosassone, "indice del Gini" o "indice Gini-Lorenz" sono, infatti, il caso particolare della curva e dell'indice di concentrazione C di una trasformata non negativa di x, $y = g(x)$, quando $g(x)=x$(*).

Mahalanobis e Iyengar, nel 1960, seguirono questa impostazione nell'analizzare i dati sulla spesa in beni di consumo in India e definirono, appunto, curva di concentrazione la curva di Lorenz generalizzata. Successivamente Kakwani (Kakwani, 1977, 1980) ha sviluppato una trattazione più generale e rigorosa.

Questa impostazione ha contribuito notevolmente allo sviluppo della metodologia statistica sulla concentrazione. Lo studio tradizionale della concentrazione di una variabile, attraverso la curva di Lorenz e l'indice del Gini, è di tipo univariato e prescinde, cioè, dalle relazioni che essa ha con altre

(*) In questa nota, per la definizione dell'indice C seguiremo uno schema molto semplice in cui la variabile y viene definita come trasformazione della variabile statistica (o casuale) non negativa x, mediante una funzione non negativa $g(x)$. Questo schema presenta carattere di generalità potendo anche essere riferito allo studio della dipendenza fra grandezze. In questo caso, infatti, date due variabili casuali non negative z e u, con funzione di ripartizione congiunta $L(z,u)$, la $g(.)$ potrebbe rappresentare la relazione $E(z|u=u_0) = g(u_0)$.

grandezze. L'analisi mediante C consente invece di recuperare queste informazioni che contribuiscono a spiegare la natura della variabile stessa (**).

Generalmente la definizione di C avviene ricorrendo al suo significato geometrico e solo attraverso alcune elaborazioni successive è possibile darne una interpretazione statistica ed economica. In questa nota si propone una diversa definizione dell'indice C riferito ad un numero finito N di osservazioni ($(N-1)/N \simeq 1$), che consente un'immediata interpretazione economica, statistica e, inoltre, lo studio della sua scomponibilità qualora le N osservazioni originarie siano raggruppate in classi. La definizione di C qui introdotta rappresenta una estensione della formulazione proposta da Pyatt (1976) per l'indice del Gini Gx.

2. Definizione di indice di concentrazione

Date N osservazioni sulla variabile $x \geqslant 0$ rilevata su N individui, ove $(N-1)/N \simeq 1$, l'indice del Gini di x è rappresentabile come (Pyatt, 1976):

$$Gx = \frac{1}{N^2\ \bar{x}}\ \Sigma_{i=1}^{N}\ \Sigma_{j=1}^{N}\ A_{ij} \tag{1}$$

dove $\bar{x}$ è il valore medio di x e:

$$A_{ij} = \begin{cases} 0 \ se \ x_j \leqslant x_i \\ x_j - x_i \ \text{altrimenti} \end{cases}$$

Per l'interpretazione di Gx, si ipotizza il seguente gioco statistico. Ogni individuo i estrae a caso dalla popolazione degli N valori x_k ($k=1,2, ..., N$) il valore x_j di cui si può, eventualmente, appropriare. Se $x_j > x_i$, ha convenienza ad appropriarsi di x_j, nel caso contrario a rifiutarlo. Nessun individuo può perdere a questo gioco: tutti, escluso il "più ricco", hanno infatti un guadagno atteso positivo. Gx rappresenta, pertanto, il guadagno atteso medio degli N individui, rapportato al valore $\bar{x}$.

Supponiamo che ad ogni valore x_i di x sia associato il valore y_i della variabile y, ottenibile attraverso una generica funzione a valori non negativi $g(.)$, tale che $y_i = g(x_i)$. Si può definire, allora, l'indice di concentrazione C di $g(x)$, rispetto alle N osservazioni, come:

$$C = \frac{1}{N^2\ \bar{y}}\ \Sigma_{i=1}^{N}\ \Sigma_{j=1}^{N}\ B_{ij} \tag{2}$$

(**) L'idea di studiare la dipendenza fra grandezze mediante misure di concentrazione è già presente in un articolo di Gini del 1910 (Gini, 1910).

dove $\bar{y}$ è il valore medio di y e:

$$B_{ij} = \begin{cases} 0 \; se \; x_j \leqslant x_i \\ y_j - y_i \text{ altrimenti} \end{cases}$$

C rappresenta una generalizzazione di Gx con cui coincide se $g(x)=x$ ed è il guadagno atteso medio di y condizionato ad un guadagno positivo di x, rapportato al valore $\bar{y}$.

Secondo la (2) il legame fra x e y può essere così interpretato: se l'individuo *i.esimo* è disposto ad appropriarsi di x_j è disposto anche a modificare l'ammontare posseduto di y. Inoltre, in base alle espressioni (1) e (2) si può ritrovare la seguente relazione già dimostrata per altra via (Pyatt, Chen e Fei, 1980; Kakwani, 1980, pp. 173 e segg.):

$$- Gy \leqslant C \leqslant Gy \tag{3}$$

dove Gy è l'indice del Gini calcolato sulla variabile y.

Il caso $C = Gy$ ($= —Gy$) si ha se e solo se a variazioni positive di x corrispondono sempre variazioni non negative (non positive) di y. Negli altri casi, un valore positivo (negativo) di C significa una concordanza "in media" positiva (negativa) fra x e y in termini di "predominanza" delle variazioni positive (negative) su quelle negative (positive), secondo la (2).

Dal punto di vista geometrico, infine, è facile dimostrare che C rappresenta il doppio dell'area compresa fra la retta di equidistribuzione e la spezzata congiungente i punti (p_k; q_k) dove (v. Fig. 1):

$$p_k = \frac{k}{N} \qquad q_k = \frac{1}{N\,\bar{y}} \Sigma_{j=1}^{k} y_{[j]}$$

e $y_{[j]}$ rappresenta l'ordinamento dei valori di y corrispondente a valori crescenti di x. Infatti, in tal caso, poiché $x_j > x_i$ se $j > i$, si ha:

$$C = \frac{1}{N^2\,\bar{y}} \Sigma_{j=2}^{N} \Sigma_{i=1}^{j-1} (y_j - y_i)$$

3. Scomposizione dell'indice di concentrazione

Supponiamo che le N osservazioni siano raggruppate in T classi rispetto a valori crescenti di x e che la numerosità n_t di ogni classe $t(t=1, 2, ..., T)$ di estremi $[x^*_{t-1}, x^*_t)$, con $x^*_0=0$ e l'ultima classe aperta, sia tale che $(n_t — 1)/n_t \simeq 1$. Siano, inoltre, $\bar{x}_t$, $\bar{y}_t$ i valori medi di x e y corrispondenti alla classe *t.esima* (*).

(*) È questa la struttura, ad esempio, dei dati derivanti dalle indagini sui bilanci familiari che vengono utilizzati nello studio della relazione fra spesa in beni di consumo e reddito.

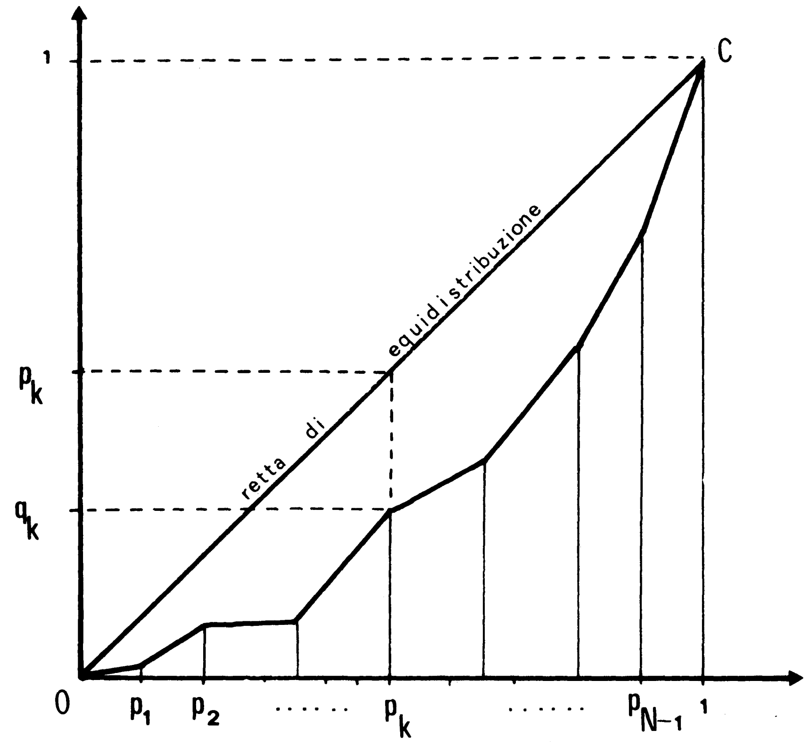

Fig. 1 - Spezzata di concentrazione di g(x)

In tal caso è noto che (Pyatt, 1976):

$$Gx = G_x^I + G_x^T \tag{4}$$

dove

$$G_x^I = \frac{1}{N^2\,\bar{x}}\, \Sigma_{t=1}^{T}\, Gx_t\, \bar{x}_t\, n_t^2$$

$$G_x^T = \frac{1}{N^2\,\bar{x}}\, \Sigma_{t=1}^{T}\, \Sigma_{k=1}^{T}\, A'_{tk}\, n_t\, n_k$$

con

$$A'_{tk} = \begin{cases} 0 \text{ se } \bar{x}_k \leq \bar{x}_t \ (k \leq t) \\ \bar{x}_k - \bar{x}_t \text{ altrimenti} \end{cases}$$

e Gx_t indice del Gini di x relativo alla classe *t.esima*.

Per la (2) e la (4), si ha:

$$C = C^I + C^T \tag{5}$$

dove

$$C^I = \frac{1}{N^2 \, \bar{y}} \, \Sigma_{t=1}^T \, C_t \, \bar{y}_t \, n_t^2$$

$$C^T = \frac{1}{N^2 \, \bar{y}} \, \Sigma_{t=1}^T \, \Sigma_{k=1}^T \, B'_{tk} \, n_t \, n_k$$

con

$$B'_{tk} = \begin{cases} 0 \text{ se } \bar{x}_k \leqslant \bar{x}_t \; (k \leqslant t) \\ \bar{y}_k - \bar{y}_t \text{ altrimenti} \end{cases}$$

e C_t indice di concentrazione di y relativo alla classe *t.esima*. Diversamente avviene per Gy. Infatti, posti:

$$G_y^I = \frac{1}{N^2 \, \bar{y}} \, \Sigma_{t=1}^T \, Gy_t \, \bar{y}_t \, n_t^2$$

$$G_y^T = \frac{1}{N^2 \, \bar{y}} \, \Sigma_{t=1}^T \, \Sigma_{k=1}^T \, A''_{tk} \, n_t \, n_k$$

con

$$A''_{tk} = \begin{cases} 0 \text{ se } \bar{y}_k \leqslant \bar{y}_t \\ \bar{y}_k - \bar{y}_t \text{ altrimenti} \end{cases}$$

e Gy_t indice del Gini di y relativo alla classe *t.esima*, vale in generale:

$$Gy \geqslant G_y^I + G_y^T$$

Possiamo quindi scrivere:

$$Gy = G_y^I + G_y^T + W \tag{6}$$

dove la grandezza non negativa W, la cui definizione esplicita non ha qui rilevanza e per la quale rimandiamo a Mehran (1975a), rappresenta la componente di Gy dovuta alla sovrapposizione delle classi rispetto ai valori di y. Non è detto, infatti, che le classi si presentino disgiunte rispetto ai valori di y.

Quando a variazioni positive di x corrispondono variazioni non negative (non positive) di y, si ha:

$$C^I = G^I_y \ (= -G^I_y) \qquad C^T = G^T_y \ (= -G^T_y) \qquad W = 0$$

Infatti, poiché la (3) può essere riferita agli indici Gy_t e C_t, allora:

$$- G^I_y \leqslant C^I \leqslant G^I_y$$

Inoltre, per la definizione di C^T e di G^T_y:

$$- G^T_y \leqslant C^T \leqslant G^T_y$$

4. Osservazioni conclusive

La scomposizione di C si presenta assai interessante riguardo a:

1) la descrizione del legame di concordanza esistente fra x e y in relazione ad un determinato criterio di classificazione delle osservazioni. In questa ottica la (6) suggerisce una particolare interpretazione della componente W che, se diversa da zero, indicherebbe la non monotonicità di $g(x)$ rispetto a x. Più in generale, si può dire che W rappresenti il potere discriminante di una variabile rispetto ad un determinato criterio di raggruppamento delle osservazioni. Più basso è il peso di W sul valore totale dell'indice Gy, più alto è il potere discriminante della variabile;

2) la possibilità di estendere a C i procedimenti utilizzati per la determinazione del valore massimo di Gx da dati raggruppati che consiste nella massimizzazione della componente G^I_x essendo G^T_x noto (Gastwirth, 1972; Gastwirth e Krieger, 1975; Mehran, 1975).

Quando le N coppie di osservazioni (x_i, y_i), $i=1,2, \ldots, N$, si presentano raggruppate in T classi rispetto a valori crescenti di x e solo le medie aritmetiche di x e y e le frequenze osservate di classe sono disponibili, due metodi vengono generalmente usati per la stima di C. Il primo consiste nel calcolo di C^T che, come abbiamo visto, coincide con C solo se $C^I=0$. Il secondo metodo consiste nell'adattamento di una funzione ai dati empirici per rappresentare la curva di concentrazione di $g(x)$ e ripropone, quindi, i consueti problemi connessi con la scelta della forma analitica e del metodo di interpolazione (Kakwani, 1980, pp. 198 e segg.; Grassini, 1982a). A differenza della curva di Lorenz di x, infatti, la curva di concentrazione di $g(x)$ non ha proprietà analitiche predeterminate poiché queste dipendono da quelle di $g(x)$. La scelta della forma analitica si basa, il più delle volte, sulla sua bontà di adattamento definita rispetto ad un determinato metodo di interpolazione.

L'ottenimento di una stima di intervallo di C consentirebbe di evitare questi inconvenienti e il procedimento di determinazione dei valori minimo e massimo di C, attraverso la massimizzazione e minimizzazione di C^I, essendo C^T noto, può consentire la formulazione di un metodo generale di stima di intervallo di C.

Questa nota ha carattere essenzialmente metodologico. L'autrice sta ultimando un lavoro in cui si affrontano gli aspetti applicativi dell'indice C e delle sue proprietà, per lo studio della relazione fra reddito e spesa in beni di consumo.

Si ringraziano i revisori per gli utili suggerimenti.

Riferimenti bibliografici

Bhattacharya N., B. Mahalanobis (1967) "Regional Disparities in Household Consumption in India", *Joruna1 of the American Statistical Association*, vol. 62.

Blitz R.C., J.A. Brittain (1964) "An Extention of the Lorenz Diagram to the Correlation of Two Variables", *Metron*, vol. 23.

Gastwirth J.L. (1972) "The Estimation of the Lorenz Curve and Gini Index", *Review of Economics and Statistics*, vol. 54.

Gastwirth J.L., A.M. Kieger (1975) "On Bounding Moments from Grouped Data", *Journal of the American Statistical Association*, vol. 70.

Gini C. (1910) "Indici di concentrazione e di dipendenza", *Biblioteca dell'Economista*, vol. XX, U.T.E.T., Torino.

Gini C. (1914) "Sulla misura della concentrazione e della variabilità dei caratteri", *Atti del Regio Istituto Veneto di Scienze, Lettere ed Arti*, Tomo LXXIII, Parte II.

Grassini L. (1982) "Un metodo deterministico per la stima delle elasticità engeliane da dati raggruppati", *Statistica*, n. 4.

Grassini L. (1982a) "On the Estimation of Engel Elasticities with Application to Italian Data", *Proceedings of the 2nd Task Force Meeting on Input-Output Modeling 1981*, CP-32-82, International Institute for Applied System Analysis, Laxenburg, Vienna.

Iyengar N.S. (1960) "On a Method of Computing Engel Elasticities from Concentration Curves". *Econometrica*, vol. 28.

Kakwani N.C. (1977) "Application of Lorenz Curves in Economic Analysis", *Econometrica*, vol. 45.

Kakwani N.C. (1980) *Income Inequality and Poverty* Oxford University Press.

Mahalanobis P.C. (1960) "A Method of Fractile Graphical Analysis", *Econometrica*, vol. 28.

Mehran F. (1975) "Bounds on the Gini Index based on Observed Points of the Lorenz Curve", *Joruna1 of the American Statistical Association*, vol. 70.

Mehran F. (1975a) "A Statistical Analysis on Income Inequality based on a Decomposition of the Gini Index", *Proceedings of the 40th Session of the International Statistical Institute*, Varsavia.

Nygard F., A. Sandstrom (1981) *Measuring Income Inequality*, Almquist & Winsell, Stockholm.

Pyatt G. (1976) "On the Interpretation and Disaggregation of Gini Coefficients", *Economic Journal*, vol. 86.

Pyatt G., C.N. Chen, J. Fei (1980) "The Distribution of Income by Factor Components", *Quarterly Journal of Economics*, vol. 94.

Summary

This paper introduces an alternative formulation of the concentration index of a non negative function $g(x)$ of a statistical variable $x \geqslant 0$, with respect to N observations. This formula offers an interesting interpretation of the index from an economic and statistical point of view and allows studying its decomposability when the N observations are grouped. The index decomposability is discussed with respect to its use in empirical applications.