Report n.102

Sul calcolo del rapporto di concentrazione del Gini

Gilberto Ghilardi

Pisa, Aprile 1996

Questa ricerca è stata finanziata in parte dal Ministero dell'Università e della Ricerca Scientifica e Tecnologica (già fondi 60%).

INDICE

Sul calcolo del rapporto di concentrazione del Gini

1. Introduzione

I cosiddetti caratteri statistici trasferibili tra più unità statistiche sono oggetto di studio in funzione di diversi aspetti. Tra questi, la concentrazione viene presa in esame da quasi tutti i testi utilizzati dagli studenti di un primo corso di statistica (ad esempio Frosini 1995, Girone e Salvemini 1987, Giusti 1986, Leti 1983). Gli indici di concentrazione sono molteplici e tra questi nella maggioranza dei casi viene proposto il rapporto di concentrazione del Gini (Gini 1912, Gini 1913-14), che può essere calcolato mediante diverse espressioni. Con questa nota, ci proponiamo di presentare una nuova espressione per il calcolo del rapporto di concentrazione del Gini su una distribuzione di unità statistiche secondo le classi di valori di un determinato carattere. Tale espressione rappresenta una modalità di calcolo, che fornisce valori più precisi di quelli ottenuti con altre espressioni ed è determinata sulla base dell'ipotesi più realistica dell'uniforme distribuzione delle unità statistiche ricadenti in ogni classe, rispetto a quella consistente nell'assumere per tutte le unità un valore corrispondente a quello centrale della classe di appartenenza.

2. Alcune espressioni per il calcolo del rapporto di concentrazione.

La misura della concentrazione attraverso il rapporto di concentrazione del Gini può essere determinata con diverse espressioni anche in funzione dei dati disponibili. Infatti, qualora si disponga di una serie (supposta ordinata) di dati,

$$x_1 \leq x_2 \leq ..... \leq x_N$$

per N unità statistiche, allora si possono definire i rapporti,

$$p_i = \frac{i}{N} \quad ; \qquad q_i = \frac{S_i}{NM} = \frac{x_1 + ... + x_i}{NM} \quad ; \qquad (i = 1,2,...,N)$$

dove M rappresenta la media aritmetica degli N dati x_i, attraverso i quali viene costruito l'indice

$$R = \frac{\sum_{i=1}^{N-1} (p_i - q_i)}{\sum_{i=1}^{N-1} p_i} \qquad [1]$$

proposto del Gini per la misura della concentrazione.

Invece, se i dati di partenza sono forniti attraverso una distribuzione secondo k classi $(x_i - x_{i+1}; i, ..., k)$ delle frequenze N_i $\left(N = \sum_{i=1}^{k} N_i \right)$; allora per il calcolo del rapporto del rapporto di

concentrazione del Gini é stata proposta la cosiddetta formula dei trapezi

$$R^* = 1 - \sum_{i=1}^{k} (p_i - p_{i-1})(q_i + q_{i-1}) \qquad [2]$$

dove per i=1 si pone $p_0 = q_0 = 0$ e le quantità p_i, q_i

$$p_i = \frac{N_1 + ... + N_i}{N} \qquad q_i = \frac{S_i'}{NM} = \frac{x_1' N_1 + ... + x_i' N_i}{NM} \quad (i = 1,...,k) \qquad [3]$$

sono determinate per ciascuna delle k classi della distribuzione mediante le frequenze relative cumulate e assumendo il valore centrale x' della classe per tutte le unità che ricadono nella classe stessa.

Naturalmente, altre possibilità per il calcolo dell'indice in questione potrebbero essere individuate, tenendo conto della relazione esistente tra la differenza media e l'indice stesso, ma anche in questi casi si perviene a delle soluzioni che nel caso di una distribuzione per classi si basano sul ricorso al valore centrale di classe, come il valore che é assunto dalle unità che ricadono nella classe considerata.

Pertanto, può essere interessante trovare un modo alternativo per calcolare il rapporto di concentrazione, a partire da una distribuzione di frequenza per classi di valori di una variabile x, facendo l'ipotesi che all'interno di ciascuna classe le unità statistiche si distribuiscano in maniera uniforme, ad una distanza

$$h_i = \frac{x_{i+1} - x_i}{N_i} \qquad [4]$$

pari al rapporto tra l'ampiezza (diversa da zero) della classe e la frequenza N_i.

In queste condizioni, il calcolo del rapporto di concentrazione può essere effettuato considerando le k coppie di valori (p_i, q_i) e i limiti inferiori x_i delle classi che figurano nell'espressione

$$R' = 1 - \sum_{i=1}^{k} (p_i - p_{i-1}) \left[2q_{i-1} + \frac{x_i N_i}{NM} + \frac{h_i (N_i + 1)(2N_i + 1)}{6NM} \right] \qquad [5]$$

in cui, come in precedenza, per i=1 si pone $p_0 = q_0 = 0$.

3. Uniforme distribuzione e calcolo del rapporto di concentrazione.

Per ottenere l'espressione [5] riportata si può operare considerando che se all'interno di ciscuna classe $(x_i - x_{i+1})$ i valori della variabile x subiscono da una unità alla successiva un incremento pari a h_i, allora la usuale spezzata di concentrazione é delimitata inferiormente da N trapezi e all'interno di ogni classe l'area A_j del j-mo trapezio $(j = 1, ..., N_i)$ é data dall'espressione

$$A_j = \frac{1}{2}\frac{1}{N} \left[2q_{i-1} + 2 \frac{(j-1)x_i + \sum_{l=1}^{j} (l-1)h_i}{NM} + \frac{x_i + jh_i}{NM} \right] \qquad [6]$$

del prodotto tra l'altezza 1/N e la somma delle basi che figura entro le parentesi quadre.

Pertanto, l'area A_i dei trapezi individuati con riferimento alla i-ma classe sarà data dalla somma

$$A_i = \sum_{j=1}^{N_i} \frac{1}{2} * \frac{1}{N} \left[2q_{i-1} + 2\frac{jx_i + h_i \sum_{\ell=1}^{j} \ell}{NM} - \frac{x_i + jh_i}{NM} \right] \qquad [7]$$

delle aree degli N_i trapezi relativi a tale classe, mentre l'area A

$$A = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \frac{1}{2} \frac{1}{N} \left[2q_{i-1} + 2j\frac{x_i}{NM}\frac{h_i}{NM} * \frac{1+j}{2} j - \frac{x_i + jh_i}{NM} \right] \qquad [8]$$

di tutti i trapezi sottostanti alla spezzata di concentrazione si ottiene sommando l'espressione [7] per tutte le k classi e può essere scritta con opportune semplificazioni nella forma

$$A = \frac{1}{2} \sum_{i=1}^{k} \left[\frac{N_i}{N} \left(2q_{i-1} + \frac{x_i N_i}{NM} + \frac{h_i}{NM} * \frac{(N_i+1)(2N_i+1)}{6} \right) \right] \qquad [9]$$

equivalente alla metà della sommatoria che compare nell'espressione [5], già proposta per il calcolo del rapporto di concentrazione, qualora si faccia l'ipotesi che le unità statistiche siano distribuite uniformemente all'interno di ciascuna classe dei valori della variabile x.

Vale la pena di osservare che nel caso in cui si facesse riferimento al valore centrale $x_i' = \frac{(x_{i+1} + x_i)}{2}$ per ciascuna unità appartenente alla classe i, allora la quantità

$$q_i - q_{i-1} = \frac{x_i' N_i}{NM} \qquad [10]$$

fornirebbe un'approssimazione per eccesso dell'espressione

$$\frac{x_i N_i}{NM} + \frac{h_i}{NM}\frac{(N_i+1)(2N_i+1)}{6} \qquad [11]$$

la quale rappresenta gli ultimi due termini che figurano nella [9] all'interno delle parentesi tonde.

Infatti, tenendo conto del fatto che per l'ipotesi formulata sulla uniforme distribuzione le unità all'interno della classe (x_i, x_{i+1}) saranno collocate ad una distanza pari a h_i e che la prima modalità é $x_i + \frac{h_i}{2}$, occorre sostituire $\frac{h_i}{2}$ al posto di h_i nella [11] e la differenza tra la [10] e la [11]

$$\frac{1}{NM}\left(x_i' N_i - x_i N_i - \frac{h_i}{2}\frac{(N_i+1)(2N_i+1)}{6}\right) \qquad [12]$$

é sempre positiva, in quanto essa é equivalente all'espressione,

$$\frac{h_i}{NM}\left[\frac{N_i^2}{2} - \frac{(N_i+1)(2N_i+1)}{12}\right] = \frac{h_i}{NM}\frac{4N_i^2 - 3N_i - 1}{12} \qquad [13]$$

in cui compare un polinomio che é funzione di N_i e che è sempre positivo per N_i maggiore di uno, risultando uguale a zero quando $N_i=1$ e l'unica modalità coincide con il valore centrale della classe x_i', cosicché le due espressioni [10] e [11] forniscono lo stesso risultato (purché l'unica unità appartenente alla classe i-ma venga collocata al centro della classe stessa, ovvero nel punto che dista $\frac{h_i}{2}$ dall'estremo inferiore x_i e da quello superiore x_{i+1}.

4. Conclusioni

Naturalmente, i valori del rapporto di concentrazione che si hanno in funzione dell'espressione definita con riferimento all'ipotesi di equidistribuzione all'interno delle varie classi potranno essere più o meno diversi da quelli che si avrebbero attraverso l'uso di altre formule, a seconda dell'ampiezza e del numero delle classi della distribuzione e dei valori delle frequenze, ma in ogni caso essi non sono inferiori a questi ultimi, perché l'area di concentrazione non viene sottostimata, come accade quando per tutte le unità ricadenti in ogni classe si considera un valore uguale a quello centrale della classe. Tuttavia, al di là delle differenze numeriche riscontrabili tra i risultati delle diverse espressioni, riteniamo che la formula qui proposta per il calcolo del rapporto di concentrazione sia più aderente alla realtà e che, pertanto, sia da preferire come una soluzione logicamente più accettabile.

Riferimenti bibliografici

FROSINI B. V. (1995). *Introduzione alla statistica*. La Nuova Italia Scientifica, Roma.

GINI C. (1912). Variabilità e mutabilità. *Studi economico-giuridici della Facoltà di Giurisprudenza della R. Università di Cagliari*, Anno III (lavoro riprodotto in: GINI C. - *Memorie di statistica metodologica: variabilità e concentrazione*. Veschi, Roma, 1955).

GINI C. (1913-14). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Ist. Veneto di Scienze Lettere ed Arti*, Tomo LXXIII (lavoro riprodotto in: GINI C. - *Memorie di statistica metodologica: variabilità e concentrazione*. Veschi, Roma, 1955).

GIRONE G., SALVEMINI T. (1987). *Lezioni di statistica*. Cacucci, Bari.

GIUSTI F. (1986). *Introduzione alla statistica*. Loescher, Torino.

LETI G. (1983). *Statistica descrittiva*. Il Mulino, Bologna.