PIETRO MULIERE

UN TEST PER L'OMOGENEITA' TRA DUE CAMPIONI BASATO SULL'INDICE DI COGRADUAZIONE DI GINI

Estratto dal
GIORNALE DEGLI ECONOMISTI E ANNALI DI ECONOMIA
Settembre-Ottobre 1977

CISALPINO - GOLIARDICA
Via Bassini 17-2
20133 MILANO

UN TEST PER L'OMOGENEITÀ TRA DUE CAMPIONI BASATO SULL'INDICE DI COGRADUAZIONE DI GINI (*)

1. - Introduzione.

Siano $Y_1, \ldots, Y_m; Y_{m+1}, \ldots, Y_{m+n}, N = m + n$, due campioni indipendenti, estratti da due popolazioni aventi funzione di ripartizione continua $F(y)$ e $G(y)$ rispettivamente. Un problema che assume rilevanza notevole nella statistica metodologica è quello del controllo dell'ipotesi:

$$H_0 : F(y) \equiv G(y) \tag{1}$$

vale a dire dell'ipotesi di omogeneità tra due campioni.

Un tale problema può essere affrontato e risolto ricorrendo a due diverse metodologie note come parametrica e non parametrica. Nella prima si assume che sia nota la forma analitica della funzione di ripartizione della popolazione e che non siano noti uno o più parametri della stessa. Nella seconda, invece, non si assume la forma analitica della funzione di ripartizione. Nella formulazione parametrica si ricorre a test quali quello di Student o di Snedecor e tali criteri sono fortemente limitati dall'ipotesi di normalità della legge di distribuzione. I criteri non parametrici più noti sono quelli che vanno sotto il nome di Smirnov-Kolmogorov, Wald-Wolfowitz, Wilcoxon - Mann - Whitney ed altri di più recente formulazione. La loro applicazione suppone solo la continuità delle funzioni di ripartizione senza ulteriori limitazioni sulla forma analitica delle stesse.

Se il test dà una risposta significativa, allora tra tutte le ipotesi alternative che si possono considerare ($F(y) \neq G(y)$) una abbastanza generale è la seguente:

$$H_1 : G(y) = F(y - \Delta) \qquad \text{per ogni } y \tag{2}$$

dove la sola differenza tra le due popolazioni è dovuta al parametro di posizione.

In questa classe di alternative il problema dell'omogeneità dei due campioni equivale a provare l'ipotesi $H_0 : \Delta = 0$ contro le specificazioni:

$$\Delta \neq 0 \qquad \text{(test bilaterale)}$$

$$\left.\begin{array}{l} \Delta > 0 \\ \Delta < 0 \end{array}\right\} \qquad \text{(test unilaterale)}$$

Subordinatamente alla suddetta ipotesi alternativa nel caso in cui l'ipotesi di nullità $\Delta = 0$ venga rifiutata sorge il problema della stima del parametro Δ.

* Desidero ringraziare il Prof. D. M. Cifarelli dell'Università di Pavia per le utili indicazioni ed osservazioni critiche.

Sotto l'ipotesi di normalità della distribuzione F lo stimatore di Δ è dato dal classico $\overline{\Delta} = \overline{Y}_1 - \overline{Y}_2$ dove $\overline{Y}_1$ e $\overline{Y}_2$ sono le medie dei due campioni. Tale stimatore porta a provare l'ipotesi (1) con il test t di Student. Con tale stimatore la stima di Δ dipende in modo marcato dai valori osservati e quindi in presenza di valori eccezionali, cioè valori fortemente devianti dalla massa dei dati, la stima ne verrà influenzata. Per ovviare a tale inconveniente si possono seguire due vie alternative, una consistente nell'eliminare le osservazioni devianti, l'altra nell'utilizzare criteri basati sui ranghi. In questa seconda direzione vanno gli stimatori proposti da Hodges-Lehmann [1, 1963] che si basa sul criterio di Wilcoxon, da Rao, Schuster e Littell [2, 1975] tramite l'indice di Smirnov-Kolmogorov, e di altri autori che si basano su indici diversi.

P. K. Sen [3, 1968] ottiene lo stimatore di Hodges-Lehmann come caso particolare della stima del coefficiente di regressione nel modello lineare semplice. Lo stimatore β^*, viene determinato utilizzando una misura della cograduazione tra $Y_i - b\,x_i$ e x_i, $i = 1, 2, \ldots, N$ data dall'indice τ di Kendall. Lo stimatore a cui giunge questo autore è quello di Theil [4, 1950] ossia la mediana delle $\binom{N}{2}$ pendenze

$$P_{ij} = \frac{Y_j - Y_i}{X_j - X_i} \qquad i < j \tag{3}$$

oppure il valore di mezzo dell'intervallo mediano nel caso di un numero pari di pendenze. Nel caso particolare in cui $x_1 = \ldots = x_m = 0$ e $x_{m+1} = \ldots = x_N = 1$ dove $(N = m + n)$ allora lo stimatore è dato dalla mediana delle mn differenze $(Y_j - Y_i)$ $i = 1, 2, \ldots m$; $j = m + 1, \ldots, N$ e coincide con quello proposto da Hodges-Lehmann.

Il suddetto problema può essere formulato anche nel modo seguente.

Immaginiamo che ogni y_i $i = 1, \ldots, N$ sia l'osservazione su Y_i con funzione di ripartizione $F(y - \Delta\, x_i)$ dipendente da x_i $i = 1, \ldots, N$ e che x_i possa assumere valore 0 e 1. Più precisamente supponiamo che:

$$x_1 = x_2 = \ldots = x_m = 0\,; \qquad x_{m+1} = \ldots = x_N = 1\,.$$

Controllare l'ipotesi che i due campioni $Y_1, \ldots, Y_m$ e $Y_{m+1}, \ldots, Y_N$ provengono dalla stessa distribuzione $F(y)$ (cioè $\Delta = 0$) equivale allora a controllare l'ipotesi di mancanza di concordanza o discordanza tra $Y_1, Y_2, \ldots, Y_N$ e $x_1, x_2, \ldots, x_N$.

Scopo del nostro lavoro, oltre che fornire un test per controllare l'ipotesi di omogeneità, è quello di formulare uno stimatore (puntuale e per intervallo) del parametro Δ nel caso di rifiuto dell'ipotesi di nullità.

Tra i tanti modi in cui è possibile misurare la concordanza tra $Y_1, \ldots, Y_N$ e $x_1, \ldots, x_N$ abbiamo scelto un indice di concordanza o discordanza tra ranghi, cioè un indice di cograduazione basato su quello di Gini.

Se $G_{m,n}$ è il valore di tale indice di cograduazione l'ipotesi di omogeneità verrà rifiutata quando $|G_{m,n}|$ è grande, vale a dire quando tra le due successioni vi è accentuata graduazione o contrograduazione. In tal caso, il procedimento che adotteremo per la stima di Δ ripeterà, eliminando l'ipotesi che le x_i siano tutte distinte, quello proposto da D.M. Cifarelli [5, s.d.] e consisterà nel trovare il valore Δ tale che $G_{m,n}(\Delta) \simeq 0$.

Nel paragrafo 2 costruiremo il test $G_{m,n}$ e ne esemplificheremo il suo calcolo. Nel paragrafo 3 calcoleremo la media e la varianza, ovviamente nel caso dell'ipotesi di nullità, e daremo la distribuzione asintotica dell'indice. Nel paragrafo 4 definiremo lo stimatore del parametro di posizione e vedremo come costruire l'intervallo di confidenza. Il paragrafo 5, infine, fornisce lo schema per un rapido calcolo della stima per ogni realizzazione campionaria.

2. - Il test.

Sia $(X_1, Y_1)\ (X_2, Y_2), \ldots (X_N, Y_N)$, $N \geqslant 2$ un campione di N elementi estratto da una popolazione bidimensionale (X, Y). Un indice per controllare la mancanza o meno di graduazione tra i due caratteri è dato dalla seguente espressione:

$$G = \frac{\sum_{i=1}^{N} \left\{ |N+1-R(X_i)-R(Y_i)| - |R(X_i)-R(Y_i)| \right\}}{\sum_{i=1}^{N} \left\{ |N+1-R^*(X_i)-R^*(Y_i)| - |R^*(X_i)-R^*(Y_i)| \right\}} \tag{4}$$

dove $R(X_i)$, $R(Y_i)$, $i = 1, 2, \ldots, N$ indicano rispettivamente il rango della i.esima osservazione della X e della i.esima osservazione della Y, mentre $R^*(X_i)$ e $R^*(Y_i)$ sono i ranghi delle osservazioni X e Y associate in termini di perfetta cograduazione (contrograduazione).

Tale indice consiste nel ragguagliare il numeratore al valore massimo che può raggiungere e può essere utilizzato anche nel caso in cui vi siano osservazioni coincidenti [6, 1939]. Nel caso di osservazioni tutte distinte il denominatore si ridurrebbe, come è noto, a:

$$\sum_{i=1}^{N} \{|N+1-2i|\} = \begin{cases} \dfrac{N^2}{2} & \text{per } N \text{ pari} \\ \dfrac{N^2-1}{2} & \text{per } N \text{ dispari} \end{cases}$$

Nel caso di osservazioni coincidenti conveniamo di assegnare loro come rango la media dei ranghi che avrebbero avuto se fossero state distinte.

Nel problema in esame vi sono osservazioni coincidenti nella sola variabile X e precisamente:

$$X_1 = X_2 = \ldots = X_m = 0 \text{ e } X_{m+1} = \ldots = X_N = 1$$

indicando le X_i l'indice di appartenenza di Y_i ad uno o all'altro dei due campioni $(Y_1, \ldots, Y_m)$, $(Y_{m+1}, \ldots, Y_N)$.

Il rango delle prime m osservazioni (X) sarà $\dfrac{m+1}{2}$ mentre il rango delle rimanenti $(N-m)$ sarà $\dfrac{n+2m+1}{2}$

L'indice assume la seguente espressione:

$$G_{m,n} = \frac{1}{D} \sum_{i=1}^{N} \left\{ |a_i - R(Y_i)| - |b_i - R(Y_i)| \right\} \tag{5}$$

dove $R(Y_i)$ $i = 1, 2, \ldots N$ indica il rango della i.esima osservazione della Y, e

$$a_i = \begin{cases} N+1-\dfrac{m+1}{2} & 1 \leqslant i \leqslant m \\ N+1-\dfrac{n+2m+1}{2} & m+1 \leqslant i \leqslant N \end{cases}$$

$$b_i = \begin{cases} \dfrac{m+1}{2} & 1 \leqslant i \leqslant m \\ \dfrac{n+2m+1}{2} & m+1 \leqslant i \leqslant N \end{cases} \tag{6}$$

D è il valore che assume $\sum_{i=1}^{N}\{|a_i - R(Y_i)| - |b_i - R(Y_i)|\}$ in caso di perfetta concordanza o discordanza.

Nel nostro lavoro d'ora innanzi supporremo che l'ampiezza dei due campioni sia uguale, ossia $N = 2n$. In tal caso D assume i seguenti valori:

$$D = \begin{cases} \dfrac{3n^2+1}{2} & \text{per } n \text{ dispari} \\ \dfrac{3n^2}{2} & \text{per } n \text{ pari} \end{cases} \tag{7}$$

L'indice $G_{n,n}$ che soddisfa la seguente doppia diseguaglianza

$$-1 \leqslant G_{n,n} \leqslant 1 \tag{8}$$

assumerà il valore $+1$ quando tutte le osservazioni del secondo campione sono superiori a quelle del primo ed il valore -1 nel caso in cui tutti gli elementi del primo campione sono superiori a quelli del secondo.

La regione critica dell'ipotesi $H_0 : F(y) = G(y)$ contro l'alternativa $H_1 : F(y) \neq G(y)$ è del tipo

$$|G_{n,n}| \geqslant G_{n,n}^{(\alpha)} \tag{9}$$

dove $G_{n,n}^{(\alpha)}$ è la soglia critica determinata con l'ausilio delle tavole della distribuzione $G_{n,n}$, nell'ipotesi di nullità, con probabilità di errore di prima specie non superiore ad α, cioè tale che

$$P\{|G_{n,n}| \geqslant G_{n,n}^{(\alpha)}\} \leqslant \alpha \tag{10}$$

oppure, volendo una regione critica di dimensione esattamente eguale ad α, si potrà seguire il normale procedimento di casualizzazione.

In parole, se il valore osservato della funzione $G_{n,n}$ è grande in modulo è plausibile pensare che le distribuzioni delle due popolazioni siano differenti.

Prima di terminare questo paragrafo illustriamo il test con un esempio. Siano, con $n = 4$

$$\underline{y}_1 = (27;\ 31;\ 33;\ 45)$$
$$\underline{y}_2 = (24;\ 30;\ 15;\ 21)$$

i due campioni.

Ordinando tutti gli elementi in un'unica successione si ha:

$$Z = (15\,;\ 21\,;\ 24\,;\ 27\,;\ 30\,;\ 31\,;\ 33\,;\ 45)$$

da cui

$$R_1 = 4\,;\quad R_2 = 6\,;\quad R_3 = 7\,;\quad R_4 = 8\,;\quad R_5 = 3\,;\quad R_6 = 5\,;\quad R_7 = 1\,;\quad R_8 = 2$$

sono i ranghi delle osservazioni Y.

Inoltre la (6) e la (7) forniscono

$$a_i = \begin{cases} \dfrac{3n+1}{2} = 6{,}5 & 1 \leqslant i \leqslant 4 \\ \dfrac{n+1}{2} = 2{,}5 & 5 \leqslant i \leqslant 8 \end{cases}$$

$$b_i = \begin{cases} \dfrac{n+1}{2} = 2{,}5 & 1 \leqslant i \leqslant 4 \\ \dfrac{3n+1}{2} = 6{,}5 & 5 \leqslant i \leqslant 8 \end{cases}$$

$$D = \frac{3n^2}{2} = 24$$

e quindi

$$G_{4,4} = -\frac{20}{24} = -0{,}8\overline{3}\,.$$

$|G_{4,4}|$ è abbastanza grande da far ritenere plausibile che i due campioni provengano da distribuzioni differenti. Verificheremo nel paragrafo 3, con l'ausilio delle tavole, che al livello $\alpha = 0{,}06$, le osservazioni contraddicono l'ipotesi H_o.

3. - Considerazioni intorno alla distribuzione di $G_{n,n}$.

Iniziamo con il determinare la media dell'indice nell'ipotesi di indifferenza ($F = G$). A tal uopo, posto

$$\begin{aligned} A_i &= \sum_{i=1}^{N} |\, a_i - R(Y_i)\,| \\ B_i &= \sum_{i=1}^{N} |\, b_i - R(Y_i)\,| \end{aligned} \tag{11}$$

il loro valor medio è dato da:

$$\begin{aligned} E(A_i) &= \sum_{i=1}^{N} \left\{ \frac{1}{N} \sum_{r=1}^{N} |\, a_i - r\,| \right\} \\ E(B_i) &= \sum_{i=1}^{N} \left\{ \frac{1}{N} \sum_{r=1}^{N} |\, b_i - r\,| \right\} \end{aligned} \tag{12}$$

essendo

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

$$P\{R_i = r;\ R_h = s\} = \frac{1}{N(N-1)} \qquad 1 \leqslant i \neq h;\ r \neq s \leqslant N \tag{13}$$

Si ha poi

$$E(A_i) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{N(N+1)}{2} - [(N+1) - a_i]\, a_i \right\}$$

$$E(B_i) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{N(N+1)}{2} - [(N+1) - b_i]\, b_i \right\} \tag{14}$$

e sostituendo i valori di a_i e b_i forniti dalla (6) si ottiene

$$E(A_i) = E(B_i) \tag{15}$$

e pertanto ne deriva che

$$E(G_{n,n}) = \frac{1}{D} E\{A_i - B_i\} = 0 \tag{16}$$

Per la varianza di $G_{n,n}$ si ha:

$$E(G_{n,n}^2) = \frac{1}{D^2} E\{ \sum_{i=1}^{N} [\,|a_i - R(Y_i)| - |b_i - R(Y_i)|\,]^2 + \\ + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} [\,|a_i - R(Y_i)| - |b_i - R(Y_i)|\,]\,[\,|a_j - R(Y_j)| - |b_j - R(Y_j)|\,]\} \tag{17}$$

ossia

$$E(G_{n,n}^2) = \\ = \frac{1}{D^2} E\{ \sum_{i=1}^{N} (a_i - R(Y_i))^2 + \sum_{i=1}^{N} (b_i - R(Y_i))^2 - 2 \sum_{i=1}^{N} [\,|a_i - R(Y_i)| \cdot |b_i - R(Y_i)|\,] + \\ + \sum_{i \neq j} [\,|a_i - R(Y_i)|\ |a_j - R(Y_j)|\,] - \sum_{i \neq j} [\,|a_i - R(Y_i)|\ |b_j - R(Y_j)|\,] + \\ - \sum_{i \neq j} [\,|b_i - R(Y_i)|\ |a_j - R(Y_j)|\,] + \sum_{i \neq j} [\,|b_i - R(Y_i)|\ |b_j - R(Y_j)|\,]\}. \tag{18}$$

e dopo qualche calcolo, utilizzando la (13), otteniamo i seguenti risultati:

$$\mathrm{Var}(G_{n,n}) = \begin{cases} \dfrac{16\, n^2 (2\, n^2 + 1)}{3\,(2\, n - 1)\,(3\, n^2 + 1)^2} & \text{per } n \text{ dispari} \\ \dfrac{8\,(2\, n + 1)}{27\, n^2} & \text{per } n \text{ pari} \end{cases} \tag{19}$$

Nella Tavola A sono riportate le funzioni di probabilità e di ripartizione di

$$G_{n,n} D = \sum_{i=1}^{N} \{\,|a_i - R(Y_i)| - |b_i - R(Y_i)|\,\} \quad \text{per} \quad N = 4, 6, 8, 10 \quad (1).$$

(1) Per la realizzazione del programma ringrazio la dott.ssa Rossella Masserizzi del Centro di Calcolo dell'Università « Luigi Bocconi » di Milano.

Poichè le distribuzioni sono simmetriche, ci siamo limitati a considerare soltanto i valori non negativi. Verifichiamo ora se nell'esempio del paragrafo 2 dobbiamo o meno rifiutare l'ipotesi H_0 con livello di significatività $\alpha = 0{,}06$.

Dalle tavole si ottiene il valore critico $G_{n,n}^{(\alpha)} = 20/24$ che è il più grande intero che soddisfa la (10). Poichè il valore osservato $G_{n,n}$ è uguale a $G_{n,n}^{(\alpha)}$ rifiutiamo l'ipotesi nulla $(F(Y) = G(Y))$ $(\Rightarrow F(Y) \neq G(Y))$.

Per quanto riguarda la distribuzione asintotica di $G_{n,n}$, sempre nel caso di nullità, osserviamo che con leggere modifiche, dovute alla differente forma dell'indice, possiamo seguire il ragionamento fatto in [5, s.d.] e dimostrare:

$$\text{a)}\quad \sqrt{n}\,(G_{n,n} - \hat{G}) \to 0 \qquad \text{in probabilità} \tag{20}$$

essendo

$$\hat{G} = \frac{n}{D} \sum_{i=1}^{N} \left\{ \left| \frac{a_i}{n} - F(Y_i) \right| - \left| \frac{b_i}{n} - F(Y_i) \right| \right\} \tag{21}$$

e $Y_1, Y_2, \ldots, Y_N$, una successione di variabili mutuamente indipendenti e somiglianti con funzione di ripartizione assolutamente continua F.

b) $\sqrt{n}\,\hat{G}$ converge in legge verso la variabile normale con media nulla e varianza 16/27.

Da ciò per la (20) si deduce che anche $\sqrt{n}\,G_{n,n}$ è asintoticamente normale con media nulla e varianza 16/27.

4. Stima del parametro di posizione.

Nel caso in cui si rifiuti l'ipotesi $H_0 : F(y) = G(y)$ e sia plausibile considerare come alternativa l'ipotesi (2) $H_1 : G(y) = F(y - \Delta)$ sorge il problema della stima del parametro Δ.

Per definire lo stimatore del parametro di posizione impiegheremo la stessa funzione utilizzata nella prova delle ipotesi.

Siano $Y_1, Y_2, \ldots, Y_N$, $N = 2n$ variabili mutuamente indipendenti con funzione di ripartizione

$$P\{Y_i \leqslant y\} = F_i(y) = F(y - \Delta x_i) \qquad i = 1, 2, \ldots, N; \quad N \geqslant 2$$

dove F è una qualunque funzione di ripartizione continua e le x sono le N costanti che assumono valore 0 per $1 \leqslant i \leqslant m$ e valore 1 per $m + 1 \leqslant i \leqslant N$, con Δ reale qualunque.

Consideriamo le nuove variabili casuali

$$U_i(\Delta) = Y_i - \Delta x_i \qquad i = 1, 2, \ldots, N$$

e misuriamo la cograduazione tra $Y_1 - \Delta x_1, \ldots, Y_N - \Delta x_N$ e $x_1, \ldots, x_N$ utilizzando l'indice

$$G_{n,n}(\underline{Y}, \Delta) = G_{n,n}(\Delta) = \frac{1}{D} \sum_{i=1}^{N} \{ |a_i - R(U_i(\Delta))| - |b_i - R(U_i(\Delta))| \} \tag{22}$$

in cui a_i e b_i assumono i valori dati dalla (6) e $R(U_i(\Delta))$ indica il rango di $U_i(\Delta)$ nella successione ordinata di $\{U_1(\Delta), U_2(\Delta), \ldots, U_N(\Delta)\}$ e D assume i valori dati nella (7).

La funzione di $\Delta\, G_{n,n}(\Delta)$ non risulta definita nell'insieme

$$A = \{Y_j - Y_i = \Delta \qquad i = 1, 2, \ldots, n; \quad j = n+1, \ldots, N\}$$

che indicheremo con

$$A = \{\Delta^{(1)}, \Delta^{(2)} \ldots \Delta^{(r)}\} \qquad 1 \leqslant r \leqslant n^2$$

con $\Delta^{(1)} < \Delta^{(2)} < \ldots < \Delta^{(r)}$

L'insieme A è evidentemente finito.

Poichè le variabili $U_i(\Delta)$ sono indipendenti e somiglianti, per ipotesi, la distribuzione di $G_{n,n}(\Delta)$ avrà la stessa distribuzione di $G_{n,n}$ data in precedenza e pertanto nel caso indifferente

$$E\{G_{n,n}(\Delta)\} = 0$$

Diventa allora abbastanza naturale proporre di stimare il parametro Δ in modo tale che $G_{n,n}(\Delta)$ risulti il più possibile prossimo a zero, cioè in modo che

$$G_{n,n}(\Delta) \simeq 0 \tag{23}$$

Le seguenti proposizioni sono facilmente dimostrabili seguendo l'impostazione data in [5, s.d.] tenendo conto delle opportune modifiche subite dall'indice.

Proposizione 1. - La funzione $G_{n,n}(\Delta)$, qualunque sia l'ennupla $\underline{Y}$ è costante all'interno degli intervalli

$$(-\infty\ \Delta^{(1)}), \qquad (\Delta^{(1)}\ \Delta^{(2)}), \ldots, (\Delta^{(r)} + \infty)$$

Proposizione 2. - Per ogni ennupla $\underline{Y}$ la funzione $G_{n,n}(\Delta)$ è non crescente.

Dalla proposizione 1 si deduce che:

$$G_{n,n}(\Delta) = \text{Massimo valore} \leqslant 1 \qquad -\infty < \Delta < \Delta^{(1)}$$

$$G_{n,n}(\Delta) = \text{Minimo valore} \geqslant -1 \qquad \Delta^{(r)} < \Delta < +\infty$$

e

$$\lim_{\Delta \to -\infty} G_{n,n} = 1 \qquad \lim_{\Delta \to +\infty} G_{n,n} = -1$$

Converremo in seguito di porre

$$G_{n,n}(\Delta^{(i)}) = \lim_{\Delta \to \Delta^{(i)}_{+}} G_{n,n}(\Delta)$$

e rendere così continue da destra tutte le realizzazioni.

La proposizione 2 ci permette di definire lo stimatore. Infatti qualunque sia l'ennupla osservata $\underline{Y}$ si potranno verificare i due casi alternativi seguenti:

a) esiste tutto un intervallo di valori di Δ per cui $G_{n,n}(\Delta) = 0$

b) esistono due intervalli consecutivi I_1, I_2 tali che

$$G_{n,n}(\Delta) > 0 \qquad \Delta \in I_1$$

$$G_{n,n}(\Delta) < 0 \qquad \Delta \in I_2$$

Nel primo caso si potrà utilizzare come stima di Δ il valore centrale dell'intervallo mentre nel secondo

$$\operatorname{Sup}_{\Delta} \{\Delta : G_{n,n}(\Delta) > 0\} = \operatorname{Inf}_{\Delta} \{\Delta : G_{n,n}(\Delta) < 0\}$$

e compattando i due casi avremo come stimatore di Δ la quantità:

$$\tilde{\Delta} = \frac{1}{2} [\operatorname{Sup}_{\Delta} \{\Delta : G_{n,n}(\Delta) > 0\} + \operatorname{Inf}_{\Delta} \{\Delta : G_{n,n}(\Delta) < 0\}] \tag{24}$$

Lo studio delle proprietà dello stimatore $\tilde{\Delta}$ sarà oggetto di un nostro prossimo lavoro, in cui effettueremo, tra l'altro, alcuni confronti di efficienza con altri stimatori già richiamati nell'introduzione ed usati allo stesso scopo. Tuttavia, ci sembra interessante dare alcune proprietà quasi immediate di $\tilde{\Delta}$.

Una semplice ma utile proposizione che riguarda lo stimatore è la seguente:

$$\tilde{\Delta}(Y_1, \ldots, Y_n, Y_{n+1} + c, , , , , , Y_N + c) = \tilde{\Delta}(Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_N) + c \tag{25}$$

per ogni reale c.

Tale proprietà è una immediata conseguenza della definizione dello stimatore $\tilde{\Delta}$.

Dalla (25) segue che

$$P_{\Delta}\{(\tilde{\Delta} - \Delta) \leqslant t\} = P_0(\Delta \leqslant t) \tag{26}$$

dove la notazione P_{Δ} indica che la probabilità è calcolata assumendo Δ come vero valore del parametro.

La (26) mostra che la distribuzione di $(\tilde{\Delta} - \Delta)$ è indipendente dal parametro Δ e pertanto questa circostanza ci permetterà, nello studio delle proprietà di $\tilde{\Delta}$, di porre $\Delta = 0$ senza perdere in generalità.

La funzione $G_{n,n}$ soddisfa la relazione di invarianza

$$G(Y_1 + c, \ldots, Y_n + c \; ; \; Y_{n+1} + c, \ldots, Y_N + c) = G(Y_1, \ldots, Y_n \; ; \; Y_{n+1}, \ldots, Y_N) \quad \text{per ogni } c \tag{27}$$

ed è ovvio che da ciò discende la medesima relazione per $\tilde{\Delta}$.

Notiamo, inoltre, che se le variabili $Y_1, Y_2, \ldots, Y_N$, hanno distribuzioni simmetriche, allora $\tilde{\Delta}$ è uno stimatore non distorto di Δ ossia

$$E(\tilde{\Delta}) = \Delta \tag{28}$$

Per la dimostrazione assumiamo, per la (26), $\Delta = 0$ ed osserviamo che

$$\tilde{\Delta}(-\underline{Y}) = -\tilde{\Delta}(\underline{Y}) \tag{29}$$

allora

$$\tilde{\Delta}(-\underline{Y}) = \frac{1}{2} [\operatorname{Sup}_{\Delta} \{\Delta : G(-\underline{Y}, \Delta) > 0\} + \operatorname{Inf}_{\Delta} \{\Delta : G(-\underline{Y}, \Delta) < 0\}]$$

$$= \frac{1}{2} [\operatorname{Sup}_{\Delta} \{\Delta : G(\underline{Y}, -\Delta) > 0\} + \operatorname{Inf}_{\Delta} \{\Delta : G(\underline{Y}, -\Delta) < 0\}]$$

$$= \frac{1}{2}\left[-\inf_{\Delta}\{\Delta : G(\underline{Y}, \Delta) < 0\} - \sup_{\Delta}\{\Delta : G(\underline{Y}, \Delta) > 0\}\right]$$

$$= -\tilde{\Delta}(\underline{Y}).$$

Per la simmetria delle variabili $Y_1, Y_2, \ldots, Y_N$, $\tilde{\Delta}(\underline{Y})$ e $\tilde{\Delta}(-\underline{Y})$ avranno la stessa distribuzione. Dalla (29) discende che $\tilde{\Delta}(\underline{Y})$ ha la distribuzione simmetrica rispetto a zero (valore di Δ) e questo implica la (28).

Per quanto riguarda l'intervallo di confidenza di Δ possiamo procedere nel modo seguente.

Essendo le variabili $U_i(\Delta)$ egualmente distribuite ed indipendenti per ipotesi $G_{n,n}(\Delta)$ si distribuirà come l'indice $G_{n,n}$ nel caso indifferente (vedi Tavola A) e potremo allora determinare un valore $G_{n,n}$ tale che (con α opportuno)

$$P\{-G^*_{n,n} \leqslant G_{n,n}(\Delta) \leqslant +G^*_{n,n}\} = 1-\alpha \tag{30}$$

dove $0 < \alpha < 1$.

Definendo

$$\tilde{\Delta}_U = \sup\{\Delta : G_{n,n}(\Delta) \geqslant -G^*_{n,n}\}$$

$$\tilde{\Delta}_L = \inf\{\Delta : G_{n,n}(\Delta) \leqslant \quad G^*_{n,n}\} \tag{31}$$

discende (vedi [5, s.d.])

$$P\{\tilde{\Delta}_L < \Delta < \tilde{\Delta}_U\} = 1-\alpha \tag{32}$$

Pertanto, la coppia di funzioni $(\tilde{\Delta}_L, \tilde{\Delta}_U)$ costituiscono gli estremi dell'intervallo di confidenza di Δ.

5. - Schema di calcolo.

Poichè non sembra possibile fornire una espressione esplicita dello stimatore $\tilde{\Delta}$, ci sembra utile predisporre un opportuno schema di calcolo.

Lo schema che qui proponiamo ricalca con varianti non irrilevanti quello proposto in [5, s.d.].

Siano $Y_1, Y_2, \ldots, Y_N$, $N = 2n$, le osservazioni corrispondenti a $x_1, x_2, \ldots, x_N$ dove le x_i assumono valore 0 per $1 \leqslant i \leqslant n$ e valore 1 per $n+1 \leqslant i \leqslant N$.

Calcoliamo le n^2 differenze

$$\Delta_{ij} = (Y_j - Y_i) \qquad i = 1, 2, \ldots, n \; ; \; j = n+1, \ldots, 2n$$

non necessariamente tutte distinte ed ordiniamole in modo crescente

$$\Delta^{(1)}_{ij} \; ; \; \Delta^{(2)}_{ij} \; ; \; \ldots \; ; \; \Delta^{(r)}_{ij} \qquad 1 \leqslant r \leqslant n^2$$

Costruiamo ora una tabella che contenga i ranghi di $U_i(\Delta)$ al variare di Δ negli intervalli

$$(-\infty\, \Delta^{(1)}) \, . \, (\Delta^{(1)} \Delta^{(2)}) \, . \ldots . \, (\Delta^{(r)} + \infty)$$

partendo dalla considerazione che il rango di $U_j(\Delta)$ è uguale al numero di osservazioni che sono al di sotto o sulla retta con coefficiente angolare Δ e passante per (X_j, Y_j).

Per meglio illustrare tale schema riprendiamo l'esempio del paragrafo 2.

Dati i due campioni, con $n = 4$

$$\underline{y_1} = (27\,;\,31\,;\,33\,;\,45)$$
$$\underline{y_2} = (24\,;\,30\,;\,15\,;\,21)$$

le 16 differenze Δ_{ij} sono date dalla seguente tabella:

		y_j			
		$y_5 = 24$	$y_6 = 30$	$y_7 = 15$	$y_8 = 21$
y_i	$y_1 = 27$	— 3	3	— 12	— 6
	$y_2 = 31$	— 7	— 1	— 16	— 10
	$y_3 = 33$	— 9	— 3	— 18	— 12
	$y_4 = 45$	— 21	— 15	— 30	— 24

Ordinando le differenze in una successione si ha:

$$\Delta^{(1)}_{47}\ ;\ \Delta^{(2)}_{48}\ ;\ \Delta^{(3)}_{45}\ ;\ \Delta^{(4)}_{37}\ ;\ \Delta^{(5)}_{27}\ ;\ \Delta^{(6)}_{46}\ ;\ \Delta^{(7)}_{38} = \Delta^{(7)}_{17}\ ;\ \Delta^{8}_{28}\ ;\ \Delta^{(9)}_{35}\ ;\ \Delta^{(10)}_{25}$$
$$\Delta^{(11)}_{18}\ ;\ \Delta^{(12)}_{36} = \Delta^{(12)}_{15}\ ;\ \Delta^{(13)}_{26}\ ;\ \Delta^{(14)}_{16}\ .$$

Costruiamo una tabella con $N = 2n$ ossia 8 righe nel modo seguente.

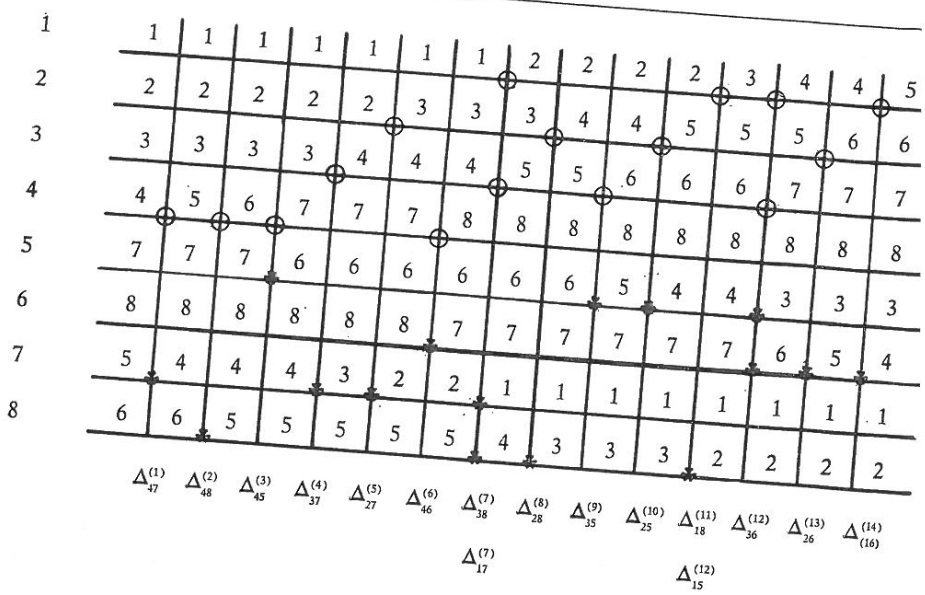
In corrispondenza ad ogni differenza Δ_{ij} innalziamo una verticale e contrassegnamo con un circoletto ($\bigcirc$) la riga i.ma e con un asterisco (*) la riga j.ma.

Si calcoli il rango di $Y_1, Y_2, \ldots, Y_n$ e si dispongano nelle prime n righe della prima colonna $(R(Y_1), R(Y_2), \ldots, R(Y_n))$.

Si calcoli il rango degli elementi del secondo campione assegnando rango $(n+1)$ al più piccolo e via via fino ad assegnare $2n$ al più grande. $(n + R(Y_j)$ con $j = n+1, \ldots, 2n)$.

In tal modo la prima colonna risulta:

$$R(Y_1)$$
$$R(Y_2)$$
$$\cdot$$
$$\cdot$$
$$R(Y_n)$$
$$n + R(Y_{n+1})$$
$$\cdot$$
$$\cdot$$
$$n + R(Y_{2n})$$



Passando da una colonna all'altra abbiamo incrementato o diminuito di una unità il numero della colonna precedente a seconda che abbiamo incontrato un circoletto o un asterisco (2).

I numeri che si leggono nella h.ma riga sono i ranghi di $Y_h - \Delta x_h$ al variare di Δ negli intervalli tra le differenze. Naturalmente per la definizione di $G_{n,n}(\Delta)$ in ogni colonna vi è una permutazione di $(1, 2, \ldots, 8)$.

Predisposta la tabella dei $R(U_i(\Delta))$ diventa molto agevole costruirne altre due e determinare così $\sum_{i=1}^{N} |a_i - R(U_i(\Delta))|$ e $\sum_{i=1}^{N} |b_i - R(U_i(\Delta))|$.

Nel nostro esempio dato che

$$a_i = \begin{cases} 6{,}5 & 1 \leqslant i \leqslant 4 \\ 2{,}5 & 5 \leqslant i \leqslant 8 \end{cases}$$

$$b_i = \begin{cases} 2{,}5 & 1 \leqslant i \leqslant 4 \\ 6{,}5 & 5 \leqslant i \leqslant 8 \end{cases}$$

le due tabelle sono rispettivamente:

---

(2) Se uno stesso incrocio è interessato da più circoletti e/o asterischi si incrementerà il numero letto nella colonna precedente del numero di circoletti e si diminuirà del numero di asterischi.

5,5	5,5	5,5	5,5	5,5	5,5	5,5	4,5	4,5	4,5	4,5	3,5	2,5	2,5	1,5
4,5	4,5	4,5	4,5	4,5	3,5	3,5	3,5	2,5	2,5	1,5	1,5	1,5	0,5	0,5
3,5	3,5	3,5	3,5	2,5	2,5	2,5	1,5	1,5	0,5	0,5	0,5	0,5	0,5	0,5
2,5	1,5	0,5	0,5	0,5	0,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5
4,5	4,5	4,5	3,5	3,5	3,5	3,5	3,5	3,5	2,5	1,5	1,5	0,5	0,5	0,5
5,5	5,5	5,5	5,5	5,5	5,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5
2,5	1,5	1,5	1,5	0,5	0,5	0,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5
3,5	3,5	2,5	2,5	2,5	2,5	2,5	1,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
32	30	28	27	25	24	24	22	20	18	16	15	13	12	11

Il totale, per colonna, rappresenta l'andamento di $\sum_{i=1}^{N} \{ \mid a_i - R(U_i(\Delta)) \mid \}$ al variare di Δ.

1,5	1,5	1,5	1,5	1,5	1,5	1,5	0,5	0,5	0,5	0,5	0,5	1,5	1,5	2,5
0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	1,5	1,5	2,5	2,5	2,5	3,5	3,5
0,5	0,5	0,5	0,5	1,5	1,5	1,5	2,5	2,5	3,5	3,5	3,5	4,5	4,5	4,5
1,5	2,5	3,5	4,5	4,5	4,5	5,5	5,5	5,5	5,5	5,5	5,5	5,5	5,5	5,5
0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	1,5	2,5	2,5	3,5	3,5	3,5
1,5	1,5	1,5	1,5	1,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
1,5	2,5	2,5	2,5	3,5	4,5	4,5	5,5	5,5	5,5	5,5	5,5	5,5	5,5	5,5
0,5	0,5	1,5	1,5	1,5	1,5	1,5	2,5	3,5	3,5	3,5	4,5	4,5	4,5	4,5
8	10	12	13	15	15	16	18	20	22	24	25	28	29	30

Il totale, per colonna, rappresenta l'andamento di $\sum_{i=1}^{N} \{ \mid b_i - R(U_i(\Delta)) \mid \}$ al variare di Δ.

L'analisi dei totali precedenti conduce immediatamente al valore di $\tilde{\Delta}$.

Nel nostro esempio $G_{n,n}(\Delta)$ è uguale a zero in tutto un intervallo, ossia per

$$-10 < \Delta < -9$$

e la stima di Δ risulta il valore centrale dell'intervallo, $\tilde{\Delta} = -9{,}5$.

Per quanto riguarda l'intervallo di confidenza di Δ, occorre determinare $\tilde{\Delta}_U$ e $\tilde{\Delta}_L$.

Osserviamo che dalle tavole della distribuzione di $G_{n,n}$, nel caso indifferente, si ha:

$$P\{-16/24 \leqslant G_{n,n} \leqslant 16/24\} = 66/70 \simeq 0{,}94$$

con l'aiuto delle tabelle precedenti si deduce immediatamente

$$\tilde{\Delta}_L = \text{Inf}\,\{\Delta : G_{n,n}(\Delta) \leqslant 16/24\} = -21$$

$$\Delta_U = \text{Sup}\,\{\Delta : G_{n,n}(\Delta) \geqslant -16/24\} = -1$$

e l'intervallo di confidenza di Δ con coefficiente di confidenza $1-\alpha = 94\%$, qualunque sia la funzione di ripartizione F, è dato da

$$-21 < \Delta < -1\,.$$

Pavia, Università

PIETRO MULIERE

RIFERIMENTI BIBLIOGRAFICI

[1] HODGES J. L. j., LEHMANN E. L., *Estimates of Location Based on Rank Tests.* « Annals of Mathematical Statistics », 34, 1963, pp. 598-611.

[2] RAO P. V., SCHUSTER E. F., LITTEL R. C., *Estimation of Shift and Center of Simmetry Based on Kolmogorov-Smirnov Statistics.* « Annals of Statistics », 3, 1975, pp. 862-873.

[3] SEN P. K., *Estimates of the Regression Coefficient Based on Kendall's Tau.* « Journal of the American Statistical Association, 63, 1968, pp. 1379-89.

[4] THEIL H., *A Rank Invariant Method of Linear and Polynomial Regression Analysis*, I, II, III. (« Nederlandische Akad. Wetensch. Proc. », 53, 1950, pp. 386-392; 521-525; 1397-1412).

[5] CIFARELLI D. M., *La stima del coefficiente di regressione mediante l'indice di cograduazione di Gini.* (In corso di stampa).

[6] SALVEMINI T., *L'indice di cograduazione del Gini nel caso di serie statistiche con ripetizione.* « Metron », 13, 1939, pp. 27-39.

TAVOLA A - *Funzioni di probabilità e di ripartizione di* $G_{n,n}\ D = \sum_{i=1}^{N} \{ |a_i - R(Y_i)| - |b_i - R(Y_i)| \}$ *per argomenti non negativi.*

$N = 4$

$G_{n,n}\ D$	Funzione di probabilità	Funzione di ripartizione
0	$0,\overline{3}$	0,66
2	$0,1\overline{6}$	0,83
6	$0,1\overline{6}$	1,00

$N = 6$

$G_{n,n}\ D$	Funzione di probabilità	Funzione di ripartizione
2	0,20	0,70
6	0,20	0,90
10	0,05	0,95
14	0,05	1,00

$N = 8$

$G_{n,n}\ D$	Funzione di probabilità	Funzione di ripartizione
0	0,14285714	0,57142857
2	0,05714286	0,62857143
4	0,05714286	0,68571429
6	0,05714286	0,74285715
8	0,07142858	0,81428573
10	0,05714286	0,87142859
12	0,01428571	0,88571430
14	0,05714286	0,94285716
16	0,02857142	0,97142858
20	0,01428571	0,98571429
24	0,01428571	1,00000000

$N = 10$

$G_{n,n}\ D$	Funzione di probabilità	Funzione di ripartizione
2	0,10714286	0,60714286
6	0,10714286	0,71428572
10	0,09523809	0,80952381
14	0,05952381	0,86904762
18	0,05952381	0,92857143
22	0,02777777	0,95634920
26	0,02777777	0,98412697
30	0,00793651	0,99206348
34	0,00396826	0,99603174
38	0,00396826	1,00000000