## La misura della variabilità per variabili qualitative ordinali

Lo studio della variabilità per variabili qualitative ordinali può essere condotto servendosi degli indici di omogeneità/eterogeneità introdotti per le variabili qualitative nominali. Tali indici, però, essendo basati sulle sole frequenze associate alle singole modalità, non sfruttano l'informazione ulteriore associata a tale tipo di variabili, ovvero la possibilità di ordinarne le modalità. A tal fine è possibile definire un'ulteriore accezione del termine variabilità tenendo conto di come le frequenze si dispongono rispetto alle modalità ordinate della variabile.

Si consideri, a titolo esemplificativo, una variabile ordinale $X$ che può assumere cinque modalità, osservata su un collettivo di $N = 100$ unità. Le seguenti tabelle riportano alcune possibili distribuzioni di tale variabile:

| **CASO 1** | |
|---|---|
| $X$ | $n$ |
| $x_{(1)}$ | 20 |
| $x_{(2)}$ | 20 |
| $x_{(3)}$ | 20 |
| $x_{(4)}$ | 20 |
| $x_{(5)}$ | 20 |
| | 100 |

| **CASO 2** | |
|---|---|
| $X$ | $n$ |
| $x_{(1)}$ | 0 |
| $x_{(2)}$ | 0 |
| $x_{(3)}$ | 0 |
| $x_{(4)}$ | 100 |
| $x_{(5)}$ | 0 |
| | 100 |

| **CASO 3** | |
|---|---|
| $X$ | $n$ |
| $x_{(1)}$ | 0 |
| $x_{(2)}$ | 50 |
| $x_{(3)}$ | 50 |
| $x_{(4)}$ | 0 |
| $x_{(5)}$ | 0 |
| | 100 |

| **CASO 4** | |
|---|---|
| $X$ | $n$ |
| $x_{(1)}$ | 50 |
| $x_{(2)}$ | 0 |
| $x_{(3)}$ | 0 |
| $x_{(4)}$ | 0 |
| $x_{(5)}$ | 50 |
| | 100 |

La distribuzione riportata come caso 1 è una distribuzione caratterizzata da massima variabilità (si parla più propriamente di massima eterogeneità o minima omogeneità) mentre quella indicata come caso 2 è caratterizzata dalla minima variabilità possibile (minima eterogeneità o massima omogeneità). Nel primo caso, infatti, le frequenze sono ripartite equamente tra tutte le modalità del carattere mentre nel secondo caso tutte le frequenze sono accentrate in una sola modalità del carattere, comportando quindi un'assenza totale di variabilità. Le due distribuzioni 3 e 4 riportano invece due casi intermedi di variabilità: in entrambi i casi le frequenze sono ripartite uniformemente tra due modalità del carattere; la misurazione degli indici di omogeneità/eterogeneità (Gini o Shannon) sulle due distribuzioni produrrà di conseguenza lo stesso valore. L'utilizzo di tali indici, non sfruttando alcuna informazione sull'ordinamento delle modalità, non permette di rilevare l'evidente diversità delle due distribuzioni: il caso 3, infatti, è caratterizzato dal fatto che le uniche due modalità presenti sono due modalità vicine alla modalità centrale mentre nel caso 4 le uniche due modalità presenti sono le due modalità estreme. Quest'ultima distribuzione è quindi caratterizzata da una maggiore dispersione (in particolare dalla massima dispersione possibile, che si ha quando le frequenze sono ripartite in maniera uniforme tra le due modalità estreme). La minima dispersione coincide con il caso di minima eterogeneità (o massima omogeneità), ovvero quando tutte le unità presentano la stessa modalità (caso 2).

Un indice di dispersione dovrà quindi assumere valore minimo nel caso della distribuzione caratterizzata da minima dispersione (caso 2) e valore massimo nel caso della distribuzione caratterizzata da massima dispersione (caso 4).

Nel caso di una variabile caratterizzata da dispersione minima, si è detto che tutte le unità presentano una stessa modalità del carattere. Indicando con $h$ tale modalità, si avrà:

| $X$ | $n$ | $f$ | $F$ | $RF$ |
|---|---|---|---|---|
| $x_{(1)}$ | 0 | 0 | 0 | 1 |
| $x_{(2)}$ | 0 | 0 | 0 | 1 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| $x_{(h-1)}$ | 0 | 0 | 0 | 1 |
| $x_{(h)}$ | $N$ | 1 | 1 | 1 |
| $x_{(h+1)}$ | 0 | 0 | 1 | 0 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| $x_{(k)}$ | 0 | 0 | 1 | 0 |
| | N | 1 | | |

In particolare, per le frequenze cumulate e retrocumulate si avrà, rispettivamente:

$$F_i = \begin{cases} 0 & \text{se } i < h \\ 1 & \text{se } i \geq h \end{cases} \qquad RF_i = \begin{cases} 1 & \text{se } i \leq h \\ 0 & \text{se } i > h \end{cases}$$

Definendo i complementi ad 1 delle frequenze cumulate e retrocumulate, si avrà ovviamente:

$$1 - F_i = \begin{cases} 1 & \text{se } i < h \\ 0 & \text{se } i \geq h \end{cases} \qquad 1 - RF_i = \begin{cases} 0 & \text{se } i \leq h \\ 1 & \text{se } i > h \end{cases}$$

Sfruttando le frequenze cumulate e retrocumulate (e i loro complementi all'unità) è possibile costruire il seguente indice di dispersione:

$$D = \sum_{i=1}^{k} \left[ F_i \left(1 - F_i\right) + RF_i \left(1 - RF_i\right) \right]$$

Essendo una somma di termini tutti non negativi, l'indice $D$ è non negativo. In particolare è immediato vedere come il valore minimo dell'indice (lo 0) è assunto nel caso di variabili caratterizzate da minima dispersione. La seguente tabella riporta i dati necessari per il calcolo dell'indice nel caso della distribuzione denotata in precedenza come caso 2:

| $X$ | $n$ | $f$ | $F$ | $1-F$ | $RF$ | $1-RF$ | $F\ (1-F)$ | $RF\ (1-RF)$ |
|---|---|---|---|---|---|---|---|---|
| $x_{(1)}$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| $x_{(2)}$ | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| $x_{(3)}$ | 100 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x_{(4)}$ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| $x_{(5)}$ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| | 100 | 1 | | | | | 0 | 0 |

Il valore massimo dell'indice, che si riscontra nel caso di distribuzione caratterizzata da dispersione massima, dipende dalla numerosità del collettivo di riferimento; in particolare si hanno i due seguenti casi (si rimanda all'appendice per la dimostrazione):

| *N* pari | *N* dispari |
|---|---|
| $\frac{k-1}{2}$ | $\frac{k-1}{2}\left(1-\frac{1}{N^2}\right)$ |

Il calcolo dell'indice D per il caso della distribuzione indicata come caso 4, ovvero il caso di massima dispersione, è riportata nella seguente tabella:

| *X* | *n* | *f* | *F* | *1-F* | *RF* | *1-RF* | *F (1-F)* | *RF (1-RF)* |
|---|---|---|---|---|---|---|---|---|
| $x_{(1)}$ | 50 | 0.5 | 0.5 | 0.5 | 1 | 0 | 0.25 | 0 |
| $x_{(2)}$ | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 |
| $x_{(3)}$ | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 |
| $x_{(4)}$ | 0 | 0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.25 | 0.25 |
| $x_{(5)}$ | 50 | 0.5 | 1 | 0 | 0.5 | 0.5 | 0 | 0.25 |
| | 100 | 1 | | | | | 1.00 | 1.00 |

Da cui si evince come l'indice *D* assume il valore 2, ovvero il suo valore massimo (essendo *N* pari e *k*= 5).

La seguente tabella riporta, infine, il calcolo dell'indice D per la distribuzione del caso 3, ovvero per la distribuzione caratterizzata da una dispersione intermedia:

| *X* | *n* | *f* | *F* | *1-F* | *RF* | *1-RF* | *F (1-F)* | *RF (1-RF)* |
|---|---|---|---|---|---|---|---|---|
| $x_{(1)}$ | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| $x_{(2)}$ | 50 | 0.5 | 0.5 | 0.5 | 0.5 | 1 | 0 | 0.25 |
| $x_{(3)}$ | 50 | 0.5 | 1 | 0 | 0.5 | 0.5 | 0.5 | 0.25 |
| $x_{(4)}$ | 0 | 0 | 1 | 0 | 0.5 | 0 | 1 | 0.25 |
| $x_{(5)}$ | 0 | 0 | 1 | 0 | 0.5 | 0 | 1 | 0.25 |
| | 100 | 1 | | | | | 0.25 | 0.25 |

In questo caso l'indice *D* vale 0.5, denotando una dispersione intermedia rispetto ai due casi di dispersione minima e dispersione massima.

E' possibile calcolare l'indice D senza far ricorso al calcolo delle frequenze retrocumulate, sfruttando la seguente equivalenza:

$$D = \sum_{i=1}^{k} \left[ F_i \left(1 - F_i\right) + RF_i \left(1 - RF_i\right) \right] = 2 \sum_{i=1}^{k-1} F_i \left(1 - F_i\right)$$

E' immediato vedere, negli esempi riportati in precedenza, come le due colonne $F_i \left(1 - F_i\right)$ e $RF_i \left(1 - RF_i\right)$ sono caratterizzate dalla stessa somma; per una dimostrazione formale dell'equivalenza delle due formule si rimanda all'appendice.

Nel caso in cui si voglia confrontare la dispersione tra due variabili caratterizzate da un numero diverso di modalità, è necessario costruire un indice normalizzato per eliminare la dipendenza del campo di variabilità dell'indice dal numero di modalità della variabile su cui si misura. E' possibile utilizzare la seguente formula:

$$I_{normalizzato} = \frac{I - \min(I)}{\max(I) - \min(I)}$$

dove con *I* si è indicato l'indice che si vuole normalizzare e di cui sono noti il valore minimo e massimo (rispettivamente indicati con min(I) e max(I)). L'indice ottenuto in seguito a tale trasformazione varia tra 0 ed 1, assumendo il valore 0 nel caso di minima dispersione e il valore 1 nel caso di massima dispersione. Per applicare tale trasformazione all'indice *D* si dovrebbe distinguere il caso di *N* pari ed *N* dispari, essendo diverso il valore massimo che l'indice può assumere nei due casi. Se *N* è sufficientemente grande, $\frac{1}{N^2}$ tende a 0: si può assumere come valore massimo di *D* sempre $\frac{k-1}{2}$ ed utilizzare tale valore nella formula di trasformazione sia nel caso di *N* pari che nel caso di *N* dispari.

## APPENDICE

**Dimostrazione dell'equivalenza delle due formule per il calcolo dell'indice di dispersione D:**

$$D = \sum_{i=1}^{k} \left[ F_i \left(1 - F_i\right) + RF_i \left(1 - RF_i\right) \right] = 2 \sum_{i=1}^{k-1} F_i \left(1 - F_i\right)$$

$$D = \sum_{i=1}^{k} \left[ F_i \left(1 - F_i\right) + RF_i \left(1 - RF_i\right) \right] =$$

$$= \sum_{i=1}^{k} F_i \left(1 - F_i\right) + \sum_{i=1}^{k} RF_i \left(1 - RF_i\right) = \quad \leftarrow \textbf{ distribuendo la sommatoria sui due addendi}$$

$$= \sum_{i=1}^{k-1} F_i \left(1 - F_i\right) + \sum_{i=2}^{k} RF_i \left(1 - RF_i\right) = \quad \leftarrow \textbf{ sfruttando il fatto che: } F_k = 1;\ \left(1 - F_k\right) = 0 \text{ e } RF_1 = 1;\ \left(1 - RF_1\right) = 0 \textbf{ si possono cambiare gli indici delle due sommatorie}$$

$$= \sum_{i=1}^{k-1} F_i\left(1-F_i\right)+\sum_{i=1}^{k-1} RF_{i+1}\left(1-RF_{i+1}\right)= \quad \leftarrow$$ **si può modificare l'indice della seconda sommatoria (e i pedici del suo argomento) in modo da uniformarne gli estremi a quelli della prima**

$$= \sum_{i=1}^{k-1} F_i\left(1-F_i\right)+\sum_{i=1}^{k-1} \left(1-F_i\right)\left(1-\left[1-F_i\right]\right)= \quad \leftarrow$$ **sfruttando la relazione che lega le frequenze retrocumulate alle frequenze cumulate:** $RF_i = 1-F_{i-1} \Rightarrow RF_{i+1} = 1-F_i$

$$= \sum_{i=1}^{k-1} F_i\left(1-F_i\right)+\sum_{i=1}^{k-1} \left(1-F_i\right)F_i =$$

$$= 2\sum_{i=1}^{k-1} F_i\left(1-F_i\right)$$

**Massimo dell'indice D**

Per trovare il valore massimo dell'indice D è necessario specificare separatamente il caso di *N* pari e di *N* dispari. Si può inoltre esprimere l'indice D usando le frequenze assolute in luogo delle relative come di seguito riportato:

$$D = 2\sum_{i=1}^{k-1} F_i\left(1-F_i\right) = 2\sum_{i=1}^{k-1} \frac{N_i}{N}\left(1-\frac{N_i}{N}\right) = 2\sum_{i=1}^{k-1} \frac{N_i}{N}\left(\frac{N-N_i}{N}\right) = \frac{2}{N^2}\sum_{i=1}^{k-1} N_i\left(N-N_i\right)$$

Per la dimostrazione si può far ricorso al seguente teorema:

FRA TUTTI I PRODOTTI DI DUE NUMERI INTERI AVENTI UNA STESSA SOMMA N, SE QUESTA È PARI, È MASSIMO QUEL PRODOTTO I CUI FATTORI SONO ENTRAMBI UGUALI AD $\frac{N}{2}$; SE N È DISPARI, È MASSIMO QUEL PRODOTTO I CUI TERMINI SONO $\frac{N-1}{2}$ ED $\frac{N+1}{2}$.

AD ESEMPIO:

| N pari → 10 = $n_1 + n_2$ | | | N dispari → 9 = $n_1 + n_2$ | | |
|---|---|---|---|---|---|
| **$n_1$** | **$n_2$** | **$n_1$ x $n_2$** | **$n_1$** | **$n_2$** | **$n_1$ x $n_2$** |
| 0 | 10 | 0 | 0 | 9 | 0 |
| 1 | 9 | 9 | 1 | 8 | 8 |
| 2 | 8 | 16 | 2 | 7 | 14 |
| 3 | 7 | 21 | 3 | 6 | 18 |
| 4 | 6 | 24 | 4 | 5 | 20 |
| 5 | 5 | 25 | | | |

**CASO DI N PARI**

Il massimo di D è assunto se, per ogni $i = 1, 2, ..., k-1$ si ha:

$$N_i = \frac{N}{2} \qquad N - N_i = \frac{N}{2}$$

Ciò si verifica soltanto se:

$$n_1 = \frac{N}{2}, \quad n_2 = 0, \quad n_3 = 0, \ldots, \; n_{k-1} = 0, \quad n_k = \frac{N}{2}$$

ossia se metà delle unità presenta la prima modalità e l'altra metà presenta l'ultima modalità. Tale caso coincide con il caso di massimo dispersione; l'indice *D*, in tale caso, vale:

$$D = \frac{2}{N^2}\sum_{i=1}^{k-1} N_i\left(N - N_i\right) = \frac{2}{N^2}\sum_{i=1}^{k-1}\frac{N}{2}\frac{N}{2} = \frac{2}{N^2}\left(k-1\right)\frac{N}{2}\frac{N}{2} = \frac{k-1}{2}$$

**CASO DI N DISPARI**

Il massimo è assunto se, per ogni $i = 1, 2, \ldots, k-1$, si ha:

$$N_i = n_1 + \ldots + n_i = \frac{N-1}{2} \qquad N - N_i = n_{i+1} + \ldots + n_k = \frac{N+1}{2}$$

ovvero se:

$$N_i = n_1 + \ldots + n_i = \frac{N+1}{2} \qquad N - N_i = n_{i+1} + \ldots + n_k = \frac{N-1}{2}$$

Queste due situazioni si verificano in uno dei seguenti tre casi:

1) $n_1 = \frac{N-1}{2}, \quad n_2 = 0, \quad n_3 = 0, \ldots, \; n_{k-1} = 0, \quad n_k = \frac{N+1}{2}$

2) $n_1 = \frac{N+1}{2}, \quad n_2 = 0, \quad n_3 = 0, \ldots, \; n_{k-1} = 0, \quad n_k = \frac{N-1}{2}$

3) $n_1 = \frac{N-1}{2}$ e $n_k = \frac{N-1}{2}$, una qualunque delle altre frequenze è uguale ad 1 e tutte le altre sono uguali a 0.

In questi casi D assume il valore massimo, avendosi:

$$\begin{aligned} D &= \frac{2}{N^2}\sum_{i=1}^{k-1} N_i\left(N - N_i\right) = \frac{2}{N^2}\sum_{i=1}^{k-1}\frac{N-1}{2}\frac{N+1}{2} = \\ &= \frac{2}{N^2}\left(k-1\right)\frac{N-1}{2}\frac{N+1}{2} = \frac{k-1}{2}\frac{\left(N-1\right)\left(N+1\right)}{N^2} \\ &= \frac{k-1}{2}\frac{\left(N-1\right)N + \left(N-1\right)}{N^2} = \frac{k-1}{2}\frac{N^2 - N + N - 1}{N^2} \\ &= \frac{k-1}{2}\frac{N^2-1}{N^2} = \frac{k-1}{2}\left(1 - \frac{1}{N^2}\right) \end{aligned}$$