Qual Quant (2015) 49:463–470
DOI 10.1007/s11135-014-0003-1

Fleiss' kappa statistic without paradoxes

Rosa Falotico · Piero Quatto

Published online: 13 February 2014

Abstract The Fleiss' kappa statistic is a well-known index for assessing the reliability of agreement between raters. It is used both in the psychological and in the psychiatric field. Unfortunately, the kappa statistic may behave inconsistently in case of strong agreement between raters, since this index assumes lower values than it would have been expected. The aim of this paper is to propose a new method to avoid this paradox through permutation techniques. Furthermore, we study the problem of kappa confidence intervals and, in particular, we suggest to use Bootstrap confidence intervals free of paradoxes.

Keywords Inter-rater agreement · Fleiss' kappa · Kappa paradoxes · Monte Carlo simulations · Bootstrap confidence intervals

1 Introduction

The kappa statistic was proposed by Cohen (1960) to measure the agreement between two raters (also called "judges" or "observers"), independently judging n subjects through a scale consisting of q categories. Kappa has become a well known index for the comparison of expert advices, especially in the psychometric field (Uttal et al. 2013; Harvey and Tang 2012; Markon et al. 2011; Östlin et al. 1990).

A comprehensive review of inter-rater agreement coefficients has been put forth by Gwet (2008) and Dijkstra and Eijnatten (2009).

The use of Cohen's kappa statistic has been increasing despite two important paradoxes (Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990): (i) the presence of high levels of raters' agreement with low kappa values (related to prevalence of the trait in the sample) and (ii) the lack of predictability of changes in the statistic with different marginals (due to the symmetry of rates in the disagreement categories). This paradoxical behaviour has

R. Falotico (✉) · P. Quatto
Department of Economics, Management and Statistics, University of Milan-Bicocca,
Piazza Ateneo Nuovo 1, 20126 Milano, Italy
e-mail: rosa.falotico@unimib.it

been widely studied (Cicchetti and Feinstein 1990; Feinstein and Cicchetti 1990; Lantz and Nebenzahl 1996; Shoukri 2004).

On the contrary, very little attention has been devoted so far to a similar problem affecting the statistic proposed by Fleiss (1971) as a multiple-raters generalization of the Cohen's kappa. As a matter of fact, in specific situations, Fleiss' kappa takes very low values even in the case of high agreement.

This paradox is due to the fact that this measure of agreement for nominal scales is not invariant under permutation of categories. In order to solve this problem, we propose a permutation-invariant version of Fleiss' kappa that is not affected by the paradox.

Since the problem depends on particular combination of category assignment and the scale is nominal, we apply permutation techniques without loss of information. In particular, we permute the dataset, we calculate Fleiss' kappa on each "permuted" dataset and we synthesize the results with a robust statistic.

In Sect. 2 we describe Fleiss' statistic, in Sect. 3 we discuss its paradoxical behaviour and in the subsequent Sections we show a method to solve the problem of paradoxes through the combined use of permutation techniques and resampling methods.

2 Fleiss' kappa statistic

We consider an inter-rater reliability study with n subjects and r rates per subject. All raters have to assign each subject in one of q exhaustive and mutually exclusive categories.

These studies involve raters who are experts in a given area (e.g. physicians—in particular psychologists and psychiatrists—archaeologists, art critics, judges, etc.). It is possible to quantify the agreement among observers who have participated to a survey.

Table 1 shows the frequency distribution of r raters by n subjects and q response categories: r_{ij} represents the number of rates assigning the ith subject ($i = 1, ..., n$) to the jth category ($j = 1, ..., q$).

In Table 1, the marginal distribution $r_{i\cdot} = \sum_{j=1}^{q} r_{ij} = r$ provides the total number of raters and the marginal $r_{\cdot j} = \sum_{i=1}^{n} r_{ij}$ provides the total number of assignments to category j.

When two or more raters agree in assigning the subject i to category j, then the agreement among raters is showed by the corresponding frequency in Table 1: $r_{ij} \geq 2$.

Using the binomial coefficient, we determine the number of concordant pairs:

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

Table 1 Distribution of raters by subject and response category

Subject	Category					Tot.
	1	⋯	j	⋯	q	
1	r_{11}	⋯	r_{1j}	⋯	r_{1q}	$r_{1\cdot} = r$
⋮	⋯	⋯	⋯	⋯	⋯	⋮
i	r_{i1}	⋯	r_{ij}	⋯	r_{iq}	$r_{i\cdot} = r$
⋮	⋯	⋯	⋯	⋯	⋯	⋮
n	r_{n1}	⋯	r_{nj}	⋯	r_{nq}	$r_{n\cdot} = r$
Tot.	$r_{\cdot 1}$	⋯	$r_{\cdot j}$	⋯	$r_{\cdot q}$	rn

We define the proportion of pairs of concordant raters assigning subject i to category j as:

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

Hence, we can calculate the proportion of concordant pairs for the ith subject for all the $r(r-1)$ possible pairs of assignments:

$$P_i = \sum_{j=1}^{q} p_{ij} = \sum_{j=1}^{q} \frac{\binom{r_{ij}}{2}}{\binom{r}{2}} = \frac{1}{r-1}\left(\frac{1}{r}\sum_{j=1}^{q} r_{ij}^2 - 1\right).$$

The overall agreement can be measured referring to Fleiss (1971) and Fleiss et al. (2003):

$$\bar{P} = \frac{1}{n}\sum_{j=1}^{n} P_i = \frac{1}{r-1}\left(\frac{1}{nr}\sum_{i,j} r_{ij}^2 - 1\right). \tag{1}$$

In general, a subject is considerate deterministically assigned to a category when, repeating several times the judgment, none of the raters changes its categorization. On the other hand, the categorization of a subject is defined as random, in case it does not depend on a shared evaluation, but is only due to chance.

The overall agreement of two or more raters has indeed to be interpreted as the observable effect of the combination of two non-observable factors: a deterministic factor and a random factor. In order to isolate the deterministic component (the object of our study), we have firstly to define the chance-agreement probability.

According to Scott (1955) and Fleiss (1971), the probability of agreement due to chance is given by the following proportion:

$$p_j = \frac{r_{\cdot j}}{nr} = \frac{1}{nr}\sum_{i=1}^{n} r_{ij}$$

and the random expected agreement is given by:

$$\bar{P}_e = \sum_{j=1}^{q} p_j^2 \in \left[\frac{1}{q}, 1\right]. \tag{2}$$

If we correct the overall agreement probability (1) for the agreement probability due to chance (2) and normalize, we obtain the statistic:

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

proposed by Fleiss (1971) as a generalization of Cohen's kappa (1960).

With respect to this point, it should be noted that Fleiss' kappa is the multiple-raters extension of Scott's π index (Scott 1955; Gwet 2008) and not of Cohen's kappa. Fleiss' kappa is one of the most common indices to quantify multiple-raters agreement (Fleiss et al. 2003), but in practice it could return inconsistent results.

3 Paradoxical behaviour of Fleiss' kappa

In Table 2 we describe a particular case of poor performance of Fleiss' kappa. All subjects are distributed in the first two categories, in equal proportion. Let M be an integer between 0 and

Table 2 Distribution of raters by subject and response category leading to the paradoxical behaviour of Fleiss' kappa

Subject	Category					Tot
	1	2	3	$\cdots$	q	
1	M	$r-M$	0	$\cdots$	0	r
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
i	M	$r-M$	0	$\cdots$	0	r
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
n	M	$r-M$	0	$\cdots$	0	r
Tot.	nM	$n(r-M)$	0	$\cdots$	0	rn

r. When M varies, we change from a situation of complete agreement among the examiners (produced by the extreme values $M = 0$ and $M = r$) to a situation of minor agreement (in correspondence of the intermediate values of M).

Expression (3), applied to Table 2, returns:

$$K_{Fleiss} = -\frac{1}{r-1}, \quad \text{for } 0 < M < r$$

and

$$K_{Fleiss} \to -\frac{1}{r-1}, \quad \text{for } M \to 0 \text{ or } M \to n$$

These results show the inadequacy of Fleiss' kappa in interpreting high level of agreement, because this index assumes a constant and negative value even when it would be expected a very high inter-rater agreement.

Fleiss' kappa does not allow to recognize different degrees of agreement when M varies. Moreover, it does not allow to discriminate the situations of perfect agreement from other situations. For instance, setting $M = 5$ and $r = 6$ in Table 2, it is obtained:

$$K_{Fleiss} = -0.2,$$

even if five out of six raters totally agree in their judgement.

Another example of Fleiss' kappa inconsistent behaviour can be showed using Table 4 (Fleiss 1971). This table shows the classification of 30 patients into 5 diagnostic categories by six psychiatrists. When we calculate Fleiss' kappa on these data, we obtain:

$$K_{Fleiss} = 0.430.$$

Merging the last three categories in a new category, we expect an increased agreement, instead the value of kappa decreases:

$$K_{Fleiss} = 0.205.$$

4 Calculating Fleiss' kappa without paradoxical behaviour

In Sect. 3 we noted some particular configurations of category assignments and we analysed the cases in which the kappa statistic underestimated the agreement. The basic idea of our

work consists in the use of permutation techniques (Mielke and Berry 2007) to solve the problem of paradoxical behaviour. Permutations do not lead to loss of information on agreement since we only consider categorical data.

Referring to Table 1, we propose, in correspondence of each row i, to randomly choose a permutation of the q frequencies r_{ij} and to substitute this new vector instead of the original vector. Finally, it is possible to calculate the Fleiss' kappa index on the *permuted frequency* table.

Repeating this procedure C times and synthesizing the C values of kappa by means of a robust index, we better quantify the inter-raters agreement. In particular, we propose the use of the median to synthesize the repeated permutation results.

As far as the choice of C is concerned, the number of all possible permuted tables, starting from Table 1, is equal to $(q!)^n$. This number can be too large for a comprehensive examination, hence, we approximate the result with a smaller number C.

We calculate the value of robust kappa, K_r, applying this technique to diagnostic data in Table 4. We now show how the proposed permutation method solves Fleiss' kappa paradoxes.

From the original dataset, we calculate the value:

$$K_r = 0.436$$

and from the merged dataset:

$$K_r = 0.454.$$

This result shows that K_r, differently from Fleiss' kappa, detects the increase of inter-rater agreement. The suggested procedure involves a high computational effort. When the size of the table and the number of iterations increase, the required computational time increases dramatically.

From the proposed method to calculate Fleiss' kappa, it also follows a new method to calculate confidence intervals not affected by paradoxes. In the following Section, we show how to construct an interval estimator based on Bootstrap techniques, according to the procedure proposed for robust kappa.

5 Bootstrap confidence intervals for robust kappa

In this Section we propose the joint use of permutation techniques and resampling methods to construct confidence intervals. The proposed Bootstrap intervals, differently from the standard one (Fleiss et al. 2003), avoid paradoxes and perform well even in case of a small number of subjects (n).

Let's indicate with $p_{ij} = \frac{r_{ij}}{r}$ (where $i = 1, \ldots, n$ and $j = 1, \ldots, q$) the proportion of categorization. The resampling of the ith row has a multinomial distribution with parameters

Table 3 Confidence intervals (at level of 95 %) for Fleiss' kappa

	Original dataset		Merged dataset	
	Inf	Sup	Inf	Sup
Asymptotic	0.382	0.478	0.135	0.274
Percentile	0.338	0.550	0.340	0.583
Bootstrap-t	0.298	0.606	0.337	0.588
Bca	0.340	0.551	0.336	0.573

Table 4 Frequency of assignment of patients to diagnostic categories (source: Fleiss 1971)

Subject	Diagnostic category				
	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

r and $p_{ij} \ldots p_{iq}$ (with $i = 1, ..., n$). We can apply to each resampled table the algorithm described in Sect. 4. In case we repeat this procedure B times, we obtain B values of robust Fleiss' kappa. Considering the resulting distributions, we can calculate the quantiles of order α and $1-\alpha$, respectively representing the lower and upper bounds of the Bootstrap percentile interval at confidence level of $1 - 2\alpha$ (Shao and Tu 1995).

Because of the computational effort of permutations, we prefer to use Bootstrap percentile with respect to Bootstrap accelerated bias-corrected percentile (Bca) and Bootstrap-t, that are more accurate (i.e. second order accurate Shao and Tu 1995) but also computationally expensive. In order to assess the results obtained in constructing Bootstrap percentile

confidence intervals (which are first order accurate Shao and Tu 1995) at level of 95 %, we compare them to intervals obtained by Bca, Bootstrap-t and Fleiss–Levin–Paik asymptotic method (Fleiss et al. 2003).

According to Fleiss et al. (2003), for n large enough, Fleiss' kappa has a Normal distribution and the estimated standard error is:

$$s_k = \frac{\sqrt{2}}{\sum_{j=1}^{q} p_j \left(1-p_j\right)\sqrt{nr\left(r-1\right)}} \sqrt{\left[\sum_{j=1}^{q} p_j \left(1-p_j\right)\right]^2 - \sum_{j=1}^{q} p_j \left(1-p_j\right)\left(1-2p_j\right)}$$

From original dataset (Table 4) and merged dataset we obtain the confidence intervals reported in Table 3 (where $C = 100$ and $B = 1000$).

The bootstrap intervals constructed with the proposed method do not lead to the paradox of Fleiss' kappa and they do not need large sample size, although they require a certain computational effort.

In particular, we can observe that the humble number of subjects is the typical case in the psychometric field and it prevents from the use of the asymptotic approximations.

6 Concluding remarks

In this work we investigated the problem of the underestimation of agreement of Fleiss' kappa statistic in assessing high levels of inter-raters agreement. Since in case of nominal variables the order of the categories is not relevant, we proposed a solution based on permutation techniques.

In order to avoid the paradoxes of this index (exposed in Sect. 3), we suggest to permute any row of the original dataset. The new permuted matrix has a level of agreement quite similar to the original one, in spite of a different configuration. If we repeat this operation C times, the "new" datasets, and the corresponding values of Fleiss' kappa can be used to assess the level of agreement of the original dataset. In order to summarize the C values of Fleiss' kappa, we have proposed the use of a robust statistic (the median), less affected by extreme values, that cause the unexpected performance of Fleiss' kappa.

The problems of this statistic involve the corresponding confidence interval proposed by Fleiss et al. (2003); this interval is affected by the same paradoxes of Fleiss' kappa and it is based on an asymptotic Normal approximation (so it is valid only for n large enough). Therefore, we proposed a Bootstrap interval not affected by paradoxes and applicable even when n is too small for Normal approximation.

References

Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**(6), 551–558 (1990). doi:10.1016/0895-4356(90)90158-L

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960). doi:10.1177/001316446002000104

Dijkstra, L., van Eijnatten, F.M.: Agreement and consensus in a q-mode research design: an empirical comparison of measures, and an application. Qual. Quant. **43**(5), 757–771 (2009). doi:10.1007/s11135-009-9249-4

Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. J. Clin. Epidemiol. **43**(6), 543–549 (1990). doi:10.1016/0895-4356(90)90159-M

Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971). doi:10.1037/h0031619

Fleiss, J.L., Levin, B., Paik, M.C.: Statistical methods for rates and proportions. Wiley, Hoboken (2003)

Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. **61**(1), 29–48 (2008). doi:10.1348/000711006X126600

Harvey, A.G., Tang, N.K.: (Mis)perception of sleep in insomnia: a puzzle and a resolution. Psychol. Bull. **138**(1), 77 (2012). doi:10.1037/a0025730

Lantz, C.A., Nebenzahl, E.: Behavior and interpretation of the k statistic: resolution of the two paradoxes. J. Clin. Epidemiol. **49**(4), 431–434 (1996). doi:10.1016/0895-4356(95)00571-4

Markon, K.E., Chmielewski, M., Miller, C.J.: The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. Psychol. Bull. **137**(5), 856 (2011). doi:10.1037/a0023678

Mielke, P.J.W., Berry, K.J.: Permutation methods: a distance function approach. Springer, New York (2007)

Östlin, P., Wärneryd, B., Thorslund, M.: Should occupational codes be obtained from census data or from retrospective survey data in studies on occupational health? Soc. Indic. Res. **23**(3), 231–246 (1990)

Scott, W.A.: Reliability of content analysis: the case of nominal scale coding. Pub. Opin. Q. (1955). doi:10.1086/266577

Shao, J., Tu, D.: The jackknife and bootstrap. Springer, New York (1995)

Shoukri, M.M.: Measures of interobserver agreement and reliability. Chapman & Hall, Boca Raton (2004)

Uttal, D.H., Meadow, N.G., Tipton, E., Hand, L.L., Alden, A.R., Warren, C., Newcombe, N.S.: The malleability of spatial skills: a meta-analysis of training studies. Psychol. Bull. **139**(2), 352 (2013). doi:10.1037/a0028446