ON AVOIDING PARADOXES IN ASSESSING INTER-RATER AGREEMENT

Rosa Falotico[1], Piero Quatto[2]

Department of Statistics, University of Milano-Bicocca, Milan, Italy

Abstract. *"Kappa" statistic has become a popular tool for measuring inter-rater agreement, despite its paradoxical behaviour. In this paper, we propose a procedure for measuring and testing the agreement among a set of judges that is based on a statistic not affected by "Kappa" paradoxes. It turns out that our procedure works well even in the case of small sample size.*

Keywords*: Inter-rater agreement, Cohen's Kappa, Fleiss' Kappa, Monte Carlo simulations.*

1. INTRODUCTION

The "Kappa" statistic was first proposed by Cohen (1960) for measuring agreement between two judges (also called "raters" or "observers"), who judge, independently, n subjects by means of a scale consisting of C categories. "Kappa" has become a widespread index, although it takes very low values even in situations of strong agreement. This paradoxical behaviour has been widely studied (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990; Lantz and Nebenzahl, 1996; Shoukri, 2004). On the contrary, very little attention has been devoted so far to such a problem affecting the statistic proposed by Fleiss (1971) as a generalization of the Cohen's "Kappa". A thorough review of inter-rater agreement coefficients is given by Krippendorff (1970; 1978; 2004).

The aim of this paper is threefold: firstly, to suggest an alternative statistic not affected by "Kappa" paradoxes (Quatto, 2004); secondly, to propose a procedure for testing chance agreement among a set of judges that is based on the alternative statistic and works well even in the case of small samples; thirdly, to highlight the inadequacy of "Kappa" statistic by means of an application in the context of psychiatric diagnoses (Fleiss, 1971).

[1] Rosa Falotico, email: rosa.falotico@unimib.it
[2] Piero Quatto, email: piero.quatto@unimib.it

2. "KAPPA" STATISTICS

Let n represents the number of subjects, M the number of judgments per subject and C the number of exhaustive and mutually exclusive categories into which assignments are made. Let x_{ij} be the number of judges who assigned the i-th subject (i = 1, ..., n) to the j-th category (j = 1, ..., C), as in Tab. 1.

Table 1: Assignment of subjects to categories

Subject	**Category**					**Tot.**
	1	…	j	…	C	
1	x_{11}	…	x_{1j}	…	x_{1C}	$x_{1.} = M$
⋮	⋮		⋮		⋮	⋮
i	x_{i1}	…	x_{ij}	…	$x_{i\,C}$	$x_{i.} = M$
⋮	⋮		⋮		⋮	⋮
n	x_{n1}	…	x_{nj}	…	$x_{n\,C}$	$x_{n.} = M$
Tot.	$x_{.1}$	…	$x_{.j}$	…	$x_{.\,C}$	nM

We may note that in the case of quantitative variables intraclass correlation coefficient is used instead of Cohen's K.

Tab. 1 shows that the row total

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

equals the number of judges and the column total

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

equals the overall number of assignments to category j. So, the proportion of agreeing pairs out of all the possible pairs of assignments for subject i is

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

(Friedman, 1920) and we may notice that its distribution is given by Rizzi (1962). Then the overall agreement can be measured with the mean

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

(Fleiss, 1971; Fleiss *et al*., 2003).

2.1 FLEISS' "KAPPA"

According to Fleiss (1971), if we assume that the proportion

$$p_j = \frac{x_{.j}}{nM} = \frac{1}{nM}\sum_{i=1}^{n} x_{ij}$$

represents the probability of random assignment to the category j, then the chance-expected agreement is

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

Therefore, the normalized agreement corrected for the amount expected by chance leads to the statistic

$$K_{\text{Fleiss}} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \in \left[-\frac{1}{M-1}, 1\right], \tag{3}$$

proposed by Fleiss (1971) as a generalization of the Cohen's "Kappa" (Cohen, 1960) described in Section 2.2.

2.2 COHEN'S "KAPPA"

In order to introduce the "Kappa" index due to Cohen (1960), we consider the special case of $M = 2$ judges. Let M_{jk} be the number of subjects assigned by the first judge to category j (j = 1, ..., C) and by the second judge to category k (k=1, ..., C), as in Tab. 2.

Table 2: Contingency table in the case of two judges

Judge 1	*Judge 2*						
	Category	1	…	k	…	C	**Tot.**
	1	M_{11}	…	M_{1k}	…	M_{1C}	$M_{1.}$
	⋮	⋮		⋮		⋮	⋮
	j	M_{j1}	…	M_{jk}	…	M_{jC}	$M_{j.}$
	⋮	⋮		⋮		⋮	⋮
	C	M_{C1}	…	M_{Ck}	…	M_{CC}	$M_{C.}$
	Tot.	$M_{.1}$	…	$M_{.k}$	…	$M_{.C}$	n

Putting

$$p_{jk} = \frac{M_{jk}}{n},$$

we can define the proportion of subjects assigned by the two judges to the same category

$$p_o = \sum_{j=1}^{C} p_{jj} = \frac{1}{n}\sum_{j=1}^{C} M_{jj}$$

and the agreement expected under the hypothesis of independence of the assignments

$$p_e = \sum_{j=1}^{C} p_{j.} p_{.j} = \frac{1}{n^2}\sum_{j=1}^{C} M_{j.} M_{.j}.$$

Normalizing with the maximum, we obtain the Cohen's "Kappa"

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

This index can assume very low values, near the lower bound of the range, in spite of an enhanced agreement between the two judges. This paradoxical behaviour has been widely studied (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990; Lantz and Nebenzahl, 1996; Shoukri, 2004).

Besides, we can see that Cohen's index is not a particular case of Fleiss' index. Indeed for M=2, applying the relations

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{n}\sum_{j=1}^{C}\sum_{i=1}^{n} P_{ij} = \frac{1}{n}\sum_{j=1}^{C} M_{jj} = p_o$$

and

$$\bar{P}_e = \sum_{j=1}^{C} p_j^2 = \frac{p_e}{2} + \frac{1}{4n^2}\sum_{j=1}^{C}\left(M_{j.}^2 + M_{.j}^2\right),$$

as shown by Krippendorf (1970), the Fleiss' index coincide with Scott's index p (Scott, 1955)

$$K_{\text{Fleiss}} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{p_o - \bar{P}_e}{1 - \bar{P}_e} = \pi_{\text{Scott}}.$$

Therefore, the Fleiss' index equals Cohen's index if and only if the marginals of Table 2 are the same:

$$K_{\text{Fleiss}} = K_{\text{Cohen}} \Leftrightarrow \bar{P}_e = p_e \Leftrightarrow \forall j \;\; M_{j.} = M_{.j}.$$

Thus, Fleiss' "Kappa" cannot be considered the generalization of Cohen's "Kappa" with M >2 judges (Fabbris and Gallo, 1993).

3. AN ALTERNATIVE STATISTIC

We assume the assignments of each subject to categories are random and independent and all categories are equally likely.

Thus, the chance-expected agreement can be expressed by the sum of the C probabilities of getting a pair of random assignments to the same category, given by

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

Finally, eliminating the chance-expected agreement (4) from the observed one (1) and normalizing, we obtain the S statistic (Quatto, 2004)

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

This index is a generalization of that proposed by Bennett, Alpert and Goldstein (1954). Furthermore, S index has the same structure and the same range of "Kappa" statistic without its paradoxes. As shown in Section 4,such paradoxes are due to overestimation of expected agreement made by the "Kappa" statistic (Quatto, 2004).

In addition, if we assume that the n groups of judges are independent Bernoulli samples of size M (Fleiss, 1971), then, under the null hypothesis H_0 of random assignment, each row of Table 1 is distributed as a multinomial with parameters M e $\pi_j = 1/C$ (j=1,…, C). In this case we obtain the following asymptotic distributions of S statistic (Quatto, 2004):

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

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

More specifically, the null hypothesis states that judges make random assignments, regardless of the characteristic of each subject. So, H_0 corresponds to a high percentage of assignment errors, namely there is a high number of subjects so that it turns out to be difficult to classify and the expected agreement is nil.

It is possible to construct a test that rejects H_0 for large values of S by means of the *p*-value given by

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

where *s* is the observed value of *S*. This *p*-value can be approximated by means of a Normal distribution when the number of subjects is large or by a Chi-square distribution when the number of judges is large.

In order to calculate critical values in the case of small sample sizes, we apply Monte Carlo method. For this purpose, we set the parameters *C*, *n* and *M* and we draw a sample from the null distribution given by a Multinomial distribution with parameters *M* and $\pi_j = 1/C$, $j = 1, \ldots, C$. So, we are able to calculate the statistic (5). Repeating that exercise *T* times, we determine critical values $s_{M,n}$, at the level of significance α, as $100 \times (1- \alpha)$-th percentile of the *T* Monte Carlo replications.

Tab. 3 provides Monte Carlo approximations (for $T = 1000$) of critical values $s_{M,n}$, that can be compared with the asymptotic critical values obtained through (6) and reported in Tab. 4.

We can see that there are no important differences between the asymptotic and Monte Carlo critical values. Moreover, these differences decrease when the number of judged subjects increases.

Table 3: Monte Carlo critical values for significance level $\alpha = 0.05$

	Monte Carlo critical values					
	***M*=2**	***M*=4**	***M*=6**	***M*=8**	***M*=10**	***M*=12**
***n* = 10**	0.250	0.104	0.083	0.054	0.042	0.034
***n* = 20**	0.188	0.083	0.054	0.038	0.028	0.023
***n* = 30**	0.167	0.062	0.042	0.030	0.021	0.018
***n* = 40**	0.125	0.057	0.035	0.026	0.019	0.017
***n* = 50**	0.100	0.050	0.032	0.023	0.018	0.016
***n* = 60**	0.104	0.045	0.031	0.022	0.017	0.014
***n* = 70**	0.107	0.042	0.026	0.019	0.015	0.013

Table 4: Asymptotic critical values for significance level $\alpha = 0.05$

	Asymptotic critical values					
	***M*=2**	***M*=4**	***M*=6**	***M*=8**	***M*=10**	***M*=12**
***n* = 10**	0.260	0.106	0.067	0.049	0.039	0.032
***n* = 20**	0.184	0.075	0.047	0.035	0.027	0.023
***n* = 30**	0.150	0.061	0.039	0.028	0.022	0.018
***n* = 40**	0.130	0.053	0.034	0.025	0.019	0.016
***n* = 50**	0.116	0.047	0.030	0.022	0.017	0.014
***n* = 60**	0.106	0.043	0.027	0.020	0.016	0.013
***n* = 70**	0.098	0.040	0.025	0.019	0.015	0.012

4. AN APPLICATION

Using the classification data of Tab. 5 (Fleiss, 1971), concerning 30 patients from 6 psychiatrists classified into 5 diagnostic categories, we calculate the following values

$$\bar{P} = 0.556 \qquad \bar{P}_e = 0.220 \qquad K = 0.430 \qquad S = 0.444$$

showing a slight difference between "Kappa" statistic and *S*.

According to the asymptotic distribution (6), the proposed test of significance rejects the null hypothesis of chance agreement if

$$s\sqrt{nM(M-1)(C-1)/2} \geq z_{1-\alpha}$$

where *s* is the determination of statistic (5) and $z_{1-\alpha}$ is the (1- α) standard Normal quantile. In particular, the *p*-value provided by

$$\tilde{\alpha} = 1 - \Phi\left(s\sqrt{\frac{nM(M-1)(C-1)}{2}} \right),$$

is close to zero, and this allows to reject H_0.

On the other hand, merging the last three categories of Table 5, the new values

$$\bar{P} = 0.640 \qquad \bar{P}_e = 0.574 \qquad K = 0.205 \qquad S = 0.460$$

show that "Kappa" does not mirror the increase in the level of agreement among judges, while the increase is highlighted by *S*.

5. CONCLUSIONS

The S statistic allows to measure the level of inter-rater agreement without incurring the paradoxes of "Kappa" statistic. Furthermore, *S* statistic enables to test the null hypothesis of random assignment both when the number of subjects is large and when the number of judges is large. Finally, we can employ Monte Carlo method to calculate critical values when either size is small.

Table 5: Frequency of assignment of patients to diagnostic categories (Source: Fleiss, 1971)

Subject	***Diagnostic categor***				
	Depression	*Personality disorders*	*Schizophrenia*	*Neurosis*	*Other*
1	0	0	0	6	0
2	0	3	0	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

References

Bennet, E.M., Alpert, R. and Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, (18): 303-308.

Cicchetti, D.V. and 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.

Fabbris, L. and Gallo, F. (1993). Bivariate coefficients of agreement among any number of observers. *Educational and Psychological Measurement*, (53): 337-349.

Feinstein, A.R. and 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. and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Hoboken.

Friedman, W.F. (1920). *The Index of Coincidence and Its Application in Cryptography*, The Riverbank Publications, Aegean Park Press, Laguna Hills.

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

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data". In: Borgatta, E.F. *Sociological Methodology*. San Francisco: Jossey-Bass: 139-150.

Krippendorff, K. (1978). Reliability of binary attribute data. *Biometrics*, 34: 142-144.

Krippendorff, K. (2004). *Analysis Introduction Methodology*. Thousand Oaks, CA: Sage.

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

Rizzi, A. (1962). Su un test statistico basato sulle frequenze. *Atti della XXI Riunione Scientifica della Società Italiana di Statistica*, Roma, 27-28 ottobre 1962.

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.