UN TEST DI CONCORDANZA TRA PIÙ ESAMINATORI

Piero Quatto

1. INTRODUZIONE

La statistica "Kappa", proposta da Fleiss (1971), costituisce uno degli strumenti più utilizzati per saggiare l'accordo fra vari esaminatori, pur essendo caratterizzata da comportamenti paradossali (Quatto, 2003).

L'obiettivo del lavoro consiste nel proporre un test sulla casualità della concordanza tra più esaminatori basato su una statistica alternativa, che non risulta affetta dai paradossi della "Kappa" e presenta una distribuzione limite nota, non solo quando è grande il numero dei soggetti (come accade per la statistica "Kappa"), ma anche quando è grande il numero degli esaminatori.

Dopo aver delineato il problema della valutazione della concordanza fra esaminatori e le principali soluzioni presenti in letteratura, vengono studiate le distribuzioni asintotiche della statistica proposta per il test di concordanza e se ne illustrano gli aspetti salienti mediante un'applicazione a dati reali.

2. LA CONCORDANZA

Al fine di valutare l'accordo tra le classificazioni espresse da più esaminatori, si considerino N soggetti, ciascuno dei quali viene classificato mediante M categorie esaustive e mutuamente esclusive da un gruppo di n esaminatori, i cui membri non sono necessariamente gli stessi per ogni soggetto. Spesso gli esaminatori sono esperti di un certo settore (come ad esempio psicologi, medici, archeologi, critici d'arte, giudici sportivi, ecc.), ma possono anche essere consumatori o utenti chiamati a valutare la qualità di un insieme di prodotti o servizi tramite un questionario a risposta chiusa.

Indicato con x_{ij} il numero di esaminatori che hanno assegnato l'i-esimo soggetto (i=1,…,N) alla j-esima categoria (j=1,…,M), le assegnazioni effettuate possono rappresentarsi nella tabella seguente.

TABELLA 1

SOGGETTI	CATEGORIE					TOT.
	1	…	*j*	…	*M*	
1	x_{11}	…	x_{1j}	…	x_{1M}	$x_{1.}=n$
⋮	⋮		⋮		⋮	⋮
I	x_{i1}	…	x_{ij}	…	x_{iM}	$x_{i.}=n$
⋮	⋮		⋮		⋮	⋮
N	x_{N1}	…	x_{Nj}	…	x_{NM}	$x_{N.}=n$
TOT.	$x_{.1}$	…	$x_{.j}$	…	$x_{.M}$	*Nn*

Si noti che nella Tabella 1 ciascuna marginale

$$x_{i.} = \sum_{j=1}^{M} x_{ij} = n$$

è pari al numero di esaminatori per soggetto, mentre la generica marginale

$$x_{.j} = \sum_{i=1}^{N} x_{ij}$$

fornisce il numero totale di assegnazioni alla categoria *j*.

Definita la proporzione delle coppie di esaminatori che hanno assegnato il soggetto *i* alla categoria *j*

$$P_{ij} = \frac{\binom{x_{ij}}{2}}{\binom{n}{2}} = \frac{x_{ij}(x_{ij}-1)}{n(n-1)},$$

è possibile calcolare la proporzione delle coppie di assegnazioni concordanti relative al soggetto *i*

$$P_i = \sum_{j=1}^{M} P_{ij} = \frac{1}{n-1}\left(\frac{1}{n}\sum_{j=1}^{M} x_{ij}^2 - 1\right)$$

e misurare l'accordo osservato tramite la media

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{n-1}\left(\frac{1}{Nn}\sum_{i,j} x_{ij}^2 - 1\right) \qquad (1)$$

(Fleiss, 1971).

Se si ammette che la proporzione

$$p_j = \frac{x_{\cdot j}}{Nn} = \frac{1}{Nn}\sum_{i=1}^{N} x_{ij}$$

sia una stima della probabilità di assegnazione casuale alla categoria *j*, allora, seguendo Scott (1955) e Fleiss (1971), l'accordo atteso per effetto del caso è dato da

$$\overline{P}_e = \sum_{j=1}^{M} p_j^2 \ . \tag{2}$$

Sottraendo dall'accordo osservato (1) l'accordo atteso casuale (2) e normalizzando, si ottiene la statistica

$$K = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}, \tag{3}$$

proposta da Fleiss (1971) come generalizzazione dell'indice "Kappa" di Cohen (1960) (Fleiss-Levin-Paik; 2003, pp. 598-617). Al riguardo, è opportuno sottolineare che la (3) rappresenta l'estensione dell'indice π di Scott (1955) al caso in cui gli esaminatori siano più di due e costituisce uno degli strumenti più usati per valutare l'accordo fra *n* esaminatori, sebbene possa comportarsi in modo paradossale (Quatto, 2003).

D'altro canto, se si suppone che le assegnazioni di ciascun soggetto alle varie categorie siano indipendenti ed avvengano in modo casuale, allora, non essendoci alcuna ragione per ritenere che il caso possa privilegiare certe categorie rispetto ad altre, appare lecito assumere che tutte le categorie siano equiprobabili. In tal modo, l'accordo atteso per effetto del caso può esprimersi tramite la somma delle *M* probabilità di ottenere una coppia di assegnazioni casuali alla medesima categoria, data da

$$M(1/M)^2 = 1/M \ . \tag{4}$$

Infine, "depurando" l'accordo osservato (1) dall'accordo atteso casuale (4) e normalizzando, si perviene all'indice di concordanza effettiva

$$S = \frac{\overline{P} - 1/M}{1 - 1/M} = \frac{M\overline{P} - 1}{M - 1} \in \left[-\frac{1}{n-1}, 1 \right] \tag{5}$$

(Quatto, 2003), che generalizza quello proposto da Bennet, Alpert e Goldstein (1954) e ha la stessa struttura ed il medesimo campo di variazione della statistica (3), senza però essere affetto dai relativi paradossi, come sarà evidenziato nel paragrafo conclusivo.

3. IL MODELLO MULTINOMIALE

Se si assume che gli N gruppi di esaminatori costituiscano altrettanti campioni bernoulliani indipendenti, ciascuno di numerosità n (Fleiss, 1971), allora le N righe della Tabella 1 possono interpretarsi come determinazioni di altrettante v.c. Multinomiali indipendenti, caratterizzate dai parametri n e $\theta_{i1},\ldots,\theta_{iM}$ (i=1,…,N).

Nell'ambito di questo modello Multinomiale, l'ipotesi nulla secondo la quale le categorizzazioni avvengono in modo casuale, prescindendo dal soggetto, può formalizzarsi come

$$H_0 : \theta_{ij} = 1/M \quad (\forall i, j), \tag{6}$$

non essendoci motivi perché il caso privilegi alcune categorie rispetto alle altre. In tale contesto, il livello di concordanza effettiva tra gli esaminatori, espresso dalla quota dell'accordo osservato non riconducibile al caso, aumenta al crescere del divario tra i valori osservati e quelli previsti dall'ipotesi nulla.

Dal punto di vista della teoria asintotica, una conveniente misura della compatibilità dei dati (x_{ij}) con l'ipotesi H_0 è data dalla somma delle N v.c. indipendenti

$$X_i^2 = \sum_{j=1}^{M} \frac{(x_{ij} - n/M)^2}{n/M} = \frac{M}{n} \sum_j x_{ij}^2 - n ,$$

che produce la statistica

$$X^2 = \sum_{i=1}^{N} X_i^2 = \frac{M}{n} \sum_{i,j} x_{ij}^2 - Nn . \tag{7}$$

Difatti, supponendo valida l'ipotesi nulla, le v.c. X_i^2 hanno distribuzione limite per $n \to \infty$ fornita da

$$X_i^2 \xrightarrow{d} \chi^2_{M-1} ,$$

cosicché

$$X^2 \xrightarrow{d} \chi^2_{N(M-1)} \qquad (n \to \infty) . \tag{8}$$

D'altra parte, fisso restando il numero n degli esaminatori, se il numero N dei soggetti tende all'infinito è possibile applicare il Teorema centrale del limite di Lindeberg-Lévy alle v.c. X_i^2, che sotto H_0 sono i.i.d. ed hanno media e varianza date rispettivamente da

$$\mathrm{E}(X_i^2) = M - 1$$

e

$$\mathrm{Var}(X_i^2) = 2(n-1)(M-1)/n\,,$$

come si può vedere utilizzando i momenti della v.c. Multinomiale (Johnson-Kotz-Balakrishnan, 1997, pp. 31-41). Più precisamente, sotto l'ipotesi nulla si ottiene

$$\frac{X^2 - N(M-1)}{\sqrt{2(n-1)N(M-1)/n}} \xrightarrow{d} N(0,1) \qquad (N \to \infty)\,,$$

essendo

$$\mathrm{E}(X^2) = N(M-1)$$

e

$$\mathrm{Var}(X^2) = 2(n-1)N(M-1)/n\,.$$

4. UN TEST DI CONCORDANZA

Sulla base della statistica (7) è possibile costruire un test di significatività che rifiuta H_0 per valori elevati di X^2 ed ha livello di significatività osservato (*p-value*) dato da

$$\hat{\alpha} = P(X^2 \geq x^2 \mid H_0)\,, \tag{9}$$

dove x^2 è il valore osservato di X^2. Tale test risulta equivalente al test basato sulla statistica (5) (Quatto, 2003), poiché

$$S = \frac{M\overline{P}-1}{M-1} = \frac{1}{M-1}\left[\frac{M}{n-1}\left(\frac{1}{Nn}\sum_{i,j} x_{ij}^2 - 1\right) - 1\right] = \frac{1}{n-1}\left[\frac{X^2}{N(M-1)} - 1\right].$$

In particolare, il *p-value* (9) è approssimabile attraverso una distribuzione Normale o Chi-quadrato, a seconda che sia grande il numero dei soggetti o quello degli esaminatori. Il primo caso si verifica quando pochi esperti sono chiamati ad esaminare molti soggetti, mentre il secondo caso si realizza quando un paniere costituito da pochi prodotti o servizi è sottoposto alla valutazione di molti consumatori o utenti.

5. UN'APPLICAZIONE

Nell'ambito di un'indagine sulla valutazione della didattica relativa a tre insegnamenti universitari (N=3), è stato estratto un campione di numerosità n=30

dalla popolazione degli studenti di ciascuno dei tre corsi e ad ogni individuo selezionato è stato richiesto di valutare le lezioni alle quali ha assistito attraverso quattro modalità ordinate (M=4), ottenendo la seguente tabella di assegnazioni.

TABELLA 2

CORSI	MODALITÀ				TOT.
	1	2	3	4	
1	7	20	3	0	30
2	8	21	1	0	30
3	1	17	9	3	30
TOT.	16	58	13	3	90

Sulla base della Tabella 2 si possono calcolare le statistiche (1), (5) e (7)

$$\overline{P} = 0.48 \qquad \mathrm{S} = 0.31 \qquad \mathrm{X}^2 = 89.2 ,$$

nonché l'approssimazione del *p-value* (9)

$$P(\chi_9^2 \geq 89.2) = 0 ,$$

stante la (8). Emerge così un livello di concordanza effettiva significativo, da cui discende il rifiuto dell'ipotesi nulla (6) che attribuisce agli intervistati un comportamento aleatorio.

Infine, si osservi che, essendo N=3, non è possibile applicare l'approssimazione Normale e che la prevalenza della seconda modalità sulle altre produce una sovrastima dell'accordo atteso casuale dato dalla (2)

$$\overline{P}_e = 0.46$$

ed una conseguente sottostima dell'accordo effettivo misurato dalla (3)

$$K = 0.04 .$$

6. CONCLUSIONI

La statistica proposta ha il duplice vantaggio di offrire una distribuzione limite nota sia quando è grande il numero dei soggetti (come avviene per la "Kappa") sia quando è grande il numero degli esaminatori, senza incorrere nei paradossi tipici della "Kappa".

D'altro canto, tale statistica è costruita, analogamente alla "Kappa", sotto la condizione che il numero degli esaminatori sia lo stesso (n) per ogni soggetto (Fleiss, 1971; Shoukri, 2004, pp. 53-57) e questa assunzione potrebbe diventare oggetto di successivi approfondimenti, anche se non appare troppo restrittiva, essendo sempre possibile estrarre con reinserimento un campione di ampiezza pre-

fissata n da ciascuna delle N popolazioni di possibili esaminatori associate agli N soggetti.

Dipartimento di Statistica PIERO QUATTO
Università Milano-Bicocca

RIFERIMENTI BIBLIOGRAFICI

E.M. BENNET, R. ALPERT, A.C. GOLDSTEIN, (1954), *Communications through limited response questioning*, "Public Opinion Quarterly", 18, pp. 303-308.

J. COHEN, (1960), *A coefficient of agreement for nominal scale*, "Educational and Psychological Measurement", 20, pp. 37-46.

J.L. FLEISS, (1971), *Measuring nominal scale agreement among many raters*, "Psychological Bulletin", 76, pp. 378-382.

J.L. FLEISS, B. LEVIN, M.C. PAIK, (2003), *Statistical Methods for Rates and Proportions*, John Wiley & Sons, Hoboken.

N.L. JOHNSON, S. KOTZ, N. BALAKRISHNAN, (1997), *Discrete Multivariate Distributions*, John Wiley & Sons, New York.

P. QUATTO, (2003), *Un test sulla natura casuale dell'accordo tra più esaminatori*, Atti del 5° Congresso Nazionale della Società Italiana di Biometria, Marina di Massa, 10-12 settembre 2003, pp. 33-36.

W.A. SCOTT, (1955), *Reliability of content analysis: the case of nominal scale coding*, "Public Opinion Quarterly", 19, pp. 321-325.

M.M. SHOUKRI, (2004), *Measures of Interobserver Agreement*, Chapman & Hall, Boca Raton.

RIASSUNTO

Un test di concordanza tra più esaminatori

L'obiettivo del presente lavoro consiste nel proporre un test sulla casualità della concordanza tra più esaminatori basato su una statistica di tipo χ^2. Il principale vantaggio di questa statistica-test risiede nella sua distribuzione asintotica, che risulta nota non solo quando è grande il numero dei soggetti, ma anche quando è grande il numero degli esaminatori.

SUMMARY

Testing agreement among multiple raters

The aim of this paper is to propose a procedure for testing chance agreement among multiple raters which is based on a χ^2 statistic. The main advantage of using the χ^2 test statistic is that it has a well-known limit distribution when either the number of subjects or the number of raters is large.