PIETRO MULIERE

UNA NOTA SUL COEFFICIENTE DI CORRELAZIONE TRA L'INDICE G DI COGRADUAZIONE DI GINI E L'INDICE τ DI KENDALL

Estratto dal
GIORNALE DEGLI ECONOMISTI E ANNALI DI ECONOMIA
Settembre-Ottobre 1976

CISALPINO - GOLIARDICA
Via Bassini 17-2
20133 MILANO

UNA NOTA SUL COEFFICIENTE DI CORRELAZIONE TRA L'INDICE G DI COGRADUAZIONE DI GINI E L'INDICE τ DI KENDALL

1. - INTRODUZIONE.

Un indice per controllare la mancanza o meno di associazione tra due caratteri osservati congiuntamente fu proposto da C. Gini (1) ed è dato dalla seguente espressione:

$$G = \frac{2}{D} \sum_{k=1}^{n} \{ | n + 1 - r(X_k) - r(Y_k) | - | r(X_k) - r(Y_k) | \} \qquad (1)$$

dove

$$D = \begin{cases} n^2 & \text{per } \underline{n} \text{ pari} \\ n^2 - 1 & \text{per } \underline{n} \text{ dispari} \end{cases}$$

e $r(X_k)$, $r(Y_k)$, $k = 1, 2, \ldots, n$, indicano rispettivamente il rango della k.ma osservazione della X e della k.ma osservazione della Y.

L'indice G è stato, ed è, oggetto di indagine e di applicazione da parte di numerosi Autori della Scuola italiana ma viene sistematicamente ignorato nella letteratura statistica internazionale.

A nostro avviso, tuttavia, non vi sono plausibili motivi che facciano preferire altri indici come il τ di Kendall o il R di Spearman e ciò perchè siamo convinti che l'indice G in parola pone in evidenza particolari aspetti dell'associazione non posti in luce né da τ, né da R. Alcune differenze tra i tre indici in questione sono state analizzate da T. Salvemini (2); D. M. Cifarelli e E. Regazzini (3) hanno mostrato che l'efficienza asintotica relativa (ARE) di R e τ nei confronti del coefficiente di correlazione r è superiore a quella di G nel caso di popolazioni normali.

Vi sono indicazioni, tuttavia, che fanno supporre che, da qualche punto di vista, G sia preferibile, ed anche nel caso dell'efficienza asintotica relativa potrebbe verosimilmente accadere che la potenza di G sia maggiore di quella di R e τ qualora si analizzassero popolazioni non normali e per ampiezze campionarie non molto elevate.

---

(1) C. GINI, *L'ammontare e la composizione della ricchezza delle nazioni*. Torino, Bocca, 1914.

(2) T. SALVEMINI, *Sui vari indici di cograduazione*, in « Statistica », n. 2, 1951.

(3) D. M. CIFARELLI, E. REGAZZINI, *Ancora sull'indice di cograduazione di C. Gini*. Torino, Istituto di Matematica finanziaria dell'Università, Serie III, n. 5, 1974.

Un altro modo per cogliere eventuali sostanziali differenze tra i suddetti indici potrebbe essere quello di studiare il comportamento del coefficiente di correlazione tra le varie coppie di indici, come ha fatto T. Salvemini nel lavoro citato.

Scopo di questa breve nota è appunto quello di stabilire l'espressione generale del coefficiente di correlazione tra G e gli altri due nel caso di indifferenza (4).

A differenza del coefficiente di correlazione tra R e τ, fornito da M. G. Kendall (5), dato dalla seguente espressione:

$$\rho(R.\tau) = \frac{2(n+1)}{\sqrt{2n(2n+5)}} \tag{2}$$

e che per $n \to \infty$ converge a 1, non vale un risultato analogo per $\rho(G,\tau)$ e $\rho(G,R)$.

Questo fatto fa cadere una congettura avanzata da T. Salvemini (6) che ha calcolato il coefficiente di correlazione tra gli indici G e τ e G e R per un numero limitato di termini $n = 4, 5, 6$.

Il fatto che, neppure asintoticamente, tali coefficienti di correlazione valgano 1, ci sembra pongano in risalto una volta di più come gli indici suddetti, mettano in evidenza, sistematicamente, aspetti differenti di uno stesso fenomeno.

Da questo punto di vista, perciò, converrebbe impiegare a coppie i vari indici onde poter meglio risolvere i problemi per cui tradizionalmente vengono utilizzati. Naturalmente occorrerebbe però conoscere la legge di distribuzione congiunta delle varie coppie. Sebbene sia possibile fornire formule ricorrenti per la determinazione di tali leggi per ogni ampiezza campionaria finita, noi ci siamo preoccupati solo di stabilire la legge di distribuzione asintotica di (G,τ) e (G,R).

2. - Il risultato.

Sia (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ $n \geq 2$, un campione di n elementi estratto da una popolazione bidimensionale (X, Y), e sia $C_{ij} = \text{sign}\ (X_j - X_i)\ (Y_j - Y_i)$ dove

$$\text{sign}\ (u) = \begin{cases} -1 & \text{per } u < 0 \\ 0 & \text{per } u = 0 \\ 1 & \text{per } u > 0 \end{cases}$$

Il τ di Kendall, come misura di associazione tra le componenti X e Y, è dato da:

$$\tau = (2 \mid n(n-1)) \sum_{1 \leq i < j \leq n} C_{ij} \tag{3}$$

Supponendo le X_i ordinate ed indicando R_h il rango dell'osservazione Y che corrisponde a X_h, l'indice τ può essere scritto nel modo seguente (7):

(4) Nel periodo intercorso tra l'accettazione della presente nota e la sua stampa l'Autore ha avuto occasione di apprendere che tali coefficienti di correlazione sono stati determinati da A. Herzel nell'articolo *Sulla distribuzione campionaria dell'indice di cograduazione del Gini* in « Metron », vol. XXX, n. 1-4, p. 137, 1972. L'A. ha ritenuto di pubblicare ugualmente la parte del lavoro che riguarda questo tema poichè la dimostrazione che viene presentata qui è differente da quella offerta dallo studioso sopra citato.

(5) M. G. Kendall, *Rank correlation methods*, IV ed. London, Griffith, 1970.

(6) T. Salvemini, *op. cit.* p. 146.

(7) J. Haiek, *A course on nonparametric statistics*, London, Holden-Day, 1969.

$$\binom{n}{2} \tau = \sum_{h=1}^{n} \sum_{i=h+1}^{n} \text{sign } (R_i - R_h). \tag{4}$$

L'altra misura di concordanza, R di Spearman, supposto le X_i ordinate, è definita dall'espressione:

$$R = 1 - \frac{6 \sum_{i=1}^{n} (i - R_i)^2}{n(n^2-1)} \tag{5}$$

Dalla (1), (2), e (5) si hanno le espressioni di $E(G, \tau)$ e $E(G, R)$

$$\begin{aligned}\binom{n}{2} \frac{D}{2} E(G,\tau) = &\sum_{h=1}^{n} \sum_{i=h+1}^{n} \sum_{\substack{k=1 \\ k \neq i \\ k \neq h}}^{n} E\{\text{sign } (R_i - R_h)\,(|\, n+1-k-R_k\,| - |\, k - R_k\,|)\} \\ &+ \sum_{h=1}^{n} \sum_{i=h+1}^{n} E\{\text{sign } (R_i - R_h)\,(|\, n+1-i-R_i\,| - |\, i - R_i\,|)\} \\ &+ \sum_{h=1}^{n} \sum_{i=h+1}^{n} E\{\text{sign } (R_i - R_h)\,(|\, n+1-h-R_h\,| - |\, h - R_h\,|)\}\end{aligned} \tag{6}$$

e

$$\frac{D\,n(n^2-1)}{12} E(G.R) = -\sum_{i=1}^{n} \sum_{k=1}^{n} E\{(i - R_i)^2\,(|\, n+1-k-R_k\,| - |\, k - R_k\,|\} \tag{7}$$

Calcoli elementari conducono ai risultati

$$\begin{cases} E(G,\tau) = \dfrac{4}{D\,n\,(n-1)} \cdot \dfrac{2(n+1)(n^2+1)}{15} \\ E_i(G,R) = \dfrac{4(n^3-n^2+n-1)}{5\,D\,(n-1)^2} \end{cases} \tag{8}$$

e quindi

$$\rho(G,\tau) = \begin{cases} \dfrac{4(n+1)(n^2+1)}{15} \cdot \dfrac{\sqrt{27}}{\sqrt{n^3(2n+5)(n^2+2)}} & \text{per } n \text{ pari} \\ \dfrac{4(n+1)(n^2+1)}{15} \cdot \dfrac{\sqrt{27}}{\sqrt{n(n^2-1)(2n+5)(n^2+3)}} & \text{per } n \text{ dispari} \end{cases} \tag{9}$$

$$\rho(G,R) = \begin{cases} \dfrac{4\sqrt{3}\,(n^2+1)}{5\sqrt{2}\,n\sqrt{n^2+2}} & \text{per } n \text{ pari} \\ \dfrac{4\sqrt{3}\,(n^2+1)}{5\sqrt{2}\,\sqrt{(n^2-1)(n^2+3)}} & \text{per } n \text{ dispari} \end{cases} \tag{10}$$

E' agevole verificare immediatamente che

$$\lim_{n\to\infty} \rho(G.\tau) = \lim_{n\to\infty} \rho(G,R) = \sqrt{\frac{24}{25}} \simeq 0{,}9798\,. \tag{11}$$

Nella Tavola 1 abbiamo fornito alcuni valori di tali coefficienti. Si può notare che anche per valori non elevati di n la correlazione è molto alta e per $n = 4, 5, 6$ i risultati coincidono con quelli dati da T. Salvemini (8).

La correlazione tra G e τ è la più bassa tra le coppie e tende molto lentamente al suo limite dal disotto. I valori di $\rho(G, R)$, invece, sono prossimi al limite già per $n \geqslant 12$.

Diamo ora la dimostrazione del risultato. Analizzeremo solamente la metodologia adottata per il calcolo del coefficiente di correlazione tra τ e G dato che il metodo di determinazione di $\rho(G, R)$ segue la stessa impostazione.

Poichè $R = (R_1 R_2, \ldots, R_n)$ si distribuisce in modo uniforme in $\mathcal{R}$ (insieme delle permutazioni di $1, 2, \ldots, n$) allora:

$$P\{R_i = r\} = \frac{1}{n} \qquad 1 \leqslant i;\ r \leqslant n$$

$$P\{R_i = r;\ R_h = s\} = \frac{1}{n(n-1)} \qquad 1 \leqslant i \neq h;\ r \neq s \leqslant n$$

$$P\{R_i = r;\ R_h = r\} = 0 \qquad 1 \leqslant i \neq h;\ r \leqslant n$$

$$P\{R_i = r \mid R_h = s\} = \frac{1}{n-1} \qquad h \neq i$$

ne segue che per $k = i$ e per ogni j la variabile condizionata

$$\{\text{sign}\ (R_i - R_h) \mid R_k = j\}$$

assume valore 1 con probabilità $\frac{j-1}{n-1}$ e valore -1 con probabilità $\frac{n-j}{n-1}$.

E' immediato verificare che

$$E\{\text{sign}\ (R_i - R_h) \mid R_k = j\} = \frac{2j-n-1}{n-1} \qquad \text{per } k = i;\ \text{per ogni } j$$

Analogamente si ha

$$E\{\text{sign}\ (R_i - R_h) \mid R_k = j\} = \frac{n+1-2j}{n-1} \qquad \text{per } k = h;\ \text{per ogni } j$$

ed inoltre

$$E\{\text{sign}\ (R_i - R_h) \mid R_k = j\} = 0 \qquad \text{per } k \neq i;\ k \neq h;\ \text{per ogni } j$$

Si raggiungono gli stessi risultati anche per la seguente espressione

$$\mathrm{E}\{\text{sign}\ (R_i - R_h) \quad |k - R_k| = |k - j|\}$$

Per una proposizione nota

$$E\left\{X \underset{X}{E}(Y \mid X)\right\} = E(XY)$$

otteniamo

$$E\{|k - R_k| \cdot \text{sign}\ (R_i - R_h)\} = 0 \qquad \text{per } k \neq i;\ k \neq h$$

---

(8) T. Salvemini, *Op. cit.*

$$= \frac{2i^3 - 3(n+1)i^2 + (3n+1)i}{3n(n-1)} + \frac{n+1}{6} \qquad \text{per } k = i$$

$$= \frac{3(n+1)h^2 - 2h^3 - (3n+1)h}{3n(n-1)} + \frac{n+1}{6} \qquad \text{per } k = h$$

ed in modo analogo

$$E\{|n+1-k-R_k| \operatorname{sign}(R_i - R_h)\} = 0 \qquad \text{per } k \neq i;\ k \neq h$$

$$= \frac{2(n+1-i)^3 - 3(n+1)(n+1-i)^2 + (3n+1)(n+1-i)}{3n(n-1)} + \frac{n+1}{6}$$

per $k = i$

$$= \frac{3(n+1)(n+1-h)^2 - 2(n+1-h)^3 - (3n+1)(n+1-h)}{3n(n-1)} + \frac{n+1}{6}$$

per $k = h$.

La loro differenza fornisce

$$E\{\operatorname{sign}(R_i - R_h)(|n+1-k-R_k| - |k-R_k|)\}$$

$$= \frac{6(n+1)i^2 - 4i^3 - 2(3n+1)i + (3n+1)(n+1) - (n+1)^3}{3n(n-1)} \qquad \text{per } k = i$$

$$= \frac{4h^3 - 6(n+1)h^2 + 2(3n+1)h + (3n+1)(n+1) - (n+1)^3}{3n(n-1)} \qquad \text{per } k = h$$

$$= 0 \qquad \text{per } k \neq i;\ k \neq h$$

Indicando con $f(i)$ l'espressione valida per $k = i$ e con $(-f(h))$ quella per $k = h$ la (6) diventa

$$\left(\frac{n}{2}\right)\frac{D}{2} E(G,\tau) = \sum_{h=1}^{n}\sum_{i=h+1}^{n} f(i) - \sum_{h=1}^{n}\sum_{i=h+1}^{n} f(h)$$

$$= 2\sum_{i=1}^{n} i f(i) - (n+1)\sum_{i=1}^{n} f(i)$$

e poichè, come è agevole mostrare,

$$\sum_{i=1}^{n} i f(i) = \frac{(n+1)(n^2+1)}{15}$$

$$\sum_{i=1}^{n} f(i) = 0$$

si ottiene finalmente il risultato enunciato.

3. - Distribuzione asintotica di (G,τ) e (G,R).

In questo numero stabiliremo la legge di distribuzione asintotica (G,τ) e (G,R).

Per la dimostrazione faremo uso del seguente teorema di Hoeffding (9).

(9) D. M. Cifarelli, E. Regazzini, *Op. cit.*

Teorema:

Siano $b_n(i,j) i, j = 1, 2, \ldots, n$, n^2 numeri reali definiti per ogni intero positivo n e sia $(R_1, \ldots, R_n)$ una variabile casuale definita su ogni permutazione di $(1, 2, \ldots, n)$ con la stessa probabilità $\frac{1}{n!}$.

Posto $L_n = \sum_{i=1}^{n} b_n(i, R_n)$

e

$$d_n(i,j) = b_n(i,j) - \frac{1}{n}\sum_{t=1}^{n} b_n(t,j) - \frac{1}{n}\sum_{s=1}^{n} b_n(i,s) + \frac{1}{n^2}\sum_{t,s=1}^{n} b_n(t,s)$$

se

$$\lim_{n\to\infty} \frac{\operatorname{Max}_{1\leqslant i;\, j\leqslant n} d_n^2(i,j)}{\frac{1}{n}\sum_{i,j=1}^{n} d_n^2(i,j)} = 0 \qquad (12$$

allora la variabile L_n si distribuisce asintoticamente secondo una legge normale

Utilizzando il teorema precedente faremo vedere che per ogni reale λ_1 e λ la variabile

$$S = \lambda_1 \sqrt{n}\, G + \lambda_2 \sqrt{n}\, R$$

è asintoticamente normale.

Da ciò, per una nota proposizione, discenderà che la coppia (G, R) è asintoticamente normale con parametri:

$$E(G) = 0, \qquad E(R) = 0$$

$$Var\,(G) \sim \frac{2}{3n}; \qquad \text{Var}\,(R) \sim \frac{1}{n}; \qquad \text{Cov}\,(G,R) \sim \frac{4}{5n}.$$

Definiamo

$$L_n = \sum_{i=1}^{n} b_n(i, R_n) = \sum_{i=1}^{n} \left\{ \frac{2\lambda_1\sqrt{n}}{D} (|\, n+1-i-R_i\,| - |\, i-R_i\,|) \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \frac{6\lambda_2\sqrt{n}}{n(n^2-1)} (i-R_i)^2 \right\}$$

da cui

$$b_n(i,j) = \left\{ \frac{2\lambda_1\sqrt{n}}{D} (|\, n+1-i-j\,| - |\, i-j\,|) - \frac{6\lambda_2\sqrt{n}}{n(n^2-1)} (i-j)^2 \right\}$$

$$i, j = 1, 2, \ldots, n; \quad n \geqslant 2.$$

Si verifica che, nel nostro caso, l'espressione di $d_n(i,j)$ è data da:

$$d_n(i,j) = \left\{ \frac{2\lambda_1\sqrt{n}}{D} (|\, n+1-i-j\,| - |\, i-j\,|) + \right.$$

$$\left. - \frac{3\lambda_2\sqrt{n}}{n(n^2-1)} (2(n+1)j + 2(n+1)i - 4ij - (n+1)^2) \right\}$$

si dimostra dopo calcoli alquanto laboriosi che

$$\lim_{n\to\infty} \operatorname{Max}_{1\leq i;j\leq n} d_n^2 \quad (i,j) = 0$$

e

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i,j=1}^{n} d_n^2 \ (i,j) = \text{costante}$$

e pertanto la condizione (12) risulta soddisfatta.

Vi è da notare, infine, che la distribuzione asintotica di (G,τ) è anch'essa normale dato che il coefficiente di correlazione tra R e τ per $n\to\infty$ converge a 1.

TAVOLA 1

Numero termini	Coefficiente di correlazione tra		
	(G,τ)	(G,R)	(R,τ)
2	1,0000	1,0000	1,0000
3	0,9847	1,0000	0,9847
4	0,9624	0,9814	0,9806
5	0,9628	0,9827	0,9798
6	0,9607	0,9801	0,9802
7	0,9619	0,9805	0,9810
8	0,9622	0,9799	0,9820
9	0,9633	0,9800	0,9829
10	0,9640	0,9798	0,9839
11	0,9649	0,9799	0,9847
12	0,9656	0,9798	0,9855
13	0,9663	0,9798	0,9862
14	0,9670	0,9798	0,9869
15	0,9676	0,9798	0,9875
20	0,9699	0,9798	0,9899
30	0,9727	0,9798	0,9927
40	0,9743	0,9798	0,9943
50	0,9753	0,9798	0,9954
100	0,9774	0,9798	0,9976
1000	0,9795	0,9798	0,9997

Pavia, Università PIETRO MULIERE