STATISTICA, anno LXII, n. 1, 2002

SCOMPOSIZIONE DELLA DISPERSIONE PER VARIABILI STATISTICHE ORDINALI

L. Grilli, C. Rampichini

1. L'INDICE DI DISPERSIONE D

Indichiamo con Y una variabile statistica ordinale, rilevata su una popolazione $\mathcal{U}$ di dimensione N, con supporto $S_Y = \{y_1, y_2, \ldots, y_K\}$, frequenze relative cumulate $F_k = \frac{n_1 + n_2 + \ldots + n_k}{N}$ e frequenze relative retrocumulate $F'_k = 1 - F_{k-1}$ ($k = 1, 2, \ldots, K$), dove n_i è la frequenza della i-ma modalità.

Leti (1983, pp. 290-297) propone la seguente misura della dispersione nel caso di variabile ordinale:

$$D^* = \sum_{k=1}^{K} [F_k(1 - F_k) + F'_k(1 - F'_k)] = 2 \sum_{k=1}^{K-1} F_k(1 - F_k) \qquad (1.1)$$

Tale indice consente di tener conto dell'informazione sull'ordinamento delle modalità, assume valore zero nel caso di minima dispersione, cioè nel caso in cui le frequenze siano tutte concentrate in un'unica modalità, ed è massimo quando la dispersione è massima, il che, per una variabile ordinale, accade quando le frequenze risultano concentrate nelle due modalità estreme y_1 e y_K. Il valore massimo di D dipende dal numero di modalità assumibili dalla variabile (numero di elementi del supporto) e dalla numerosità della popolazione osservata N (solo nel caso in cui N sia dispari). In particolare si ha:

$$D^*_{\max} = \begin{cases} \dfrac{K-1}{2} & \text{se N è pari} \\ \dfrac{K-1}{2}\left(1 - \dfrac{1}{N^2}\right) & \text{se } N \text{ è dispari} \end{cases}.$$

Per N sufficientemente grande, il valore massimo di D" può assumersi pari a $\frac{K-1}{2}$ qualunque sia N. Si definisce allora il seguente indice di dispersione relativo:

$$d = \frac{D^*}{\frac{K-1}{2}},$$

con $d \in [0, 1]$, a meno dell'approssimazione per il caso di N dispari.

Ai fini della nostra analisi il fattore moltiplicativo 2 presente nella (1.1) è superfluo e quindi dimostriamo la scomposizione relativamente all'indice D così definito:

$$D = D^*/2 = \sum_{k=1}^{K-1} F_k(1 - F_k). \tag{1.2}$$

2. SCOMPOSIZIONE DELLA DISPERSIONE

Si supponga ora che la popolazione $\mathcal{U}$ che stiamo esaminando sia suddivisa in J sottopopolazioni (gruppi) $\mathcal{U}_j$, di numerosità N_j, definite per esempio in termini di modalità assunte da una seconda variabile X, con supporto $S_X = \{x_1, x_2, \ldots, x_J\}$, in modo tale che le sottopopolazioni considerate formino una partizione della popolazione $\mathcal{U}$. Con riferimento a tale partizione possiamo definire J variabili condizionate $Y|\mathcal{U}_j$, con frequenze relative cumulate (funzione di ripartizione) $F_{k|j} = \frac{n_{1j} + n_{2j} + \ldots + n_{kj}}{N_j}$, $k = 1, 2, \ldots, K$, dove n_{ij} è la frequenza della i-ma modalità nella j-ma sottopopolazione. Si osservi che è possibile esprimere la funzione di ripartizione marginale come mistura delle distribuzioni condizionate:

$$F_k = \sum_{j=1}^{J} \pi_j F_{k|j}, \tag{2.1}$$

dove $\pi_j = N_j/N$ è la proporzione di unità statistiche appartenenti alla j-ma sottopopolazione.

Definiamo ora con D_j l'indice di dispersione (1.2) calcolato per la j-ma sottopopolazione:

$$D_j = \sum_{k=1}^{K-1} F_{k|j}(1 - F_{k|j}).$$

Dimostreremo ora che, analogamente a quanto accade per la varianza di variabili quantitative, l'indice D riferito all'intera popolazione può essere scomposto nella somma di due quote: la prima relativa alla dispersione di Y *interna* alle sottopopolazioni considerate e la seconda interpretabile come una misura della dispersione *tra* le sottopopolazioni.

Ricordando che le frequenze marginali F_k possono essere espresse in termini delle frequenze condizionate $F_{k|j}$ in base alla mistura (2.1), l'indice D diviene:

$$\boldsymbol{D} = \sum_{k=1}^{K-1} F_k(1-F_k) = \sum_{k=1}^{K-1}\sum_{j=1}^{J} \pi_j F_{k|j}(1-F_k) = \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K-1} F_{k|j}(1-F_k)$$

$$= \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K-1} F_{k|j}(1-F_{k|j}+F_{k|j}-F_k)$$

$$= \sum_{j=1}^{J} \pi_j \left[\sum_{k=1}^{K-1} F_{k|j}(1-F_{k|j}) + \sum_{k=1}^{K-1} F_{k|j}(F_{k|j}-F_k)\right]$$

$$= \sum_{j=1}^{J} \pi_j D_j + \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K-1} F_{k|j}(F_{k|j}-F_k). \qquad (2.2)$$

Si osservi che il primo termine della (2.2), che indichiamo con D_W, è una media ponderata degli indici di dispersione calcolati nelle J sottopopolazioni, con pesi pari a π_j, ed è pertanto una misura della dispersione media entro i gruppi. Resta da esaminare il secondo termine della (2.2), che indicheremo con il simbolo D_B:

$$D_B = \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K-1} F_{k|j}(F_{k|j}-F_k) = \sum_{k=1}^{K-1}\sum_{j=1}^{J} \pi_j F_{k|j}(F_{k|j}-F_k). \qquad (2.3)$$

Ricordando dalla (2.1) che, per ogni k, F_k rappresenta la media (ponderata) degli $F_{k|j}$, ognuno dei (K - 1) addendi della (2.3) è semplicemente la varianza degli $F_{k|j}$, si ha cioè

$$\sum_{j=1}^{J} \pi_j F_{k|j}(F_{k|j}-F_k) = \sum_{j=1}^{J} \pi_j (F_{k|j}-F_k)^2$$

Pertanto la (2.3) può scriversi come

$$D_B = \sum_{k=1}^{K-1}\sum_{j=1}^{J} \pi_j (F_{k|j}-F_k)^2 = \sum_{j=1}^{J} \pi_j \sum_{k=1}^{K-1} (F_{k|j}-F_k)^2 = \sum_{j=1}^{J} \pi_j Z_{2j}^2 ,$$

dove Z_{2j}^2 è il quadrato dell'indice quadratico di dissomiglianza per variabili ordinali (Leti, 1983, p. 531) tra la j-ma distribuzione condizionata e la distribuzione marginale. $\boldsymbol{D}_r$ può quindi essere interpretato come una misura della dispersione tra gruppi.

Si osservi che D_B è una quantità non negativa che assume il valore zero quando le distribuzioni condizionate sono tutte simili a quella marginale, cioè $F_{k|j} = F_k$ per ogni k (nel caso in cui la partizione sia indotta dalla variabile concornitante X, ciò equivale all'indipendenza stocastica fra Y e X). Un caso particolare di distribuzioni simili si ha quando la distribuzione marginale e, di conseguenza, tutte le distribuzioni condizionate sono degeneri (frequenze concentrate in un'unica modalità).

Viceversa, $\boldsymbol{D}_r$ coincide con $\boldsymbol{D}$ quando $D_j = 0$ per ogni j, cioè quando tutte le distribuzioni condizionate sono degeneri.

Si noti l'analogia, formale e interpretativa, della scomposizione proposta,

$$D = \sum_{j=1}^{J} \pi_j D_j + \sum_{j=1}^{J} \pi_j Z_{2j}^2 ,$$

con la nota scomposizione della varianza (Leti, 1783, p. 783):

$$\sigma^2 = \sum_{j=1}^{J} \pi_j \sigma_j^2 + \sum_{j=1}^{J} \pi_j (\overline{Y}_j - \overline{Y})^2 ,$$

dove σ_j^2 e $\overline{Y}_j$ sono, rispettivamente, la varianza e la media della j-ma sottopopolazione, mentre Y è la media generale.

Dati i risultati precedenti, risulta sensato costruire un indice che misuri quanta parte della dispersione della variabile ordinale Y sia spiegata dalla partizione considerata (ovvero dall'eventuale variabile X che definisce la partizione). In analogia con il rapporto di correlazione, tale indice può essere costruito come rapporto fra la quota di dispersione tra gruppi D_B e la dispersione totale D:

$$\delta = \frac{D_B}{D} , \qquad (2.4)$$

con $6 \in [0, 1]$.

3. UN ESEMPIO APPLICATIVO: VALUTAZIONE DELLA DIDATTICA

Consideriamo, a titolo di esempio, la variabile ordinale Y relativa al livello di soddisfazione globale degli studenti iscritti al primo anno presso la Facoltà di Ingegneria di Firenze, in merito ai corsi frequentati nel secondo semestre dell'a.a. 1999/2000. Tale variabile può assumere le seguenti 4 modalità, corrispondenti al giudizio espresso: 1 = *decisamente no;* 2 = *più no che si;* 3 = *più* si *che no;* 4 = *decisamente si.*

La tabella 1 riporta (in termini percentuali) le distribuzioni condizionate per corso e la distribuzione marginale della variabile Y, la proporzione di osservazioni per corso π_j ($j = 1, 2, \ldots, 30$) e i valori degli indici di dispersione D_j e D (si noti che in questa applicazione si ha $\frac{K-1}{4} = 0.75$, per cui tali indici assumono valori compresi tra 0 e 0.75).

L'indice D_j è massimo per il corso n. 26 (0.534), che presenta frequenze distribuite tra le quattro modalità, e minimo per il corso n. 28 (0.188), per il quale i giudizi sono tutti positivi e concentrati nella quarta modalità. Una misura sintetica della dispersione *entro* i corsi è data dall'indice D_W, che risulta pari a 0.433, mentre la dispersione *tra* i corsi si ottiene sottraendo dal valore globale $D = 0.508$ il valore di D_W, per cui $D_B = 0.075$. La proporzione di dispersione legata al corso, secondo la (2.4), è dunque $\delta = 0.148$. Questo valore piuttosto basso significa che, nello spiegare le differenze nei giudizi formulati sul corso, la variabilità individuale gioca un ruolo preminente rispetto a quella legata alle caratteristiche del corso.

Dipartimento di Statistica "G. Parenti",
Università degli Studi di Firenze

LEONARDO GRILLI
CARLA RAMPICHINI

TABELLA 1

Soddisfazione globale del corso. Ingegneria primo anno, Ateneo di Firenze – A.A. 1999/2000, II semestre. Distribuzioni percentuali condizionate, percentuali di osservazioni per corso e indici di dispersione

	giudizio					
corso	1	2	3	4	π_i	D_i
1	18.60	41.86	25.58	13.95	4.93	0.511
2	10.61	30.30	43.94	15.15	7.56	0.465
3	17.24	51.72	24.14	6.90	3.32	0.421
4	10.00	26.00	34.00	30.00	5.73	0.530
5	17.65	52.94	17.65	11.76	1.95	0.451
6	11.43	14.29	40.00	34.29	4.01	0.518
7	5.88	11.76	47.06	35.29	1.95	0.429
8	20.93	30.23	37.21	11.63	4.93	0.518
9	3.03	12.12	59.09	25.76	7.56	0.349
10	0	0	50.00	50.00	0.46	0.250
11	0	0	60.00	40.00	0.57	0.240
12	4.17	12.50	45.83	37.50	2.75	0.413
13	13.64	50.00	27.27	9.09	2.52	0.432
14	12.50	33.33	45.83	8.33	2.75	0.434
15	4.55	18.18	54.55	22.73	2.52	0.395
16	2.94	19.12	45.58	32.35	7.79	0.419
17	20.00	31.11	35.56	13.33	5.15	0.525
18	0	8.11	29.72	62.16	4.24	0.310
19	25.58	6.98	58.14	9.30	4.93	0.494
20	15.71	50.00	30.00	4.29	8.02	0.399
21	0	25.00	68.75	6.25	1.83	0.246
22	10.00	50.00	40.00	0	1.15	0.330
23	15.38	34.62	38.46	11.54	2.98	0.182
24	0	7.41	55.56	37.04	3.09	0.302
25	25.00	50.00	25.00	0	0.46	0.375
26	50.00	19.23	23.08	7.69	2.98	0.554
27	22.22	33.33	44.44	0	1.03	0.120
28	0	0	25.00	75.00	0.92	0.188
29	0	50.00	37.50	12.50	0.92	0.359
30	0	44.44	33.33	22.22	1.03	0.420
TOT	12.14	26.92	40.32	20.62	100.00	0.508

RIFERIMENTI BIBLIOGRAFICI

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

RIASSUNTO

Scomposizione della dispersione per variabili statistiche ordinali

In questo lavoro si mostra come il noto principio della scomposizione della varianza entro e tra gruppi possa essere esteso al caso di un particolare indice di dispersione per variabili statistiche ordinali.

SUMMARY

Decomposition of dispersion for ordered variables

In the paper we show how the well-known principle of variance decomposition within and between groups can be extended to the case of a specific dispersion index for ordinal variables.