UN TEST SULLA NATURA CASUALE DELL'ACCORDO TRA PIÙ ESAMINATORI

Piero Quatto

Dipartimento di Statistica, Università degli Studi di Milano – Bicocca

Riassunto

La statistica "Kappa" costituisce uno degli strumenti più utilizzati per saggiare l'accordo fra esaminatori, sebbene possa comportarsi in modo paradossale, assumendo lo stesso valore in corrispondenza di situazioni molto differenti.

L'obiettivo del presente lavoro consiste nel proporre un test sulla casualità della concordanza tra più esaminatori, basato su una statistica alternativa che non risulta affetta dai paradossi tipici della "Kappa".

1. INTRODUZIONE

La statistica "Kappa" è stata introdotta da Cohen (1960) per misurare il livello di concordanza tra due esaminatori, ma può assumere valori molto bassi anche in situazioni di forte accordo. Questi comportamenti paradossali sono stati oggetto di studi approfonditi (Feinstein - Cicchetti, 1990; Cicchetti - Feinstein, 1990; Shoukri, 2004), a differenza degli analoghi paradossi che caratterizzano la statistica proposta da Fleiss (1971) come generalizzazione della "Kappa" di Cohen al caso di n esaminatori (Fleiss - Levin - Paik, 2003).

Il presente lavoro intende evidenziare i paradossi della statistica di Fleiss e proporre una statistica alternativa che, senza incorrere in tali paradossi, permetta di valutare la concordanza tra più esaminatori e di verificare l'ipotesi nulla secondo la quale l'accordo osservato è riconducibile al caso.

2. L'ACCORDO OSSERVATO

Si considerino n esaminatori, ciascuno dei quali deve ripartire N soggetti in M categorie esaustive e mutuamente esclusive. Tipicamente gli esaminatori sono esperti di un certo settore (come per esempio psicologi, medici, archeologi, critici d'arte, ecc.), ma possono anche essere consumatori o utenti chiamati a valutare la

qualità di un insieme di prodotti o servizi tramite un questionario con domande 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), il complesso delle assegnazioni effettuate dagli n esaminatori trova una rappresentazione naturale nella Tab. 1.

Tab. 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 osservi che nella Tab. 1 ciascuna marginale

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

rappresenta il numero degli esaminatori, mentre la generica marginale

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

fornisce il numero totale di assegnazioni alla categoria j.

Se due o pi esaminatori sono concordi nell assegnare il soggetto i alla categoria j, allora tale situazione di accordo tra esaminatori si manifesta nella Tab.°1 attraverso la corrispondente frequenza interna $x_{ij}\geq 2$, che consente di determinare il numero delle coppie di esaminatori concordanti

$$\binom{x_{ij}}{2} = \frac{x_{ij}\left(x_{ij}-1\right)}{2}.$$

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

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

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

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

(Fleiss, 1971).

3. LA STATISTICA "KAPPA"

Si dice che un esaminatore assegna in modo deterministico un soggetto ad una categoria, se l'assegnazione non varia quando il soggetto viene ripetutamente sottoposto alla valutazione dell'esaminatore. In caso contrario, la categorizzazione del soggetto viene detta casuale perché assume l'attitudine a variare tipica dei fenomeni aleatori.

Ne discende che, in generale, la concordanza di due o più esaminatori nel classificare uno stesso soggetto può essere interpretata come l'effetto osservabile della combinazione di due fattori non osservabili, uno deterministico e l'altro casuale. Emerge così la necessità di scorporare dall'accordo osservato la quota di origine casuale, al fine di isolare la componente di natura deterministica, che rappresenta l'oggetto di studio.

Se si ammette che la proporzione

$$p_j = \frac{x_{\cdot j}}{Nn} = \frac{1}{Nn}\sum_{i=1}^{N} x_{ij} \tag{2}$$

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

$$\bar{P}_e = \sum_{j=1}^{M} p_j^2. \tag{3}$$

Infine, "depurando" l'accordo osservato (1) dall'accordo atteso casuale (3) e normalizzando, si perviene alla statistica

$$K = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \in \left[-\frac{1}{n-1}, 1\right], \tag{4}$$

proposta da Fleiss (1971) come generalizzazione dell'indice "Kappa" di Cohen (1960). Al riguardo, è opportuno sottolineare che la (4) rappresenta l'estensione dell'indice π di Scott (1955) al caso di n esaminatori.

Si noti che l'accordo atteso casuale può superare l'accordo osservato e in tal caso si ottiene $K<0$; altrimenti, si ha $K\geq 0$ e in particolare $K=1$ se c'è pieno accordo fra gli n esaminatori, ovvero quando l'accordo osservato raggiunge il valo e massimo.

4. I PARADOSSI

Si consideri la Tab. 2 nella quale tutti i soggetti sono ripartiti nelle prime due categorie secondo la medesima proporzione.

Tab. 2

Soggetti	Categorie					Tot.
	1	2	3	...	M	
1	m	$n-m$	0	...	0	n
⋮	⋮	⋮	⋮		⋮	⋮
i	m	$n-m$	0	...	0	n
⋮	⋮	⋮	⋮		⋮	⋮
N	m	$n-m$	0	...	0	n
Tot.	Nm	$N(n-m)$	0	...	0	Nn

Esaminando la Tab. 2 si può rilevare come, al variare di m da 0 a n, si passi da situazioni di completo accordo fra gli esaminatori, prodotte dai valori estremi $m=0$ e $m=n$, a casi di minore accordo, in corrispondenza dei valori intermedi di$m$. In particolare, le formule (1), (3) e (4) applicate alla Tab. 2 danno, rispettivamente,

$$\bar{P} = 1 - 2\frac{m(n-m)}{n(n-1)},$$

$$\bar{P}_e = 1 - 2\frac{m(n-m)}{n^2}$$

e

$$K = \frac{-2\frac{m(n-m)}{n^2(n-1)}}{2\frac{m(n-m)}{n^2}} = -\frac{1}{n-1} \qquad (5)$$

per $0<m<n$. D'altro canto, se $m \to 0$ o $m \to n$

$$K \to -\frac{1}{n-1}.$$

Da questi risultati si evince un'eccessiva prossimità fra l'accordo osservato $\bar{P}$ e l'accordo atteso $\bar{P}_e$, che determina l'invarianza della statistica K rispetto a m (ed a M) e la conseguente inadeguatezza nella valutazione del livello di concordanza manifestato dagli n esaminatori. Più precisamente, il valore costante e negativo assunto dalla "Kappa" non solo impedisce di riconoscere i diversi gradi di concordanza che si realizzano al variare di m, ma non permette neppure di discriminare le situazioni di accordo perfetto dalle altre. Ad esempio, con $m=5$ e $n=6$ si ha $k=-0.2$, pur essendoci 5 esaminatori su 6 d'accordo tra loro.

5. UN TEST SULLA CASUALITÀ DELL'ACCORDO

Se si suppone che gli N soggetti formino un campione bernoulliano e che valga l'ipotesi nulla H_0, secondo la quale le categorizzazioni avvengono in modo casuale prescindendo dal soggetto, allora le N righe della Tab. 1 possono interpretarsi come determinazioni di altrettante v.c. Multinomiali indipendenti e identicamente distribuite, caratterizzate dai parametri

$$n;\theta_1,\theta_2,\ldots,\theta_M.$$

Inoltre, non essendoci ragioni per ritenere che il caso possa privilegiare alcune categorie rispetto alle altre, appare lecito assumere

$$\theta_1 = \theta_2 = \ldots = \theta_M = \frac{1}{M}.$$

In tal modo, si ha che sotto H_0

$$E\left(\bar{P}\right) = \frac{1}{N}\sum_{i=1}^{N} E\left(P_i\right) = \frac{1}{M} \tag{6}$$

e

$$Var\left(\bar{P}\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var\left(P_i\right) = \frac{2\left(M-1\right)}{Nn\left(n-1\right)M^2},$$

in quanto

$$E\left(P_i\right) = \frac{1}{M}$$

e

$$Var\left(P_i\right) = \frac{2\left(M-1\right)}{n\left(n-1\right)M^2},$$

come si può vedere utilizzando i momenti della v.c. Multinomiale (Johnson - Kotz - Balakrishnan, 1997).

Poiché sotto l'ipotesi nulla le v.c. P_i sono indipendenti e identicamente distribuite, è possibile applicare il Teorema centrale del limite di Lindeberg - Lévy ed ottenere la distribuzione limite (per $N \to \infty$) di $\bar{P}$:

$$\left(M\bar{P}-1\right)\sqrt{\frac{Nn\left(n-1\right)}{2\left(M-1\right)}} = \frac{\bar{P}-\frac{1}{M}}{\sqrt{\frac{2\left(M-1\right)}{Nn\left(n-1\right)M^2}}} \xrightarrow{d} N\left(0,1\right). \tag{7}$$

In analogia con la (4) si può definire la statistica

$$S = \frac{\bar{P}-\frac{1}{M}}{1-\frac{1}{M}} = \frac{M\bar{P}-1}{M-1} \in \left[-\frac{1}{n-1},1\right], \tag{8}$$

che, grazie alla (7), risulta caratterizzata dalla distribuzione asintotica sotto H_0

$$S\sqrt{\frac{Nn\left(n-1\right)\left(M-1\right)}{2}} \xrightarrow{d} N\left(0,1\right). \tag{9}$$

Tale statistica costituisce una generalizzazione dell indice di Bennett - Alpert - Goldstein (1954) e pu essere impiegata per valutare la concordanza fra pi esaminatori e per verificare l ipotesi nulla che attribuisce l accordo osservato all azione del caso.

Più precisamente, il test di significatività basato sul divario tra l'accordo osservato (1) e l'accordo atteso casuale (6) rifiuta H_0 per valori elevati della (8). In particolare, grazie alla (9), il livello di significatività osservato $P\left(S \geq s \mid H_0\right)$ è approssimabile mediante

$$\tilde{\alpha} = 1-\Phi\left(s\sqrt{\frac{Nn\left(n-1\right)\left(M-1\right)}{2}}\right), \tag{10}$$

essendo s la determinazione di S e Φ la funzione di ripartizione della Normale standard.

6. L'ELIMINAZIONE DEI PARADOSSI

La statistica (8) ed il relativo test non producono i paradossi tipici della "Kappa", poiché questi sono riconducibili alla sovrastima dell'accordo atteso basata sulle marginali (2).

Infatti, il valore $1/M$ rappresenta il minimo di $\bar{P}_e$, come si può vedere osservando che il punto di minimo della funzione (3) sotto i vincoli

$$\sum_{j=1}^{M} p_j = 1, \; 0 \le p_j \le 1$$

corrisponde al vettore $p = (p_1,\ldots,p_M)$ con lunghezza

$$|p| = \sqrt{\sum_{j=1}^{M} p_j^2}$$

minima tra i vettori M-dimensionali con proiezione ortogonale su $u = (1,\ldots,1)$ data da

$$p \cdot \frac{u}{|u|} = \frac{1}{\sqrt{M}}. \tag{11}$$

Si tratta, dunque, del vettore parallelo a u

$$p = \lambda u \tag{12}$$

con il coefficiente $\lambda = 1/M$ ricavato sostituendo la (12) nella (11). Ne discende che le coordinate $p_j = 1/M$ individuano il punto in cui la funzione (3) assume il valore minimo $|p|^2 = 1/M$.

In conclusione, la statistica S permette sia la valutazione del livello di concordanza tra pi esaminatori che la verifica dell ipotesi di casualit dell accordo osservato, impiegando la medesima logica della Kappa senza per incorrere nei relativi paradossi. Per esempio, con i dati della Tab. 2 si ha

$$S = 1 - \frac{2m(n-m)M}{n(n-1)(M-1)},$$

che, a differenza della (5), dipende in modo coerente da m (e da M).

7. APPLICAZIONI

Il test di significatività proposto rifiuta l'ipotesi H_0 sulla natura casuale dell'accordo osservato se

$$s\sqrt{Nn(n-1)(M-1)2} \ge z_{1-\alpha},$$

dove s è la determinazione della statistica (8) e $z_{1-\alpha}$ è il quantile di ordine $1-\alpha$ della Normale standard.

Tale ipotesi prevede che gli esaminatori assegnino le categorie in modo casuale, prescindendo dalle caratteristiche specifiche dei singoli soggetti. Nel caso in cui gli esaminatori siano consumatori o utenti che rispondono a domande sulla qualità di prodotti o servizi, l'ipotesi nulla attribuisce agli intervistati un comportamento aleatorio. Se invece gli esaminatori sono degli esperti, allora l'ipotesi H_0 corrisponde a un'alta percentuale di errori di valutazione, ovvero a una forte presenza di soggetti difficili da classificare.

Ad esempio, rivolgendo l'attenzione ai dati della Tab. 3 (Fleiss, 1971), concernenti 30 pazienti classificati da 6 psichiatri in 5 categorie diagnostiche, si possono calcolare i valori

$$\overline{P} = 0.556$$
$$\overline{P}_e = 0.220$$
$$K = 0.431$$
$$S = 0.444,$$

che mostrano una lieve differenza tra la statistica "Kappa" e l'alternativa S.

Inoltre, il p-value fornito dalla (10) risulta così prossimo a zero da consentire il rifiuto di H_0.

D'altra parte, se si collassano le ultime tre categorie della Tab. 3, allora i nuovi valori

$$\overline{P} = 0.640$$
$$\overline{P}_e = 0.548$$
$$K = 0.204$$
$$S = 0.460$$

evidenziano l'incapacità della "Kappa" di cogliere l'incremento del livello di accordo fra gli esaminatori, messo in rilievo dalla statistica S.

Tab. 3

Soggetti	Categorie diagnostiche				
	Depressione	**Disturbi della personalità**	**Schizofrenia**	**Nevrosi**	**Altro**
1	0	0	0	6	0
2	0	0	3	0	3
3	0	1	4	0	1
4	0	0	0	0	6
5	0	3	0	3	0
6	2	0	4	0	0
7	0	0	4	0	2
8	2	0	3	1	0
9	2	0	0	4	0
10	0	0	0	0	6
11	1	0	0	5	0
12	1	1	0	4	0
13	0	3	3	0	0
14	1	0	0	5	0
15	0	2	0	3	1
16	0	0	5	0	1
17	3	0	0	1	2
18	5	1	0	0	0
19	0	2	0	4	0
20	1	0	2	0	3
21	0	0	0	0	6
22	0	1	0	5	0
23	0	2	0	1	3
24	2	0	0	4	0
25	1	0	0	4	1
26	0	5	0	1	0
27	4	0	0	0	2
28	0	2	0	4	0
29	1	0	5	0	0
30	0	0	0	0	6
Tot.	26	26	30	55	43

RIFERIMENTI BIBLIOGRAFICI

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

CICCHETTI D.V. - FEINSTEIN A.R. (1990), High agreement but low kappa: II. Resolving the paradoxes, *Journal of Clinical Epidemiology,* 43, 551-558.

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

FEINSTEIN A.R. - CICCHETTI D.V. (1990), High agreement but low kappa: I. The problems of two paradoxes, *Journal of Clinical Epidemiology,* 43, 543-549.

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

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

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

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

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

TESTING CHANCE AGREEMENT AMONG MANY RATERS

Summary

"Kappa" statistic has become a very popular tool for assessing agreement between raters, in spite of its paradoxical behaviour which happens when different situations give rise to the same "Kappa" estimate.

The aim of this paper is to propose a procedure for testing chance agreement among multiple raters that is based on a test statistic not affected by "Kappa" paradoxes.