Rivista di Statistica Applicata Vol. 19 n. 4 Dicembre 1986

ROSTRA

Rubrica a cura di Benito V. Frosini

Su una variante della differenza media assoluta

Benito V. Frosini, Università Cattolica del Sacro Cuore – Milano

1. Premessa

Mi è capitato più volte, negli ultimi anni, che studenti e anche colleghi mi sottoponessero alcune idee concernenti la differenza media assoluta, e più precisamente una variante della stessa formula che viene suggerita in modo naturale da una delle sue possibili scritture (la (3) più sotto). Con mia sorpresa, ho trovato una certa resistenza da parte dei proponenti ad accettare le mie osservazioni, che sostanzialmente escludevano tale variante dall'ambito degli indici di variabilità, come comunemente intesi. Oltre a ciò, appariva che nessuno degli interlocutori era a conoscenza di un lavoro di G. Leti (1967) in cui tale variante viene discussa correttamente, sia pure senza derivarne in modo esplicito tutte le implicazioni dal punto di vista applicativo; per la verità, il volume della Biblioteca del "Metron" in cui appare questo lavoro (come pure altri della stessa serie) non esiste in molte biblioteche. Vi è quindi qualche ulteriore motivo per questa breve nota.

2. Sulle caratterizzazioni degli indici di variabilità

Non vi è dubbio che molte incomprensioni fra i maggiori esponenti della Statistica italiana, a cominciare dalla diatriba che vedeva contrapposti negli anni 1914–16 C. Gini e G. Pietra da un lato, e C. Bresciani Turroni e U. Ricci dall'altro, sono in buona parte derivate dall'assoluta

mancanza di definizioni operative generali, cioè concernenti le distribuzioni e non i singoli indici (cfr. Avondo Bodino, 1963, p. 2). Come si può infatti parlare di indici di variabilità senza possedere un criterio oggettivo per riconoscere se un dato indice è un *valido* indice di variabilità, oppure no? Questa istanza è ovviamente preliminare a quella del riconoscimento dei particolari aspetti della variabilità che un dato indice riesce ad evidenziare. Né si può sottacere che il più grande motivo di confusione è stata la commistione inestricabile fra i concetti di variabilità e di concentrazione.

Ogni caratterizzazione degli indici di variabilità (e lo stesso uso della parola *variabilità* in questo contesto) ha ovviamente qualcosa di arbitrario, ed è forse inevitabile che porti ad escludere da questa categoria anche indici che sono usualmente indicati con questa denominazione (cfr. Herzel, 1967; Frosini, 1986, pp. 129–131, 164–165). Confido comunque che i lettori possano essere d'accordo almeno con la prima delle due seguenti definizioni, che a motivo della successiva applicazione vengono riferite a due serie di n valori $x_1 \leq x_2 \leq \ldots \leq x_n$ e $y_1 \leq y_2 \leq \ldots \leq y_n$ (per un riferimento generale a variabili casuali vedi Frosini, 1981 e 1984).

Variabilità globale – La variabile statistica $X = (x_1, \ldots, x_n)$ è detta *più variabile* (globalmente) di $Y = (y_1, \ldots, y_n)$, se vale $|x_i - x_j| \geq |y_i - y_j|$ per ogni i, j, con diseguaglianza stretta per qualche i, j; in tal caso, per ogni indice di variabilità V deve essere $V(X) \geq V(Y)$.

Dispersione – La variabile statistica $X = (x_1, \ldots, x_n)$ ha *maggiore dispersione* di $Y = (y_1, \ldots, y_n)$ *intorno all'i-esimo valore* se vale $|x_i - x_j| \geq |y_i - y_j|$ per $j = 1, \ldots, n$, con diseguaglianza stretta per qualche j; in tal caso, per ogni indice di dispersione D deve essere $D(X) \geq D(Y)$ (cfr. Frosini, 1986, pp. 143 e 160).

Sembra evidente che la prima definizione debba essere accolta per ogni indice di variabilità: se tutte le distanze fra valori x_i sono maggiori o al più uguali alle corrispondenti distanze fra valori y_i, qualunque indice di variabilità deve essere coerente col suddetto ordinamento delle distribuzioni X e Y. D'altra parte, in tale situazione anche un indice di dispersione deve rispettare l'ordinamento, cioè anche gli indici di dispersione sono indici di variabilità (secondo le suddette definizioni).

3. La differenza media secondo le modalità

Come è noto, la differenza media assoluta Δ è definita da

$$\Delta = \sum_{i,j}^{n} |x_i - x_j| / (n^2 - n); \tag{1}$$

se la serie $x_1, \ldots, x_n$ (con $x_i \leq x_{i+1}$) dà luogo alla seriazione dei valori

crescenti $x'_1, \ldots x'_k$ con rispettive frequenze assolute $n_1, \ldots, n_k$, dall'eguaglianza

$$\sum_{i,j}^{n} |x_i - x_j| = \sum_{i,j}^{k} |x'_i - x'_j| n_i n_j$$

si ricava la scrittura di Δ

$$\Delta = \sum_{i,j}^{k} |x'_i - x'_j| n_i n_j / (n^2 - n) \tag{2}$$

e anche

$$\Delta = \sum_{i \neq j}^{k} |x'_i - x'_j| n_i n_j / (n^2 - n). \tag{3}$$

Le scritture (1), (2) e (3) sono equivalenti; tuttavia la (3) può far sorgere una curiosità, se non una pretesa: non sarebbe meglio calcolare una media aritmetica ponderata dei valori non nulli $|x'_i - x'_j|$ con rispettivi pesi $n_i n_j$? Tenuto conto che

$$\sum_{i \neq j}^{k} n_i n_j = \sum_{i,j}^{k} n_i n_j - \sum_{i=j}^{k} n_i n_j = n^2 - \sum_{i=1}^{k} n_i^2,$$

una tale media si scriverebbe

$$\Delta^* = \sum_{i \neq j}^{k} |x'_i - x'_j| n_i n_j / \left(n^2 - \sum_{i=1}^{k} n_i^2 \right). \tag{4}$$

È questa la "differenza media secondo le modalità" studiata da Leti (1967, §6), come uno degli indici che rispondono alla domanda: "Di quanto ciascuna unità della popolazione è in media diversa, rispetto al carattere considerato, da ogni altra unità nella quale il carattere presenta modalità diversa?".

Si osserva subito che se $k = n$, ovvero le osservazoni sono tutte distinte, vale l'eguaglianza $\Delta^* = \Delta$; in tutti gli altri casi $\Delta^* > \Delta$. Rinviando a Leti (1967) per ciò che riguarda alcune interessanti relazioni concernenti Δ^*, preme qui sottolineare che Δ^* *non è un indice di variabilità* secondo le definizioni date più sopra. Lo si intuisce osservando che se due valori coincidenti vengono distanziati, con conseguente aumento della variabilità (riflesso correttamente da un aumento del numeratore), il denominatore aumenta anch'esso, e può chiaramente aumentare anche in proporzione maggiore del numeratore; variazioni opposte avvengono se valori diversi vengono portati a coincidere. Ad esempio, dati i quattro valori 0,9; 1; 2;

2,1, si calcola $\Delta = \Delta^* = 9,2/12 = 0,7\overline{6}$. Passando a una nuova distribuzione che assume i due soli valori 1 e 2, entrambi con frequenza 2, la variabilità della distribuzione è diminuita; ciò è correttamente riflesso dal ridotto valore di $\Delta = 8/12 = 0,\overline{6}$, mentre $\Delta^* = 8/8 = 1$ è aumentato.

Si può notare che Δ^* resta un indice di variabilità se (condizione sufficiente) i confronti avvengono fra distribuzioni in cui restano immutate le frequenze $n_1, \ldots, n_k$; in tal caso si ha infatti che il denominatore resta costante. Ma in generale, come si è detto, Δ^* non può essere impiegato come indice di variabilità. Il suo uso (come quello di indici analoghi) deve essere limitato a quei casi in cui si voglia rispondere alla domanda di Leti più sopra citata. Avendo escluso di usare per un tale indice la denominazione di "indice di variabilità", si può ricavare dallo stesso articolo di Leti la denominazione di "indice di variabilità secondo le modalità", da contrapporre a quella di "indice di variabilità secondo le unità" o più semplicemente "indice di variabilità".

Riferimenti bibliografici

Avondo Bodino G., Su alcune fondamentali questioni riguardanti la misura della variabilità, *Rendiconti dell'Istituto di Statistica*, Università L. Bocconi, Milano, 1963.
Frosini B.V., Sul concetto e sulla misura della variabilità, *Atti del Convegno 1981 della Società Italiana di Statistica*, vol. 2, 11–35.
Frosini B.V., *Characterizations of variability measures*, Istituto di Statistica dell'Università Cattolica del S. Cuore di Milano, Serie E.P., N. 1, 1984.
Frosini B.V., *Lezioni di Statistica, Parte prima*, Vita e Pensiero, Milano, 1986.
Herzel A., Considerazioni sulla variabilità, Vol. III di *Note e commenti*, Serie C della biblioteca del *Metron*, Roma, 1967, 187–203.
Leti G., Sugli indici di variabilità e di mutabilità basati sul confronto a due a due delle unità del collettivo, Vol. III di *Note e commenti*, Serie C della biblioteca del *Metron*, Roma, 1967, 45–69.

Summary

If the denominator of the mean difference in formula (3) is replaced by the sum of *weights* $n_i n_j$, for $i \neq j$, the new formula (4) comes up, with interesting features, both formal and regarding specific applications, as Leti (1967) already pointed out.

However, it must be noted that – according to a widely accepted characterization of variability measures – the above mentioned formula cannot be used as a measure of variability of a distribution, because, for example, it can increase as some positive differences $|x_i - x_j|$ are reduced to zero, without modifying any other difference.