***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 109-116

Indici e test di concordanza con un'applicazione alla valutazione della didattica universitaria

Piero Quatto[§]

Summary: *In spite of its paradoxical behaviour, "Kappa" statistic has become a popular tool for measuring interobserver agreement.*
The aim of this paper is twofold: firstly to point out the inadequacy of the "Kappa" statistic in the context of University Student Satisfaction; secondly to propose a procedure for assessing and testing agreement among multiple raters which is based on a statistic not affected by "Kappa" paradoxes.
Another advantage of the proposed statistic is that it has a well-known limit distribution when either the number of subjects or the number of raters is large.

Keywords: *Measures of Agreement, Chance Agreement Test, "Kappa" Statistics.*

1. Introduzione

La statistica "Kappa" è stata introdotta da Cohen (1960) per misurare il livello di concordanza fra due esaminatori chiamati a classificare N oggetti secondo M categorie e, a tal fine, costituisce uno degli strumenti più diffusi, sebbene possa assumere valori molto bassi anche in situazioni di forte accordo.

Questo comportamento paradossale è stato oggetto di molti studi (Feinstein - Cicchetti, 1990; Cicchetti - Feinstein, 1990; Lantz - Nebenzahl, 1996; Shoukri, 2004), a differenza dell'analogo problema che riguarda la statistica proposta da Fleiss (1971) come generalizzazione della "Kappa" di Cohen.

Nel presente lavoro si intende innanzitutto rimarcare che la statistica di Fleiss non costituisce l'estensione al caso di più esaminatori della "Kappa" di Cohen, ma bensì dell'indice π di Scott (1955).

Inoltre, con riferimento ai dati relativi ad un'indagine sulla valutazione della didattica universitaria, si vuole mettere in evidenza il comportamento

[§] Dipartimento di Statistica, Università degli Studi di Milano - Bicocca
via Bicocca degli Arcimboldi, 8 – 20126 Milano (e-mail: piero.quatto@unimib.it).

paradossale della statistica di Fleiss e proporre l'applicazione di una statistica alternativa (Quatto, 2004), che non risulta affetta da tale problema.

In particolare, tale statistica consente sia la misura del livello di concordanza tra più esaminatori, sia la verifica dell'ipotesi nulla di casualità delle classificazioni.

2. Indici di concordanza

Si considerino N oggetti (che possono essere, ad esempio, prodotti o servizi), 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 oggetto.

Indicato con x_{ij} il numero di esaminatori che hanno assegnato l'i-esimo oggetto ($i = 1,\ldots,N$) alla j-esima categoria ($j = 1,\ldots,M$), le assegnazioni effettuate possono rappresentarsi come nella Tabella 1.

Tabella 1. *Frequenze di assegnazione degli oggetti alle categorie*

Oggetti	**Categorie**					**Tot.**
	1	...	j	...	M	
1	x_{11}	⋯	x_{1j}	⋯	x_{1M}	$x_{1\cdot} = n$
⋮	⋮		⋮		⋮	⋮
i	x_{i1}	⋯	x_{ij}	⋯	x_{iM}	$x_{i\cdot} = n$
⋮	⋮		⋮		⋮	⋮
N	x_{N1}	⋯	x_{Nj}	⋯	x_{NM}	$x_{N\cdot} = n$
Tot.	$x_{\cdot 1}$	⋯	$x_{\cdot j}$	⋯	$x_{\cdot M}$	Nn

Si osservi che nella Tabella 1 la marginale

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

fornisce il numero degli esaminatori, mentre la marginale

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

rappresenta il numero totale di assegnazioni alla categoria j.

Definita la proporzione delle coppie di esaminatori che hanno assegnato l'oggetto i alla categoria j

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

è possibile calcolare la proporzione delle coppie di assegnazioni concordanti relative all'oggetto *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 (Fleiss, 1971)

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

2.1 *La statistica "Kappa" di Fleiss*

Se si ammette che la proporzione

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

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

$$\overline{P}_e = \sum_{j=1}^{M} p_j^2 \,. \qquad (2)$$

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

$$K_{\text{Fleiss}} = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \in \left[-\frac{1}{n-1}, 1\right], \qquad (3)$$

proposta da Fleiss (1971) come generalizzazione dell'indice "Kappa" di Cohen (1960).

2.2 *La statistica "Kappa" di Cohen*

Per introdurre l'indice di Cohen (1960), si consideri il caso particolare in cui gli *N* oggetti sono classificati secondo le *M* categorie da $n = 2$ esaminatori, che sono gli stessi per ogni oggetto.

Denotando con n_{jk} il numero di oggetti assegnati dal primo esaminatore alla categoria j (j = 1,…,M) e dal secondo esaminatore alla categoria k (k = 1,…,M), le classificazioni effettuate trovano una rappresentazione naturale nella Tabella 2.

Tabella 2. *Tabella di contingenza nel caso di due esaminatori*

Esaminatore 2

Esaminatore 1	**Categorie**	1	…	k	…	M	**Tot.**
	1	n_{11}	…	n_{1k}	…	n_{1M}	$n_{1.}$
	⋮	⋮		⋮		⋮	⋮
	j	n_{j1}	…	n_{jk}	…	n_{jM}	$n_{j.}$
	⋮	⋮		⋮		⋮	⋮
	M	n_{M1}	…	n_{Mk}	…	n_{MM}	$n_{M.}$
	Tot.	$n_{.1}$	…	$n_{.k}$	…	$n_{.M}$	N

Definite le frequenze relative

$$\pi_{jk} = \frac{n_{jk}}{N},$$

è possibile sottrarre dalla proporzione di oggetti assegnati da entrambi gli esaminatori alla medesima categoria,

$$\pi_o = \sum_{j=1}^{M} \pi_{jj} = \frac{1}{N} \sum_{j=1}^{M} n_{jj},$$

l'accordo atteso sotto l'ipotesi di indipendenza delle assegnazioni,

$$\pi_e = \sum_{j=1}^{M} \pi_{j.} \pi_{.j} = \frac{1}{N^2} \sum_{j=1}^{M} n_{j.} n_{.j},$$

e normalizzare, pervenendo alla statistica di Cohen (1960):

$$K_{\text{Cohen}} = \frac{\pi_o - \pi_e}{1 - \pi_e} \in [-1,1].$$

Tale indice può assumere valori prossimi all'estremo inferiore del campo di variazione anche in presenza di un accentuato accordo tra i due esaminatori, e questo comportamento paradossale è stato oggetto di molte ed approfondite analisi (Feinstein - Cicchetti, 1990; Cicchetti - Feinstein, 1990; Lantz - Nebenzahl, 1996; Shoukri, 2004).

Se, però, si particolarizza l'indice di Fleiss (3) al caso n = 2, grazie alle relazioni

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N} P_{ij} = \frac{1}{N}\sum_{j=1}^{M} n_{jj} = \pi_o$$

e

$$\overline{P}_e = \sum_{j=1}^{M} p_j^2 = \frac{\pi_e}{2} + \frac{1}{4N^2}\sum_{j=1}^{M}\left(n_{j.}^2 + n_{.j}^2\right),$$

si ottiene l'indice π di Scott (1955)

$$K_{\text{Fleiss}} = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} = \frac{\pi_o - \overline{P}_e}{1 - \overline{P}_e} = \pi_{\text{Scott}},$$

che coincide con quello di Cohen solo quando le marginali della Tabella 2 sono identiche:

$$K_{\text{Fleiss}} = K_{\text{Cohen}} \Leftrightarrow \overline{P}_e = \pi_e \Leftrightarrow \forall j \;\; n_{j.} = n_{.j}.$$

Dunque, in generale, la "Kappa" di Fleiss non può ritenersi l'estensione della "Kappa" di Cohen al caso di $n > 2$ esaminatori.

2.3 *Una statistica alternativa*

Se si suppone che le assegnazioni di ciascun oggetto alle varie categorie siano indipendenti ed avvengano in modo casuale, allora, non essendoci motivi perché il caso privilegi alcune categorie rispetto alle 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

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

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, 2004), che generalizza quello proposto da Bennet, Alpert e Goldstein (1954) e ha la stessa struttura ed il medesimo campo di variazione della statistica "Kappa", senza però essere affetto dai relativi paradossi (come sarà illustrato nel paragrafo 4).

3. Test di concordanza

Se si assume che gli N gruppi di esaminatori costituiscano altrettanti campioni bernoulliani indipendenti, ciascuno di numerosità n (Fleiss, 1971), allora, sotto l'ipotesi nulla H_0 secondo la quale le categorizzazioni avvengono in modo casuale, si ottengono le seguenti distribuzioni asintotiche della statistica S (Quatto, 2004):

$$S\sqrt{Nn(n-1)(M-1)/2} \xrightarrow{d} N(0,1) \qquad (N \to \infty); \tag{6}$$

$$N(M-1)[(n-1)S+1] \xrightarrow{d} \chi^2_{N(M-1)} \qquad (n \to \infty). \tag{7}$$

Sulla base della statistica (5) è possibile costruire un test di significatività che rifiuta H_0 per valori elevati di S ed ha livello di significatività osservato (*p-value*) dato da

$$\hat{\alpha} = P(S \geq s \mid H_0), \tag{8}$$

dove s è il valore osservato di S.

Questo *p-value* è approssimabile attraverso una distribuzione Normale o Chi-quadrato, a seconda che sia grande il numero degli oggetti o quello degli esaminatori.

D'altro canto, la statistica (5) è costruita, come la "Kappa", sotto la condizione che il numero degli esaminatori sia lo stesso per ogni oggetto (Fleiss, 1971; Shoukri, 2004), dato che questa assunzione non appare troppo restrittiva, essendo sempre possibile estrarre con reinserimento un campione di ampiezza prefissata n da ciascuna delle N popolazioni di possibili esaminatori associate agli N oggetti.

4. Un'applicazione alla valutazione della didattica universitaria

Nell'ambito di un'indagine sulla valutazione della didattica relativa a 16 insegnamenti universitari (N = 16), condotta nell'A.A. 2003/2004 presso la Facoltà di Scienze Statistiche dell'Università degli Studi di Milano – Bicocca, è stato estratto un campione di numerosità n = 30 dalla popolazione degli studenti di ciascuno dei 16 corsi. Ad ogni individuo selezionato è stato richiesto di esprimere la propria soddisfazione nei confronti delle lezioni alle quali ha assistito attraverso quattro modalità (M = 4) ordinate (in ordine crescente di soddisfazione), ottenendo le assegnazioni rappresentate nella Tabella 3.

Sulla base di questa si possono calcolare le statistiche (1) e (5)

$$\overline{P} = 0.5125 \qquad S = 0.35,$$

nonché le approssimazioni del *p-value* (8) determinate mediante i limiti (6) e (7), entrambe pari a 0 (sebbene la prima non garantisca una buona approssimazione con $N = 16$). Ne discende il rifiuto dell'ipotesi nulla che attribuisce agli intervistati un comportamento aleatorio.

D'altra parte, si osservi che la prevalenza delle due modalità intermedie sulle altre produce una sovrastima dell'accordo atteso casuale dato dalla (2)

$$\overline{P}_e = 0.5086$$

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

$$K_{\text{Fleiss}} = 0.0079 .$$

In particolare, il test relativo alla statistica "Kappa" (Fleiss - Levin - Paik, 2003) non conduce al rifiuto di H_0, essendo il corrispondente *p-value* pari a 0.1416. Emerge così il comportamento paradossale della "Kappa", che può assumere un valore estremamente basso anche in una situazione di marcato accordo.

Tabella 3. *Frequenze di valutazione degli insegnamenti*

Insegnamenti	Modalità				Tot.
	1	2	3	4	
1	1	8	20	1	30
2	1	12	16	1	30
3	0	7	21	2	30
4	0	8	20	2	30
5	0	12	17	1	30
6	0	8	19	3	30
7	3	10	16	1	30
8	1	9	19	1	30
9	1	4	22	3	30
10	0	4	24	2	30
11	0	2	26	2	30
12	0	4	23	3	30
13	2	10	18	0	30
14	2	10	17	1	30
15	0	7	20	3	30
16	0	5	21	4	30
Tot.	11	120	319	30	480

Inoltre, unendo le due modalità intermedie (2 e 3) della Tabella 3, l'indice "Kappa" diminuisce, mentre il valore della statistica alternativa cresce, coerentemente con l'aumentato livello di concordanza:

$$K_{\text{Fleiss}} = -0.015 \qquad (p\text{-}value = 0.918),$$

$$S = 0.7578 \qquad (p\text{-}value = 0).$$

5. Conclusioni

Da quanto esposto nei paragrafi precedenti si può concludere che la statistica alternativa (5) non solo consente di valutare il livello di concordanza tra più esaminatori senza incorrere nei paradossi della "Kappa", ma permette anche di verificare l'ipotesi nulla di casualità delle assegnazioni, sia quando è grande il numero degli oggetti (come accade per la "Kappa"), sia quando è grande il numero degli esaminatori.

Riferimenti Bibliografici

Bennet 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 scale. *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.C. (2003). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Hoboken.

Lantz C.A., Nebenzahl E. (1996). Behavior and interpretation of the k statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, **49**, 431-434.

Quatto P. (2004). Un test di concordanza tra più esaminatori. *Statistica*, **LXIV**, 1, 145-151.

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.