ISSN: 2037-6367

MatematicaMente

Pubblicazione mensile della sezione veronese della MATHESIS – Società Italiana di Scienze Matematiche e Fisiche – Fondata nel 1895 – Autorizzazione del Tribunale di Verona n. 1360 del 15 – 03 – 1999 – I diritti d'autore sono riservati. Direttore: Luciano Corso - Redazione: Alberto Burato, Elisabetta Capotosto, Carlo Marchiori, Giovanna Tessari – Via IV Novembre, 11/b – 37126 Verona – tel e fax (045) 8344785 – 338 6416432 – e-mail: lcorso@iol.it – Stampa in proprio - Numero 164 – Pubblicato il 07 – 12 – 2011

L'indice di eterogeneità del Gini e una mia variante «fuzzy»

di Luciano Corso

Corrado Gini (1884 - 1965) fu il più importante statistico italiano della prima metà del Novecento. Durante il Fascismo ebbe un ruolo importante nella politica demografica italiana e dette contributi notevoli allo sviluppo della statistica. I più importanti contributi scientifici del Gini si trovano nelle misure della variabilità. È noto il rapporto di concentrazione del Gini. Un po' meno noti sono altri indici da lui stesso inventati. Tra questi, uno molto interessante è l'indice di eterogeneità.

Consideriamo un carattere statistico X che può assumere le modalità x_1 con frequenza assoluta n_1, x_2 con frequenza assoluta n_2, … , x_k con frequenza assoluta n_k, in relazione a un certo esperimento. Confrontiamo quindi queste modalità tra di loro mediante la seguente metrica (distanza)

$$I(x_i,x_j)=\begin{cases}0 & se \quad x_i=x_j\\ 1 & se \quad x_i\neq x_j\end{cases} \qquad \forall i,j\,. \tag{1}$$

Costruiamo la seguente matrice degli $I(x_i, x_j)$:

$$\begin{pmatrix} I(x_i,x_j) & x_1 & x_2 & ... & x_j & ... & x_k\\ x_1 & 0 & 1 & ... & 1 & ... & 1\\ x_2 & 1 & 0 & ... & 1 & ... & 1\\ ... & ... & ... & ... & ... & ... & ...\\ x_i & 1 & 1 & ... & 0 & ... & 1\\ ... & ... & ... & ... & ... & ... & ...\\ x_k & 1 & 1 & ... & 1 & ... & 0 \end{pmatrix}. \tag{2}$$

Siccome i confronti sono a due a due, essi vanno ponderati con il prodotto delle frequenze assolute associate alle modalità x_i e x_j. Calcoliamo quindi la media aritmetica ponderata delle distanze così ottenute:

$$M\left(I(x_i,x_j)\right)=\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}I(x_i,x_j)\cdot n_i\cdot n_j}{\sum_{i=1}^{k}\sum_{j=1}^{k}n_i\cdot n_j}=$$

$$=\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}I(x_i,x_j)\cdot n_i\cdot n_j}{n^2}, \tag{3}$$

dove

$$n=\sum_{i=1}^{k}n_i\,. \tag{4}$$

Da (2) e (3), si ottiene [B.1]:

$$M\left(I(x_i,x_j)\right)=\frac{1}{n^2}\cdot[0\cdot n_1\cdot n_1+1\cdot n_1\cdot n_2+...+1\cdot n_1\cdot n_k+$$

$$+1\cdot n_2\cdot n_1+0\cdot n_2\cdot n_2+1\cdot n_2\cdot n_3+...+1\cdot n_2\cdot n_k+...$$

$$...+1\cdot n_{k-1}\cdot n_1+1\cdot n_{k-1}\cdot n_2+...+0\cdot n_{k-1}\cdot n_{k-1}+...+0\cdot n_k\cdot n_k]=$$

$$=\frac{1}{n^2}\cdot[n_1\cdot(n-n_1)+n_2\cdot(n-n_2)+...+n_k\cdot(n-n_k)]=$$

$$=\frac{1}{n^2}\cdot\left[\sum_{i=1}^{k}n_i\cdot(n-n_i)\right]=\sum_{i=1}^{k}\frac{n_i}{n}\cdot\left(1-\frac{n_i}{n}\right)=$$

$$=\sum_{i=1}^{k}\frac{n_i}{n}-\sum_{i=1}^{k}\frac{n_i}{n}\cdot\frac{n_i}{n}=1-\sum_{i=1}^{k}f_i^2\,, \tag{5}$$

dove $f_i = n_i / n$ è la frequenza relativa alla modalità i-esima. Questo risultato è detto indice di eterogeneità del Gini e lo si scrive convenzionalmente con la lettera G:

$$G=1-\sum_{i=1}^{k}f_i^2\,. \tag{6}$$

È facile normalizzare (6). Infatti, il suo massimo è dato da

$$Max[G]=1-\sum_{i=1}^{k}\left(\frac{1}{k}\right)^2=1-\frac{k}{k^2}=\frac{k-1}{k}\,. \tag{7}$$

Infatti il massimo di G deve corrispondere al massimo dell'eterogenietà e ciò accade quando tutte le modalità x_j (j = 1, 2, …, k) presentano la stessa frequenza relativa $1/k$. L'indice normalizzato, quindi, corrisponde al rapporto di (6) con (7). Si ottiene:

$$G'=\frac{G}{Max(G)}=\left(1-\sum_{i=1}^{k}f_i^2\right)\cdot\frac{k}{k-1}\,. \tag{8}$$

Supponiamo, ora, di avere n = 4 celle ciascuna delle quali occupate da 4 , 3 , 2 , 1 particelle (Fig. 2). Diamo di seguito la matrice delle distanze con associata la frequenza assoluta congiunta $n_i \cdot n_j$ di ogni coppia di modalità a confronto:

$$\begin{pmatrix} I(x_i,x_j) & x_1 & x_2 & x_3 & x_4\\ x_1 & (0,16) & (1,12) & (1,8) & (1,4)\\ x_2 & (1,12) & (0,9) & (1,6) & (1,3)\\ x_3 & (1,8) & (1,6) & (0,4) & (1,2)\\ x_4 & (1,4) & (1,3) & (1,2) & (0,1) \end{pmatrix} \tag{9}$$

Confrontando i risultati che si ottengono applicando (7) e (8) con quelli che si determinano applicando l'entropia e l'entropia normalizzata [si veda B.4] si scopre che le differenze sono minime e che entrambe le misure portano alle stesse conclusioni. Tuttavia il numero di operazioni elementari coinvolte aumenta se si usa l'entropia di Shannon.

Le due misure del Gini sono:

$$G=1-\sum_{i=1}^{k}f_i^2=1-\left[\left(\frac{4}{10}\right)^2+\left(\frac{3}{10}\right)^2+\left(\frac{2}{10}\right)^2+\left(\frac{1}{10}\right)^2\right]=\frac{70}{100},$$

$$G'=\frac{70}{100}\cdot\frac{4}{3}=\frac{28}{30}\approx 0{,}93333. \tag{10}$$

La relazione per il calcolo dell'entropia di Shannon [B.1 e 2] è

$$H=\frac{1}{Ln2}\cdot\sum_{i=1}^{k}p_i\cdot Ln\frac{1}{p_i}\,, \tag{11}$$

dove a p_i sostituiremo f_i. Otteniamo:

$$H=\frac{1}{Ln2}\cdot\left(\frac{4}{10}\cdot Ln\frac{10}{4}+\frac{3}{10}\cdot Ln\frac{10}{3}+\frac{2}{10}\cdot Ln\frac{10}{2}+\frac{1}{10}\cdot Ln\frac{10}{1}\right)\approx$$

$$\approx 1{,}84644.$$

L'entropia normalizzata è data da

$$Max[H]=\frac{1}{Ln2}\cdot\left(4\cdot\frac{1}{4}\cdot Ln\,4\right)=2,$$

$$H'=H/Max[H]\approx 0{,}92322. \tag{12}$$

Il numero di operazioni che abbiamo fatto per il calcolo dei due indici è: |*operazione Gini*|=12; |*operazioni entropia*|=20. Questo semplice esempio ci fa capire perché conviene usare l'indice di eterogeneità del Gini, invece dell'entropia di Shannon, quando si vuole misurare l'eterogeneità delle modalità con cui

si presenta un carattere statistico.
La Fig. 2 dà la rappresentazione grafica di ciò che abbiamo calcolato.

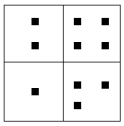


Fig. 2. Nella figura sono presentate le 4 celle dello spazio piano che contengono le 10 particelle; quest'ultime si trovano dislocate nelle varie cellette come indicato dalla figura. Dal punto di vista dell'eterogeneità ogni rotazione di questo spazio, che conservi la posizione delle particelle all'interno delle celle, ha associata la stessa entropia e lo stesso valore dell'indice normalizzato del Gini.

Eterogeneità "Fuzzy"

Diamo per noti i fondamenti della teoria *fuzzy-set* (grado di appartenenza, logica e insiemi *fuzzy*, ecc.), rinviando per eventuali chiarimenti o approfondimenti al testo [B.3] richiamato in bibliografia.

Supponiamo ora che un gruppo di n oggetti sia sottoposto a indagine rispetto a un carattere statistico X. Sia x_i la valutazione (qualitativa o quantitativa) attribuibile all'oggetto i con riferimento alla peculiarità X. Definiamo la funzione

$$Y:\left(x_i,x_j\right)\mapsto[0,1]; \tag{13}$$

essa misura il grado di dissomiglianza di x_i rispetto a x_j. Con Y possiamo comparare tutti gli n oggetti del gruppo a due a due. La misura Y è:

$$Y\left(x_i,x_j\right)=\begin{cases}0 & se \quad x_i=x_j\\ 0<Y\le 1 & se \quad x_i\ne x_j\end{cases} \qquad \forall i,j\,. \tag{14}$$

La matrice (2), in questo caso, diventa:

$$\begin{pmatrix} Y\left(x_i,x_j\right) & x_1 & x_2 & \dots & x_j & \dots & x_n\\ x_1 & 0 & y_{12} & \dots & y_{1j} & \dots & y_{1n}\\ x_2 & y_{21} & 0 & \dots & y_{2j} & \dots & y_{2n}\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ x_i & y_{i1} & y_{i2} & \dots & 0 & \dots & y_{in}\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ x_n & y_{n1} & y_{n2} & \dots & y_{nj} & \dots & 0 \end{pmatrix} \tag{15}$$

dove $0\le y_{ij}\le 1$.
L'indice di Gini subisce una variazione sia nella sua relazione generale (3), sia nella sua interpretazione. La nuova misura tiene conto di una logica non dicotomica e diventa

$$M\left(Y\left(x_i,x_j\right)\right)=\frac{1}{n^2}\cdot\sum_{i=1}^{n}\sum_{j=1}^{n}y_{i,j}\quad, \tag{16}$$

se ogni individuo presenta una valutazione x_i diversa dagli altri. Se, invece, il numero delle modalità distinte osservate nel gruppo rispetto al carattere X è k, allora il gruppo viene diviso in k parti e la matrice (15) diventa

$$\begin{pmatrix} Y\left(c_i,c_j\right) & c_1 & c_2 & \dots & c_j & \dots & c_k\\ c_1 & 0 & y_{12} & \dots & y_{1j} & \dots & y_{1k}\\ c_2 & y_{21} & 0 & \dots & y_{2j} & \dots & y_{2k}\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ c_i & y_{i1} & y_{i2} & \dots & 0 & \dots & y_{ik}\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ c_k & y_{k1} & y_{k2} & \dots & y_{kj} & \dots & 0 \end{pmatrix} \tag{17}$$

dove c_i è la modalità relativa al gruppo di oggetti i.
In corrispondenza della matrice (17), la (16) diventa

$$M\left(Y\left(c_i,c_j\right)\right)=\frac{1}{n^2}\cdot\sum_{i=1}^{k}\sum_{j=1}^{k}y_{i,j}\cdot n_i\cdot n_j\quad. \tag{18}$$

La (16) e la (18) assumono un significato più ampio di quello assunto dalla (3); esse, infatti, valgono sia nel caso di una logica dicotomica, sia in quello di una logica *fuzzy*.

Come esempio, consideriamo $n = 100$ individui e le loro 100 valutazioni x_i, con riferimento al carattere X. Supponiamo, inoltre, che gli n individui, dopo essere stati valutati, presentino $k = 4$ valutazioni distinte c_1, c_2, c_3, c_4 con frequenze assolute associate pari a n_1=10, n_2=20, n_3=45, n_4=25 e tali che $n_1 + n_2 + n_3 + n_4 = 100$ (per la (4)). Confrontando a due a due tali individui in dettaglio, si compone la tabella (19) che presenta i gradi di dissomiglianza sperimentali ottenuti da questo confronto.

$$\begin{pmatrix} Y\left(c_i,c_j\right) & c_1 & c_2 & c_3 & c_4\\ c_1 & 0 & 0{,}4 & 0{,}3 & 0{,}8\\ c_2 & 0{,}4 & 0 & 0{,}6 & 0{,}5\\ c_3 & 0{,}3 & 0{,}6 & 0 & 0{,}1\\ c_4 & 0{,}8 & 0{,}5 & 0{,}1 & 0 \end{pmatrix}\,. \tag{19}$$

Con la (18) determiniamo la media dei gradi di dissomiglianza $M(c_{ij})$ e assegniamo a questa media il simbolo C:

$$\begin{aligned} C = (\quad & 0\cdot10\cdot10 + 0{,}4\cdot10\cdot20 + 0{,}3\cdot10\cdot45 + 0{,}8\cdot10\cdot25 +\\ & + 0{,}4\cdot20\cdot10 + 0\cdot20\cdot20 + 0{,}6\cdot20\cdot45 + 0{,}5\cdot20\cdot25 +\\ & + 0{,}3\ 45\ 10 + 0{,}6\ 45\ 20 + 0\ 45\ 45 + 0{,}1\ 45\ 25 +\\ & + 0{,}8\cdot25\cdot10 + 0{,}5\cdot25\cdot20 + 0{,}1\cdot25\cdot45 + 0\cdot25\cdot25) / 100^2 = \\ = {} & 0{,}2635. \end{aligned}$$

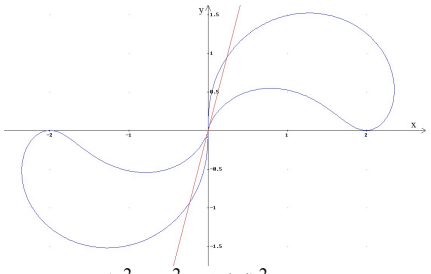
Poiché il valore massimo che può assumere C è, anche in questo caso, $(k-1)/k$, otteniamo subito il suo valore normalizzato:

$C' = 0{,}2635\cdot(4/3) \approx 0{,}3513.$

Il valore suggerisce che le modalità osservate presentano uno scarso livello di dissomiglianza.

Bibliografia: [B.1] Piccolo Domenico, Statistica, ed. Il Mulino, Bologna, 2010. [B.2] Ballatori Enzo, *Statistica e metodologia della ricerca*, Galeno editrice, Perugia, 1988. [B.3] Dumitrescu D., Lazzerini B. Jain L. C., *Fuzzy sets and their application to clustering and training*, CRC Press, Boca Raton (Florida), 2000. [B.4] Corso Luciano (2011), *Entropia come misura dell'eterogeneità*, Periodico di matematiche numero 2 Mag-Ago 2011, Volume 3 Serie XI, Anno CXXI, pag. 97 e segg., Ed. Mathesis, Caserta UNINA II.

Biflecnodo



$$(x^2+y^2-2\cdot|x|)^2 = x\cdot y$$

Tangenti nel biflecnodo: $x = 0$ e $y = 4\cdot x$. (*di Nazario Magnarelli*)

Non c'è scatto nel cielo

di Mariangela Gualtieri

Non c'è scatto nel cielo.
Solo il fulmine ha spigoli e fuoco.
Solo il fulmine viaggia nervoso.
Ma guarda ora - che pace.
A me pare di averlo percorso tutto a volo
questo azzurro che si dispiega pacato.
Mi pare un luogo che conosco.
Che è stato di me.
E lo è ancora.
Se guardo - entra nella radice
dà da bere al mio
alimenta il mio fuoco.