GTND
ISSN 0394-9362

Giornale di Tecniche Nefrologiche e Dialitiche 2016; 28 (4): 289-292
DOI: 10.5301/GTND.2016.16402

EPIDEMIOLOGIA E STATISTICA

Valutazione della riproducibilità di una misura: la statistica Kappa

Francesco Franco[1], **Anteo Di Napoli**[2]

[1] Regione Lazio, Roma
[2] Comitato Tecnico-Scientifico RIDT, Roma

Reliability assessment of a measure: the kappa statistic

The Kappa coefficient is a measure of inter-rater agreement. This statistic measures observed agreement greater than chance and can range from -1 to 1. A value of zero indicates statistical independence and a value of 1 indicates a perfect agreement between observers. The value of Kappa is influenced by the prevalence of the evaluated condition: two observers can have a high observed agreement but low Kappa if the prevalence is very high or very low (paradox of the Kappa statistic).

Keywords: Cohen's Kappa, Inter-rater agreement, Intra-rater agreement, Reliability, Weighted Kappa

Francesco Franco

Introduzione

Dopo aver trattato, nei precedenti articoli, il tema della validità di una misura in tutte le sue dimensioni (1-3), nel presente lavoro ci occuperemo della riproducibilità di una misura soffermandoci sul caso di una misura espressa su scala nominale od ordinale (variabile qualitativa). La riproducibilità riguardante il caso di variabili continue e il metodo di Bland-Altman, che valuta la variabilità fra misure in termini di concordanza (4), potranno essere oggetto di un successivo numero di questa rubrica.

La riproducibilità di una misura rappresenta la probabilità di ottenere sempre lo stesso risultato se ripetuta più volte nelle stesse condizioni; più simili sono i valori ripetuti, più è alta la riproducibilità della misura. Riproducibilità e variabilità sono due concetti complementari, nel senso che, se una misura è molto variabile, è poco riproducibile e, all'opposto, se è poco variabile è altamente riproducibile. Una misura poco riproducibile è poco precisa e, quindi, non affidabile.

Quando si valuta la riproducibilità di una misura, si assume che tutti i *test* o gli osservatori siano tra loro equivalenti e indipendenti. La riproducibilità di un *test* o di un osservatore implica l'abilità degli stessi di ottenere valori di misure ripetute che differiscono poco tra loro a causa esclusivamente di un piccolo errore casuale.

Quando si considera la riproducibilità di un *test*, si distingue tra quella intra-osservatore (*intra-rater agreement*) e quella inter-osservatore (*inter-rater agreement*).

Si parla di riproducibilità intra-osservatore quando lo stesso *test*, ripetuto più volte sullo stesso paziente dal medesimo operatore, fornisce lo stesso risultato; tale riproducibilità è, generalmente, elevata e migliora con l'esperienza dell'operatore.

Si parla di riproducibilità inter-osservatore quando lo stesso *test*, ripetuto più volte sullo stesso paziente da operatori diversi, fornisce lo stesso risultato; tale riproducibilità varia molto a seconda della complessità del *test* e migliora con l'esperienza comune e l'uniformità nell'interpretazione degli *standard* di riferimento. Negli studi clinici che utilizzano valutazioni qualitative di più operatori, ci si attende che i giudizi espressi dagli stessi siano coerenti tra loro e costanti durante tutto il periodo dello studio. Quindi, una revisione periodica dell'accordo *inter-rater* e *intra-rater* dovrebbe far parte del controllo di qualità degli studi clinici. Se gli osservatori risultassero essere in disaccordo l'uno con l'altro, sarebbe necessario ripetere la formazione dei valutatori (5).

Anche in presenza di due o più osservatori tra loro indipendenti, una quota parte dell'accordo nei giudizi espressi potrebbe essere dovuta solo al caso (non reale). Per superare questo inconveniente, Cohen ha proposto l'uso della statistica Kappa, che opera una correzione della misura dell'accordo (misura del grado di accordo oltre il caso).

Di seguito, vengono descritte le formule per il calcolo della statistica Kappa (Tabb. I e II) e fornito un esempio pratico di calcolo del Kappa (RIQUADRO 1).

Accepted: October 27, 2016
Published online: November 24, 2016

Indirizzo per la corrispondenza:
Dr. Anteo Di Napoli
anteo.dinapoli@libero.it
Dr. Francesco Franco
franco_francesco@fastwebnet.it

© 2016 Wichtig Publishing

TABELLA I - Caso di una variabile qualitativa con 2 modalità (SÌ/No), con 2 valutatori e N soggetti

		Osservatore 2		Totale
		SÌ	**NO**	
Osservatore 1	SÌ	*a*	*b*	**a + b**
	NO	*c*	*d*	**c + d**
Totale		**a + c**	**b + d**	**N**

$$\textit{Accordo osservato} \quad A_o = \frac{a+d}{N}$$

TABELLA II - Frequenze attese per effetto del caso in assenza di una reale concordanza tra osservatori

		Osservatore 2		Totale
		SÌ	**NO**	
Osservatore 1	SÌ	$\frac{(a+c)*(a+b)}{N}$	$\frac{(b+d)*(a+b)}{N}$	**a + b**
	NO	$\frac{(a+c)*(c+d)}{N}$	$\frac{(b+d)*(c+d)}{N}$	**c + d**
Totale		**a + c**	**b + d**	**N**

$$\textit{Accordo atteso} \quad A_a = \frac{[(a+c)*(a+b)]+[(b+d)*(c+d)]}{N^2}$$

$$\textit{Kappa di Cohen} = \frac{A_o - A_a}{1 - A_a}$$

dove: **1** - A_a rappresenta il valore massimo dell'accordo osservato al netto dell'accordo dovuto al caso.

Tuttavia, la statistica Kappa è fortemente influenzata dalla prevalenza della condizione oggetto della valutazione. Pertanto, due osservatori con un alto accordo osservato possono determinare, in realtà, un basso valore di Kappa. Questa forte dipendenza del Kappa dalla prevalenza della caratteristica esaminata, con il conseguente sbilanciamento dei totali marginali, ne complica l'interpretazione come indice di qualità di una misura, rendendo difficile confrontare due o più valori del Kappa quando le prevalenze delle condizioni comparate sono diverse.

Il paradosso di un Kappa relativamente basso, nonostante un risultato relativamente elevato nell'accordo osservato, si verifica quando i totali marginali sono altamente sbilanciati, ma in modo simmetrico (Tabb. III e IV). Questa situazione si verifica quando il totale ($a + b$) è molto diverso da quello ($c + d$) oppure quando il totale ($a + c$) è molto diverso da quello ($b + d$).

In pratica, si parla di sbilanciamento simmetrico nel caso in cui la prevalenza del totale dei positivi (presenza) sia maggiore di quello dei negativi (assenza) per entrambi gli osservatori: $(a + b)>(c + d)$ e $(a + c)>(b + d)$ oppure quando la prevalenza del totale dei positivi (presenza) è minore di quello dei negativi (assenza) per entrambi gli osservatori: $(a + b)<(c + d)$ e $(a + c)<(b + d)$. Invece, si parla di sbilanciamento asimmetrico sia nel caso in cui per un osservatore è maggiore il totale dei positivi e per l'altro quello dei negativi: $(a + b) >(c + d)$ e $(a + c)<(b + d)$ sia quando per un osservatore è minore il totale dei positivi e per l'altro quello dei negativi $(a + b)<(c + d)$ e $(a + c)>(b + d)$ (6).

In sostanza, si ha un effetto della prevalenza quando la proporzione dell'accordo sui casi positivi (*Sì*) è diversa da quella sui casi negativi (*No*). Se l'indice di prevalenza, calcolato come $|a\text{-}d|/N$ (differenza in valore assoluto sulla diagonale dell'**accordo/totale**), è alto, la prevalenza sui casi positivi è molto alta o molto bassa; in questa situa-

RIQUADRO 1 - Esempio di calcolo del Kappa di Cohen

Presenza/assenza di lesioni addominali valutate da due medici sulla base di tomografie computerizzate effettuate su 300 pazienti

Valori Osservati

	Medico - 1		
Medico - 2	*Presenza*	*Assenza*	*Totale*
Presenza	14	20	34
Assenza	24	242	266
Totale	38	262	300

Valori Attesi

	Medico - 1		
Medico - 2	*Presenza*	*Assenza*	*Totale*
Presenza	4.3	29.7	34
Assenza	33.7	232.3	266
Totale	38	262	300

A_o = 0.85
A_a = 0.79
Kappa di Cohen = 0.31

© 2016 Wichtig Publishing

TABELLA III - Esempio con totali marginali molto sbilanciati (bassa prevalenza)

		Osservatore 2		Totale
		SÌ	NO	
Osservatore 1	**SÌ**	**1**	**3**	**4**
	NO	**2**	**94**	**96**
Totale		**3**	**97**	**100**

$\mathbf{A_o}$ = (1 + 94)/100 = 0.95.
$\mathbf{A_a}$ = [(3*4) + (97*96)]/100^2 = (12 + 9312)/10000 = (9324/10000) = 0.9324.
Kappa = (0.95-0.9324)/(1-0.9324) = 0.26.

TABELLA IV - Esempio con totali marginali molto sbilanciati (alta prevalenza)

		Osservatore 2		Totale
		SÌ	NO	
Osservatore 1	**SÌ**	**94**	**2**	**96**
	NO	**3**	**1**	**4**
Totale		**97**	**3**	**100**

$\mathbf{A_o}$ = (94 + 1)/100 = 0.95
$\mathbf{A_a}$ = [(97*96) + (3*4)]/100^2 = (9312 + 12)/10000 = 0.9324
Kappa = (0.95-0.9324)/(1-0.9324) = 0.26

zione, l'accordo atteso, per effetto del caso, diventa alto e si riduce, di conseguenza, il valore di Kappa, che misura il grado di accordo oltre il caso. L'effetto della prevalenza sul Kappa è maggiore per grandi valori di Kappa (7).

Quando si passa a considerare l'accordo tra due valutatori che classificano un campione di soggetti secondo le modalità di una variabile ordinale con più di due categorie, bisogna tenere conto del fatto che una differenza di giudizio tra gli osservatori pari a una sola categoria di distanza indica una discordanza inferiore rispetto a quella rappresentata da una differenza pari a due o a tre categorie.

In tale situazione, l'indice di accordo adeguato da utilizzare è il cosiddetto "Kappa ponderato" che, nel misurare la concordanza, attribuisce ai soggetti un peso, p_i, in base alle ***i*** categorie di differenza di classificazione fra i giudici. Nel caso di concordanza fra i giudici, il peso assume un valore unitario, cioè $p_0 = 1$.

In presenza di una variabile di classificazione ordinale con k categorie, la discordanza massima osservabile sarà pari a *k-1* categorie; in tal caso, il peso attribuito ha valore pari a *0*, cioè nullo (8).

Di solito, i pesi utilizzati per i valori intermedi sono equispaziati e sono definiti con la formula:

$$p_i = 1 - \frac{i}{k-1}$$

La concordanza osservata ponderata sarà data da:

$$A_{op} = \frac{\sum p_i \times s_{io}}{n}$$

dove s_{io} : numero di soggetti osservati per i quali i giudici differiscono di ***i*** categorie.

La concordanza attesa ponderata sarà data da:

$$A_{ap} = \frac{\sum p_i \times s_{ia}}{n}$$

dove s_{ia}: numero di soggetti attesi, per effetto del caso, per i quali i giudici differiscono di ***i*** categorie.

Il Kappa ponderato sarà dato da:

$$Kappa\ ponderato = \frac{A_{op} - A_{ap}}{1 - A_{ap}}$$

Per il calcolo del Kappa ponderato si deve fare riferimento, quindi, a una tabella predefinita di pesi (matrice dei pesi) che misurano il grado di accordo tra i due valutatori (più distanti sono i giudizi più bassi sono i pesi assegnati). La tabella dei pesi deve essere una matrice simmetrica con gli elementi sulla diagonale principale che hanno tutti valore uno (dove c'è accordo tra i due giudici) e valori decrescenti positivi al di fuori della stessa (RIQUADRO 2).

Interpretazione del Kappa

Nel caso di concordanza perfetta fra valutatori, il Kappa assume un valore pari a 1 (valore massimo), mentre, in caso di concordanza solamente casuale, assume un valore pari a zero (indipendenza statistica). In presenza di totale disaccordo tra gli osservatori, il Kappa non assume il valore di -1. Per tale motivo la statistica Kappa è una misura del grado di accordo e non di disaccordo (9).

Esistono diverse classificazioni internazionali utilizzate per valutare la riproducibilità di un *test* o di accordo tra osservatori, attraverso l'interpretazione del grado di concordanza misurato dalla statistica Kappa.

Di seguito si riporta la classificazione della concordanza in base al valore del Kappa proposta da Landis JR e Koch GG (1977) (10):

Statistica Kappa	Forza della Concordanza
< 0	*Poor* (Nulla)
0-0.20	*Slight* (Scarsa)
0.21-0.40	*Fair* (Modesta)
0.41-0.60	*Moderate* (Moderata)
0.61-0.80	*Substantial* (Sostanziale)
0.81-1	*Almost Perfect* (Quasi perfetta)

© 2016 Wichtig Publishing

RIQUADRO . 2 - Esempio di calcolo del Kappa ponderato.

Classificazione della gravità clinica di 100 pazienti (bassa, media, alta) effettuata da due medici

Valori Osservati

Medico - 2 \ Medico - 1	*Alta*	*Media*	*Bassa*	*Totale*
Alta	32	12	4	**48**
Media	8	20	2	**30**
Bassa	6	0	16	**22**
Totale	**46**	**32**	**22**	**100**

Matrice dei pesi

	Alta	*Media*	*Bassa*
Alta	1	0.5	0
Media	0.5	1	0.5
Bassa	0	0.5	1

Valori Attesi

Medico - 2 \ Medico - 1	*Alta*	*Media*	*Bassa*	*Totale*
Alta	22.08	15.36	10.56	**48**
Media	13.80	9.60	6.60	**30**
Bassa	10.12	7.04	4.84	**22**
Totale	**46**	**32**	**22**	**100**

$A_{op} = 0.79$

$A_{ap} = 0.58$

Kappa ponderato = 0.50

Nel caso di patologie di difficile diagnosi, come può esserlo quella di NEC (enterocolite necrotizzante) nei neonati altamente pretermine (<32 settimane di età gestazionale), il grado di accordo tra medici, sulla base di un singolo segno radiologico, può essere scarso. Informazioni cliniche aggiuntive e l'analisi di più di un segno radiologico possono ridurre il margine di errore dell'osservatore, che scaturisce, principalmente, dall'alta soggettività di valutazione e dalla bassa specificità del segno radiologico.

In tale contesto, il ricorso ad analisi più complesse di tipo multidimensionale, come quelle costituite dalle tecniche di segmentazione gerarchica, può notevolmente migliorare l'individuazione di profili omogenei di segni radiologici più stringenti ai fini della diagnosi, nonché far emergere scuole diverse di interpretazione dei segni radiologici fra osservatori (11).

Disclosures

Financial support: No financial support was received for this submission.
Conflict of interest: The authors have no conflict of interest.

Bibliografia

1. Franco F, Di Napoli A. Introduzione alla valutazione di un test diagnostico: sensibilità, specificità, valore predittivo. Giornale di Tecniche Nefrologiche e Dialitiche. 2016;28(1):53-5.
2. Franco F, Di Napoli A. Rapporto di verosimiglianza del risultato positivo e negativo di un test diagnostico e teorema di Bayes. Giornale di Tecniche Nefrologiche e Dialitiche. 2016;28(2):134-6.
3. Franco F, Di Napoli A. Valutazione in parallelo e in serie di *test* diagnostici multipli. Giornale di Tecniche Nefrologiche e Dialitiche. 2016;28(3):212-4.
4. Sardanelli F, Di Leo G. Biostatistica in Radiologia. Milano: Springer. 2008:24-131.
5. Lu Y, Zhao S. Statistics used in quality control, quality assurance, and quality improvement in radiological studies. In: Lu Y, Fang J, a cura di, Advanced Medical Statistics (1st edition). River Edge, NJ: World Scientific. 2003:139-42.
6. Shoukri MM. Measures of interobserver agreement. Boca Raton, FL: Chapman & Hall/CRC Press. 2004:35-7.
7. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. https://www.ncbi.nlm.nih.gov/pubmed/8501467\o Journal of clinical epidemiology. J Clin Epidemiol. 1993;46:423-9.
8. Armitage P, Berry G. Statistica Medica, metodi statistici per la ricerca in Medicina (III edizione). Milano: Mc Graw-Hill. 1996:448-50.
9. Indrayan A. Medical biostatistics (2nd edition). Boca Raton, FL: Chapman & Hall/CRC Press. 2008:599.
10. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33:159-74.
11. Di Napoli A, Di Lallo D, Perucci CA, et al. Inter-observer reliability of radiologic signs of necrotizing enterocolitis in a population of high-risk newborns. Paediatr Perinat Epidemiol. 2004;18(1):80-7.

© 2016 Wichtig Publishing