Asymptotic distributions of kappa statistics and their differences with many raters, many rating categories and two conditions

Biometrical Journal **52** (2017) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Asymptotic distributions of kappa statistics and their differences with many raters, many rating categories and two conditions.

Luca Grassano [*,1], **Guido Pagana** [2] [3], **Marco Daperno** [4], **Enrico Bibbona** [1], and **Mauro Gasparini** [1]

[1] Politecnico di Torino, Department of Mathematical Sciences *(Torino, Italy)*

[2] Politecnico di Torino, Department of Automatics and Informatics *(Torino, Italy)*

[3] Istituto Superiore Mario Boella *(Torino, Italy)*

[4] Ospedale Ordine Mauriziano di Torino Umberto I *(Torino, Italy)*

Received zzz, revised zzz, accepted zzz

In clinical research and in more general classification problems, a frequent concern is the reliability of a rating system. In the absence of a gold standard, agreement may be considered as an indication of reliability. When dealing with categorical data, the well-known kappa statistic is often used to measure agreement. The aim of this paper is to obtain a theoretical result about the asymptotic distribution of the kappa statistic with multiple items, multiple raters, multiple conditions and multiple rating categories (more than two), based on recent work. The result settles a long lasting quest for the asymptotic variance of the kappa statistic in this situation and allows for the construction of asymptotic confidence intervals. A recent application to clinical endoscopy and to the diagnosis of Inflammatory Bowel Diseases (IBDs) is shortly presented to complement the theoretical perspective.

Key words: Agreement; Correlated kappa statistics; Inflammatory Bowel Diseases; de Finetti representation theorem.

1 Introduction

Consider a situation where each one of N items (subjects, biopsies, etc…) is assigned by the same n raters (physicians, biologists, teachers etc…) to one out of K mutually exclusive levels of a categorical variate, possibly ordinal (disease type, diagnosis, class etc…), under two different conditions (treatment, times etc…).

If recording the true category for each item is not feasible (absence of gold standard), then a good level of agreement is often desirable in order to achieve certain conclusions about the reliability and reproducibility of the rating or to show improvement of one condition over another.

A widely used index of agreement is the kappa statistic, as introduced by Cohen (1960) and Fleiss (1971). We give explicit asymptotic distributions for the kappa statistic when the rating levels are more than two and similar asymptotic distributions for the difference of kappa statistics between two conditions.

1.1 The motivating case study

The motivation for this work comes from a group of gastroenterologists within the Italian IGIBDendo project who diagnose patients with inflammatory bowel diseases (IBDs) through endoscopy. The IBDs are chronic autoimmune non-infectious inflammatory conditions affecting the gastrointestinal tract and, in particular, the colon and small intestine. Ulcerative colitis (UC) and Crohn's disease (CD) are the principal types of IBD. Aiming at a more objective and robust evaluation, some scoring systems for endoscopic outcomes have been introduced to deal with the different diseases. Nevertheless, duplicability of endoscopic

*Corresponding author: e-mail: luca.grassano@studenti.polito.it, Phone: +39-340-282-2053

scoring systems used to categorize endoscopic exams is far from being optimal. Inter-rater agreement of non-dedicated gastroenterologists on IBD endoscopic scoring systems is explored in this work.

Among the different diseases and few scoring systems available, we focus on UC and on the so-called Mayo score, an ordinal factor taking values in $\{0, 1, 2, 3\}$ (the larger the score, the more severe the injuries observed in the endoscopy). Although the scale is an ordinal one, we have worked with it as if it were just categorical, since we were not confident assigning weights to the differences between categories, as it is usually done in the literature on weighted kappa statistics. The non-dedicated gastroenterologists assigned Mayo scores to the same patients in two different conditions, before and after a training event guided by more experienced specialists. Such design may have been affected by biases such as confounding with time and learning effects by the gastroenterologists, but in any case a primary issue was to evaluate whether the training event significantly improved rater agreement in terms of Mayo scores. More detailed motivations and results are described in Daperno *et al.* (2016), where the complete case study is described, similar analyses are performed for CD (with the related Rutgeerts score) and biannual pooled results are shown.

In this paper we focus on the relevant statistical methods, and in particular on an extension of results about certain kappa statistics, as described in the next section.

1.2 Correlated kappa statistics to measure difference of agreement

The most commonly used measure of agreement for categorical ratings is the kappa statistic, based on an original proposal by Cohen (1960) and having the general form

$$\hat{\kappa} = \frac{\hat{p}_o - \hat{p}_e}{1 - \hat{p}_e} \tag{1}$$

where $\hat{p}_0$ is the observed proportion of agreement and $\hat{p}_e$ is an estimate of the expected agreement due to chance alone. The definition of observed and chance-expected agreement depends on the different sampling scenarios, and in particular on the number of categories, whether two or more raters are considered, which raters rate which subjects and so on.

Following our motivating example, in this paper the same subjects are rated by the same raters in two different conditions, creating a strong correlation between the two kappa statistics. The difference between the two kappa statistics is a measure of how the agreements among raters differ in the two conditions. Subjects give rise to independent observations, each subject being characterized by a certain level of disease status and by a personal profile which controls the probability of being given the different ratings; naturally such probability, for each subject, may also change between the first and the second condition. On the other hand, given a specific subject and a specific condition, the ratings provided by the raters are (conditionally) i.i.d., since in this situation a group of gastroenterologists with similar expertise is considered.

A very similar situation has been considered in this journal by Cao *et al.* (2016), who considered ratings into $K = 2$ possible categories only. We were able to extend their methods to more than two possible categories, settling in this way a long lasting quest for the asymptotic variance of the kappa statistic with many categories and many raters. Working with more than two categories is important in the biomedical literature as well as in engineering applications, where for example production faults can be classified as "machine related", "operator related", "material related" and so on (see e.g. De Mast (2007)). The literature on the various developments and variants of kappa is reviewed in Chapter 18 of the reference by Fleiss *et al.* (2003) and in Congalton and Green (2008). See also Gwet (2008) for an alternative randomization approach different from our model-based approach.

We therefore follow closely the symbolism in Cao *et al.* (2016), defined in the next section, and their techniques, to extend the relevant asymptotic distributions to the case of $K \geq 2$ categories, applying them to both one-condition (Section 2.1) and two-condition (Section 2.2) agreement problems. Section 3.1 contains MonteCarlo results which confirm the validity of our theoretical findings and Section 3.2 gives the

relevant results for the IGIBDendo project. Since the extension to $K \geq 2$ categories is rather technical, proofs of theorems are stored into the online Supplementary Material together with all R programs supporting our simulations and our results.

2 Asymptotic distributions of kappa statistics

2.1 The kappa statistic for a single condition

Let $X_{ijc}, i = 1, \ldots, N; j = 1, \ldots, n; c = 1, \ldots, K$ be the indicator that subject i has been assigned score c by rater j and let $n_{ic} = \sum_{j=1}^{n} X_{ijc}$ be the number of raters classifying subject i into class c. Then, for each i, the vector $n_{i\cdot} = (n_{i1}, \ldots, n_{iK})'$ is multinomial with parameters n and $\pi_{i\cdot} = (\pi_{i1}, \ldots, \pi_{iK})'$, where $\pi_{ic}, c = 1, \ldots, K$ is the probability that the $i-$th subject is assigned score c by the generic rater. Equality of the (conditional) distributions of the ratings for a given subject can be described as the absence of rater bias, i.e. the raters are homogeneous and none of them polarizes the ratings differently from the other raters. This assumptions is rather reasonable for our setup, but it is different from Davies and Fleiss (1982), who allow for different distributions for the different raters and provide only numerical solutions. The only overlap between the model presented here and Davies and Fleiss (1982) is the absence of rater agreement, i.e. $\kappa = 0$; further comments below on this uninteresting case.

Following therefore Cao *et al.* (2016), in our setup it is sensible to define the observed proportion of agreement as

$$\hat{p}_{o,n} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{K} \binom{n_{ic}}{2} / \binom{n}{2} = \frac{\sum_{i=1}^{N} \sum_{c=1}^{K} n_{ic}^2 - Nn}{Nn(n-1)}, \tag{2}$$

with $\binom{n}{2} = 0$ for $n < 2$, and the estimated probability of chance agreement as

$$\hat{p}_{e,n} = \sum_{c=1}^{K} (\frac{1}{nN} \sum_{i=1}^{N} n_{ic})^2. \tag{3}$$

At this point our notation differs from Cao *et al.* (2016), simplifying it quite a bit; their results coincide in any case with ours when $K = 2$. Define then

$$p_o = \frac{1}{N} \sum_{c=1}^{K} \sum_{i=1}^{N} \pi_{ic}^2$$

$$\bar{\pi}_c = \frac{1}{N} \sum_{i=1}^{N} \pi_{ic}, \qquad c = 1, \ldots, K$$

$$p_e = \sum_{c=1}^{K} \bar{\pi}_c^2,$$

and the population kappa parameter as

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{4}$$

Notice in particular that we can define κ for any finite N and, if appropriate, as $N \to \infty$. We now easily obtain the following laws of large numbers:

$$\hat{p}_{o,n} \overset{n\to\infty}{\longrightarrow} p_o \tag{5}$$

$$\hat{p}_{e,n} \overset{n\to\infty}{\longrightarrow} p_e \tag{6}$$

$$\hat{\kappa}_n \overset{n\to\infty}{\longrightarrow} \kappa \tag{7}$$

almost surely for any finite N, by the almost sure convergence of the (conditional) multinomial frequencies for any given subject. The population kappa κ is a measure of the heterogeneity of the rating probabilities $\pi_{i\cdot}$ of the different subjects. It can be interpreted as a measure of performance of the rating system in evaluating subjects since the more heterogeneous the subjects are, the more systematic agreement we will have among raters. On the other hand, a sufficient condition for the population kappa to be null is identical subjects: if $\pi_{ic} = \pi_c = \bar{\pi}_c,\ \ i = 1, \ldots, N,\ c = 1, \ldots, K$, then $p_o = p_e$ and $\kappa = 0$; with i.i.d. ratings on identical subjects any observed agreement would be due to chance. The population kappa is not dependent on sampling variability: it is a parameter, a fixed number which plays a key role in our modeling approach and an estimand of primary importance.

The asymptotic distribution of the kappa statistic $\hat{\kappa}$ can now be derived as in the following theorem.

Theorem 2.1 *Assuming that, given subject i, the ratings are i.i.d. with rating probabilities $\pi_{i\cdot} = (\pi_{i1}, \ldots, \pi_{iK})'$ and that the ratings across different sujects are independent, the following normal asymptotic results hold:*

$$\sqrt{n}\left[\begin{pmatrix}\hat{p}_{o,n}\\ \hat{p}_{e,n}\end{pmatrix} - \begin{pmatrix}p_o\\ p_e\end{pmatrix}\right] \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0, \begin{pmatrix}\sigma_{oo} & \sigma_{oe}\\ \sigma_{oe} & \sigma_{ee}\end{pmatrix}\right), \qquad \text{as } n \to \infty$$

with

$$\begin{aligned}
\sigma_{oo} &= \frac{4}{N^2}\sum_{i=1}^{N}\left(\sum_{c=1}^{K}\pi_{ic}^3(1-\pi_{ic}) - \sum_{c=1}^{K}\sum_{c'\neq c}\pi_{ic}^2\pi_{ic'}^2\right)\\
\sigma_{ee} &= \frac{4}{N^2}\sum_{i=1}^{N}\left(\sum_{c=1}^{K}\bar{\pi}_c^2\pi_{ic}(1-\pi_{ic}) - \sum_{c=1}^{K}\sum_{c'\neq c}\bar{\pi}_c\pi_{ic}\bar{\pi}_{c'}\pi_{ic'}\right)\\
\sigma_{oe} &= \frac{4}{N^2}\sum_{i=1}^{N}\left(\sum_{c=1}^{K}\bar{\pi}_c\pi_{ic}^2(1-\pi_{ic}) - \sum_{c=1}^{K}\sum_{c'\neq c}\pi_{ic}^2\bar{\pi}_{c'}\pi_{ic'}\right)
\end{aligned} \tag{8}$$

and the main result is

$$\sqrt{n}\,(\hat{\kappa}_n - \kappa) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau) \qquad \text{as } n \to \infty$$

with

$$\tau = \frac{\sigma_{oo}}{(1-p_e)^2} + \frac{\sigma_{ee}(1-p_o)^2}{(1-p_e)^4} - 2\frac{\sigma_{oe}(1-p_o)}{(1-p_e)^3}. \tag{9}$$

Proof. The proof is an application of the multivariate central limit theorem to the multinomial frequencies $n_{i\cdot}/n$ and of the multivariate delta method. The detailed but tedious computations are given in the Appendix as online supplementary material.

□

The result allows for the construction of approximate asymptotic tests and confidence intervals for the population kappa. The proof of Theorem 2.1 is contained as an Appendix in the online Supplementary Material.

2.2 The difference of kappa statistics between two conditions

When considering two different conditions A and B it becomes necessary to keep into account the dependency structure between ratings in the two conditions. Consequently, for each i, let $m_{ic_1c_2}$ represent the number of raters assigning subject i into category c_1 under condition A and into category c_2 under condition B. The resulting contingency table for the generic subject i is

B / A	1	2	...	K	
1	m_{i11}	m_{i12}	...	m_{i1K}	n_{i1A}
2	m_{i21}	m_{i22}	...	m_{i1K}	n_{i2A}
⋮	...	...	...	...	...
K	m_{iK1}	m_{iK2}	...	m_{iKK}	n_{iKA}
	n_{i1B}	n_{i2B}	...	n_{iKB}	n

where the marginal counts are defined as $n_{icA} = \sum_{c'} m_{icc'}$ and $n_{icB} = \sum_{c'} m_{ic'c}, c = 1, \ldots, K$. Accordingly, in this section, we switch to the natural notation $\hat{\kappa}_A$, $p_{oA} \ldots \hat{\kappa}_B$, $p_{oB} \ldots$ to indicate the two conditions respectively.

The counts $m_{ic_1c_2},\ c_1 = 1, \ldots, K,\ c_2 = 1, \ldots, K$ can be collected in a vector $m_{i\cdot\cdot}$ having, for a given i and in the lexicographic order, a multinomial distribution with parameters n and $\theta_{i\cdot\cdot}$ where, by definition, the c_1,c_2-th component $\theta_{ic_1c_2}$ of the vector $\theta_{i\cdot\cdot}$ is the probability for a generic rater to assign subject i to category c_1 under condition A and to category c_2 under condition B. Marginalizing, we obtain $\pi_{icA} = \sum_{c'=1}^{K} \theta_{icc'}$ and $\pi_{ic'B} = \sum_{c=1}^{K} \theta_{icc'},\ \ c, c' = 1, \ldots, K$.

The focus of the following main theorem is on the difference $\hat{\kappa}_A - \hat{\kappa}_B$, which turns out to be useful when building inferential procedures to compare the agreements in conditions A and B.

Theorem 2.2 *Assuming that, given subject i, the ratings are i.i.d. and that the ratings across different subjects (but not different conditions) are independent, the following asymptotic normal results hold:*

$$\sqrt{n}\left[\begin{pmatrix}\hat{p}_{oA,n}\\ \hat{p}_{eA,n}\\ \hat{p}_{oB,n}\\ \hat{p}_{eB,n}\end{pmatrix} - \begin{pmatrix}p_{oA}\\ p_{eA}\\ p_{oB}\\ p_{eB}\end{pmatrix}\right] \xrightarrow{\mathcal{L}} \mathcal{N}_4\left(0, \left(\begin{array}{cc:cc} \sigma_{oA,oA} & \sigma_{oA,eA} & \sigma_{oA,oB} & \sigma_{oA,eB} \\ \sigma_{eA,oA} & \sigma_{eA,eA} & \sigma_{eA,oB} & \sigma_{eA,eB} \\ \hdashline \sigma_{oB,oA} & \sigma_{oB,eA} & \sigma_{oB,oB} & \sigma_{oB,eB} \\ \sigma_{eB,oA} & \sigma_{eB,eA} & \sigma_{eB,oB} & \sigma_{eB,eB} \end{array}\right)\right) \quad \text{as } n \to \infty$$

where the north-western quadrant of the covariance matrix (the AA part), and the south-eastern quadrant (the BB part) are given in equation (8) *and the north-eastern corner (the AB part) is in turn given by*

$$\begin{aligned}
\sigma_{oA,oB} &= \frac{4}{N^2}\left(\sum_{i=1}^{N}\sum_{c=1}^{K}\sum_{c'=1}^{K}\theta_{icc'}(\pi_{icA} - \sum_{k=1}^{K}\pi_{ikA}^2)\pi_{ic'B}\right)\\
\sigma_{eA,oB} &= \frac{4}{N^2}\left(\sum_{i=1}^{N}\sum_{c=1}^{K}\sum_{c'=1}^{K}\theta_{icc'}(\bar{\pi}_{cA} - \sum_{k=1}^{K}\bar{\pi}_{kA}\pi_{ikA})\pi_{ic'B}\right)\\
\sigma_{oA,eB} &= \frac{4}{N^2}\left(\sum_{i=1}^{N}\sum_{c=1}^{K}\sum_{c'=1}^{K}\theta_{icc'}(\pi_{icA} - \sum_{k=1}^{K}\pi_{ikA}^2)\bar{\pi}_{c'B}\right)\\
\sigma_{eA,eB} &= \frac{4}{N^2}\left(\sum_{i=1}^{N}\sum_{c=1}^{K}\sum_{c'=1}^{K}\theta_{icc'}(\bar{\pi}_{cA} - \sum_{k=1}^{K}\bar{\pi}_{kA}\pi_{ikA})\bar{\pi}_{c'B}\right).
\end{aligned}$$

The main result is

$$\sqrt{n}\{(\hat{\kappa}_{A,n} - \hat{\kappa}_{B,n}) - (\kappa_A - \kappa_B)\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau_\Delta),\ \textit{as}\ \ n \to \infty. \tag{10}$$

with the asymptotic variance of the kappa difference given by

$$\tau_\Delta = \tau_A + \tau_B - 2\tau_{AB},$$

where τ_A and τ_B are derived from expression (9)*, while*

$$\tau_{AB} = \frac{\sigma_{oA,oB}}{(1-p_{eA})(1-p_{eB})} + \frac{\sigma_{eA,oB}(p_{oA}-1)}{(1-p_{eA})^2(1-p_{eB})} + \frac{\sigma_{oA,eB}(p_{oB}-1)}{(1-p_{eA})(1-p_{eB})^2} + \frac{\sigma_{eA,eB}(p_{oA}-1)(p_{oB}-1)}{(1-p_{eA})^2(1-p_{eB})^2}.$$

The proof of Theorem 2.2 is contained as an Appendix in the online Supplementary Material and it is a non-trivial application of the delta method.

The theorem allows for the construction of approximate asymptotic tests and confidence intervals for the difference of population kappa statistics. It is easy to show that the formula in Cao *et al.* (2016) is a special case of Theorem (2.2); when $K = 2$ the two-condition problem requires the specification of only 3 parameters per subject θ_{i11}, π_{i1A} and π_{i1B}, while the remaining ones may be expressed in terms of complementary probabilities (i.e. $\pi_{i2} = 1 - \pi_{i1}$, $\theta_{i12} = \pi_{i1A} - \theta_{i11}$, $\theta_{i21} = \pi_{i1B} - \theta_{i11}$ and $\theta_{i22} = 1 - \theta_{i11} - \theta_{i12} - \theta_{i21}$).

3 Applications

3.1 Monte Carlo simulations

Monte Carlo methods are used in this section to confirm the validity of the formulas obtained in Theorems 2.1 and 2.2 via simulations.

The central simulating scenario is obtained with $K = 3$ possible categories of rating, when half of the N subjects are characterized by multinomial probability parameters equal to $\pi_{i\cdot} = (0.09, 0.07, 0.84)$ and the other half by $\pi_{i\cdot} = (0.84, 0.07, 0.09)$, leading to $\kappa = 0.4999$, approximately 0.5. Two other scenarios are studied with lower and higher levels of agreement: the former, leading to $\kappa = 0.15$, is obtained with $\pi_{i\cdot} = (0.18, 0.20, 0.62)$ and $\pi_{i\cdot} = (0.62, 0.20, 0.18)$, the latter, leading to $\kappa = 0.85$, with $\pi_{i\cdot} = (0.02, 0.02, 0.96)$ and $\pi_{i\cdot} = (0.96, 0.02, 0.02)$. For each scenario, 10000 samples are generated for to provide $\hat{\kappa}^*_1, \ldots, \hat{\kappa}^*_{10000}$ for different values of n and N. The results are shown in Table 1 and in Table 2.

In Table 1, the theoretical asymptotic variance τ obtained from formula (9) is compared with the following empirical version:

$$\tau_{MC} = n \times \frac{\sum(\hat{\kappa}^*_j - \kappa)^2}{10000}$$

where κ is the true population kappa, obtained from the true probabilities used in the simulations. It can be seen that, for each N, the approximations get better and better for increasing values of n, validating Theorem 2.1.

Table 2 concerns the following simulated confidence intervals for kappa:

$$\hat{\kappa}^*_j \pm z_{\alpha/2} \sqrt{\hat{\tau}^*_j / n}, \;\; j = 1, \ldots, r \tag{11}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\hat{\tau}^*_j$ is obtained from equation (9), when all population parameters are estimated through their corresponding empirical versions, sample by sample. The empirical coverage over 10000 simulations of the above confidence intervals under the same scenarios as in Table 1 is reported in Table 2 and is shown to approximate well the nominal level 95%.

A framework similar to the one of Table 1 is exploited to produce Table 3 for the two-condition problem: we compare τ_Δ from Theorem 2.2 to its corresponding empirical $\tau_{\Delta,MC}$ in three scenarios designed to explore different values of the population kappa difference $\Delta\kappa = \kappa_A - \kappa_B$. The underlying population probabilities of the three scenarios, which can be read out the code in the online Supplementary Material, provide $\Delta\kappa \approx 0, 0.05, 0.10$ respectively. The same population parameters are used to derive Table 4, where we assess the performance of the test of the null hypothesis

$$H_0 : \Delta\kappa = 0 \tag{12}$$

based on the asymptotic distribution of the kappa difference statistic. As for Table 2, standard deviations are estimated from the data, to reproduce the situation researchers usually face. Determining the proportion of simulations where the null hypothesis is rejected, we are able to evaluate the type-I error rate (against

		$\kappa = 0.1512$		$\kappa = 0.4999$		$\kappa = 0.8506$	
N	n	τ_{MC}	RD (%)	τ_{MC}	RD (%)	τ_{MC}	RD (%)
4	10	0.1016	35.8	0.2185	11.6	0.1169	0.2
	50	0.0793	5.9	0.1964	0.3	0.1175	0.7
	100	0.0793	5.9	0.1965	0.3	0.1158	-0.8
	500	0.0761	1.7	0.1957	-0.1	0.1158	-0.8
	1000	0.0738	-1.5	0.1952	-0.3	0.1185	1.5
		$\tau = 0.0749$		$\tau = 0.1958$		$\tau = 0.1167$	
10	10	0.04	33.6	0.0831	6.1	0.0468	0.2
	50	0.0312	4.2	0.0785	0.2	0.0467	0
	100	0.0312	4.2	0.0807	3.1	0.0469	0.4
	500	0.0312	4.3	0.0785	0.2	0.0462	-1
	1000	0.0303	1.1	0.0774	-1.2	0.0466	-0.3
		$\tau = 0.0299$		$\tau = 0.0783$		$\tau = 0.0467$	
100	10	0.0039	28.6	0.0083	6.2	0.0047	0.4
	50	0.0031	4.2	0.008	1.6	0.0046	-1
	100	0.003	1.7	0.0081	3	0.0047	1.3
	500	0.003	0.4	0.0078	-0.4	0.0047	-0.1
	1000	0.003	1.3	0.0077	-2.1	0.0048	2.3
		$\tau = 0.003$		$\tau = 0.0078$		$\tau = 0.0047$	

Table 1 For different values of κ (corresponding to three scenarios described in the text) the table contains the simulated variance τ_{MC}, its relative difference against τ (i.e. RD=$100(\tau_{MC} - \tau)/\tau$) and the asymptotic variances τ from Theorem 2.1, for various combinations of the number of subjects N and the number of raters n.

the 5% nominal value) and the power when $\Delta\kappa = 0$ and $\Delta\kappa \approx 0.05, 0.10$ respectively. One can see that the type-I error rate is approximated better and better as n increases, whereas for $\Delta\kappa \geq 0.10$, moderate N and $n \geq 100$, the power of the test reaches already its maximum value 1, making higher values of $\Delta\kappa$ not necessary to simulate.

3.2 Summary of results for the IGIBDendo case study

One of the training activities of the IGIBDendo group in 2013 was a series of educational events in several cities in Italy aimed at improving the use of the Mayo score to rate endoscopic videos. The events took place in similar ways: first (situation B) the participants rated $N = 5$ endoscopic videos on the Mayo score with $K = 4$ categories, then some training took place; finally the same participants rated again the same videos (situation A) on the same scale. A total of $n = 121$ participants provided full data (missing data were ignored under a missing-completely-at-random assumption). The goal was to assess whether the training intervention significantly increased interrater agreement. In order to do so, approximate confidence intervals were computed in the usual way by inverting the asymptotically normal pivotal quantities in Theorem 2.1 and 2.2 and by plugging empirical estimates of $\pi_{icA}, \pi_{icB}, \theta_{icc'}, c, c' = 1, \ldots, K$ in the formulas obtained for the different asymptotic variances. These are the sort of intervals illustrated in Tables 2 and 4. The interest in the single kappa statistics lied in a rough comparison with the historical level of agreement usually found in this area of research, whereas the main focus was on the confidence interval for $\kappa_A - \kappa_B$, which can also be used as a test of the hypothesis $\kappa_A = \kappa_B$. Such hypothesis can be rejected at level α in favour of the hypothesis $\kappa_A > \kappa_B$ if the corresponding $100(1 - \alpha)\%$ confidence interval for $\kappa_A - \kappa_B$ is entirely positive, which actually was the case. Approximate confidence intervals for the difference $\kappa_A - \kappa_B$ were reported.

N	n	$\kappa = 0.1512$ Coverage (%)	$\kappa = 0.4999$ Coverage (%)	$\kappa = 0.8506$ Coverage (%)
4	10	86.1	83.7	78.5
	50	92.5	93.3	90.2
	100	93.5	94.3	93.8
	500	95.2	94.8	94.5
	1000	95.3	95.2	94.9
10	10	89	85.8	85.3
	50	93.6	93.7	92.2
	100	94.5	94.7	94.2
	500	94.7	94.6	94.3
	1000	94.9	95.2	94.7
100	10	90.2	86.7	86.2
	50	94.5	93.4	93.7
	100	94.8	94.6	94.2
	500	95.4	94.8	94.9
	1000	95.1	94.9	94.5

Table 2 Number of subjects N, number of raters n and empirical coverage of the confidence intervals described in Equation (11) for different values of κ (corresponding to three scenarios described in the text).

N	n	$\Delta\kappa = 0$ $\tau_{\Delta,MC}$	RD (%)	$\Delta\kappa = 0.0509$ $\tau_{\Delta,MC}$	RD (%)	$\Delta\kappa = 0.1083$ $\tau_{\Delta,MC}$	RD (%)
4	10	0.1304	10.1	0.1293	15	0.1178	12.6
	50	0.1253	5.8	0.1178	4.8	0.1055	0.8
	100	0.1176	-0.7	0.1167	3.8	0.107	2.3
	500	0.1176	-0.7	0.1122	-0.2	0.1055	0.9
	1000	0.1196	1	0.1118	-0.6	0.1053	0.7
		$\tau = 0.1185$		$\tau = 0.1124$		$\tau = 0.1046$	
10	10	0.0521	10	0.0505	12.4	0.0468	12
	50	0.0474	0.1	0.0458	1.9	0.0434	3.7
	100	0.0487	2.7	0.0467	3.9	0.0412	-1.5
	500	0.0485	2.4	0.0452	0.4	0.0415	-0.9
	1000	0.0469	-1	0.0451	0.3	0.0419	0.2
		$\tau = 0.0474$		$\tau = 0.0450$		$\tau = 0.0418$	
100	10	0.0051	8	0.005	10.3	0.0046	10
	50	0.0049	3.5	0.0046	1.5	0.0044	4.5
	100	0.0048	0.4	0.0046	1.9	0.0042	-0.4
	500	0.0048	1.1	0.0045	-0.9	0.0041	-2.9
	1000	0.0047	-1.5	0.0046	1.3	0.0042	-0.2
		$\tau = 0.0047$		$\tau = 0.0045$		$\tau = 0.0042$	

Table 3 For three different values of $\Delta\kappa \approx 0, 0.05, 0.10$, corresponding to three different true scenarios, the table contains the simulated variance $\tau_{\Delta,MC}$, its relative difference against τ_Δ (i.e. RD=$100(\tau_{\Delta,MC} - \tau_\Delta)/\tau_\Delta$) and the corresponding asymptotic variances τ_Δ from Theorem 2.2, for various combinations of the number of subjects N and the number of raters n.

N	n	$\Delta\kappa = 0$ Type-I error (%)	$\Delta\kappa = 0.0509$ Power (%)	$\Delta\kappa = 0.1083$ Power (%)
4	10	NA	NA	NA
4	50	5.6	20.8	66.3
4	100	5.4	33.2	90.8
4	500	5.1	91.3	100
4	1000	5.4	99.8	100
10	10	13.2	22.7	51.3
10	50	6	41.0	95.9
10	100	6	66.2	100
10	500	4.9	100	100
10	1000	5	100	100
100	10	12.3	77.0	100
100	50	6	100	100
100	100	5.4	100	100
100	500	5.1	100	100
100	1000	4.8	100	100

Table 4 Number of raters n, number of subjects N, type-I error rate (when $\Delta\kappa = 0$, with 5% nominal value) and power for two different values of $\Delta\kappa \approx 0.05, 0.10$, corresponding to three different true scenarios.

	lower	mean	upper
κ_B	0.41	0.46	0.50
κ_A	0.70	0.74	0.77
$\kappa_A - \kappa_B$	0.22	0.28	0.33

Table 5 Confidence intervals for kappa statistics and their difference in the IGIBDendo project

The results are shown in Table 5. Further details on the IGIBDendo project (metanalysis of more than one training event, clustering of participants) can be found in Daperno *et al.* (2016).

4 Conclusion

On the Internet and in several manuscripts, the following quote is attributed to Fleiss and Cuzick (1979):

> Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third.

Actually, the attribution seems to be a fake, but the quote reflects the many trials and many errors in the history of this subject. Many years later, the situation has not changed a lot for the computation of the variance of the kappa statistic for the many-rater situation. Building on a method of proof due to Cao *et al.* (2016), we have tried to give a sensible contribution to the third human endeavor mentioned above.

Conflict of Interest

The authors have declared no conflict of interest.

Acknowledgements The paper has improved a lot with the reviews of the Associate Editor and two referees, whom we thank.

References

Cao, H., Sen, P., Peery, A. and Dellon, E. (2016). Assessing agreement with multiple raters on correlated kappa statistics. *Biometrical Journal*, 58, 4, 935-943.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, 37–46.

Daperno, M., Comberlato, M., Bossa, F., Armuzzi, A., Biancone, L., Bonanomi, A.G., Cosintino, R., Lombardi, G., Mangiarotti, R., Papa, A., Pica, R., Grassano, L., Pagana, G., D'Incà, R., Orlando, A. and Rizzello, F. (2016). Training programs on endoscopic scoring systems for inflammatory bowel disease lead to significant increase in inter-observer agreement among community Gastroenterologists. *Journal of Crohn's and Colitis*, `http://dx.doi.org/10.1093/ecco-jcc/jjw181`

Congalton, R.G. and Green, K. (2008). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press.

Davies, M. and Fleiss, J.L. (1982). Measuring Agreement for Multinomial Data. *Biometrics*, 38, 1047–1051.

De Mast, J. (2007). Agreement and kappa-type indices. *The American Statistician*, 61, 2, 148–153.

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

Fleiss, J.L., Cohen, J. and Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323–327.

Fleiss, J.L. and Cuzick, J. (1979). The reliability of dichotomous judgments: unequal number of judges per subject. *Applied Psychological Measurement*, 3:4, 537–542.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical methods for rates and proportions*. John Wiley & Sons.

Gwet, K.L. (2008). Variance Estimation of Nominal-Scale Inter-Rater Reliability with Random Selection of Raters. *Psychometrika*, 73, 407–430.

Appendix *(proofs of theorems)*

A.1. Proof of Theorem 2.1

The multivariate central limit theorem for the multinomial frequencies $f_{i\cdot} = n_{i\cdot}/n$ guarantees that for any $i = 1 \dots N$

$$\sqrt{n}\,(f_{i\cdot} - \pi_{i\cdot}) \xrightarrow{\mathcal{L}} \mathcal{N}_K(0, \Sigma_i) \qquad \text{for } n \to \infty \tag{1}$$

where Σ_i is a $K \times K$ matrix with diagonal entries $\Sigma_{cc} = \pi_{ic}(1 - \pi_{ic})$ and off diagonal entries $\Sigma_{cc'} = -\pi_{ic}\pi_{ic'}$ ($c \neq c'$). The $f_{i\cdot}$ and $f_{j\cdot}$ are independent if $i \neq j$. Equation (1) holds for any i, so we can reorganize the frequencies into a single vector and obtain

$$\sqrt{n}\left[\begin{pmatrix} f_{11} \\ \vdots \\ f_{1K} \\ \vdots \\ f_{N1} \\ \vdots \\ f_{NK} \end{pmatrix} - \begin{pmatrix} \pi_{11} \\ \vdots \\ \pi_{1K} \\ \vdots \\ \pi_{N1} \\ \vdots \\ \pi_{NK} \end{pmatrix}\right] \xrightarrow{\mathcal{L}} \mathcal{N}_{N\times K}\,(0, \Sigma) \qquad \text{for } n \to \infty$$

where Σ is a $(N \times K) \times (N \times K)$ block-diagonal matrix with the $K \times K$ matrices Σ_i as diagonal blocks. The estimator $\hat{p}_{o,n}$ can be recast into the following form

$$\hat{p}_{o,n} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} f_{ic}^2 + O_P\left(\frac{1}{n}\right)$$

where the term $O_P\left(\frac{1}{n}\right)$ does not contribute to the asymptotic distribution since it converge to zero and can then be disregarded. The vector $\left(\begin{smallmatrix}\hat{p}_{o,n}\\ \hat{p}_{o,n}\end{smallmatrix}\right)$ is a differentiable functions of the multinomial frequencies $f_{i\cdot}$ with Jacobian matrix composed of

$$\frac{\partial \hat{p}_{o,n}}{\partial f_{ic}} \approx \frac{2}{N} f_{ic} \tag{2}$$

$$\frac{\partial \hat{p}_{e,n}}{\partial f_{ic}} = \frac{2}{N} \cdot \frac{1}{N}\sum_{i=1}^{N} f_{ic} \tag{3}$$

By the delta method therefore, the following asymptotic normality holds

$$\sqrt{n}\left[\begin{pmatrix}\hat{p}_{o,n}\\ \hat{p}_{e,n}\end{pmatrix} - \begin{pmatrix}p_o\\ p_e\end{pmatrix}\right] \xrightarrow{\mathcal{L}} \mathcal{N}_2\Big(0, \Gamma\Big)$$

with the matrix Γ given by the following matrix product

$$\frac{4}{N^2}\begin{pmatrix} \pi_{11} & \cdots & \pi_{1K} & \cdots & \pi_{N1} & \cdots & \pi_{NK} \\ \bar{\pi}_1 & \cdots & \bar{\pi}_K & \cdots & \bar{\pi}_1 & \cdots & \bar{\pi}_K \end{pmatrix} \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \Sigma_N \end{pmatrix} \begin{pmatrix} \pi_{11} & \bar{\pi}_1 \\ \vdots & \vdots \\ \pi_{1K} & \bar{\pi}_K \\ \vdots & \vdots \\ \pi_{N1} & \bar{\pi}_1 \\ \vdots & \vdots \\ \pi_{NK} & \bar{\pi}_K \end{pmatrix}$$

which after some algebra simplifies to the expression stated in Theorem 2.1, formula (8) in the main text. Finally, κ_n is a differentiable function of $\begin{pmatrix} \hat{p}_{o,n} \\ \hat{p}_{o,n} \end{pmatrix}$ and one further application of the delta method give us that

$$\sqrt{n}\,(\hat{\kappa}_n - \kappa) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tau)$$

with τ given by the following matrix product

$$\tau = \left(\tfrac{1}{1-p_e} - \tfrac{1-p_o}{(1-p_e)^2} \right) \begin{pmatrix} \sigma_{oo} & \sigma_{oe} \\ \sigma_{oe} & \sigma_{ee} \end{pmatrix} \begin{pmatrix} \frac{1}{1-p_e} \\ -\frac{1-p_o}{(1-p_e)^2} \end{pmatrix}$$

which again is easily manipulated into the expression given in the main text.

A.2. Proof of Theorem 2.2

The proof of Theorem 2.2 is a simple, but computationally demanding, generalization of the proof of Theorem 2.1.

First, the multivariate central limit theorem is applied to the multinomial frequencies $q_{i\cdot\cdot} = m_{i\cdot\cdot}/n$:

$$\sqrt{n}(q_{i\cdot\cdot} - \theta_{i\cdot\cdot}) \xrightarrow{\mathcal{L}} \mathcal{N}_2(0, \Lambda_i) \qquad \text{for } n \to \infty \tag{4}$$

Diagonal elements of Λ_i are given by $\theta_{ic_1c_2}(1 - \theta_{ic_1c_2})$ and the off-diagonal ones by $-\theta_{ic_1c_2}\theta_{ic_1'c_2'}$.

Since (4) holds for any i we could also reformulate the theorem by organizing the frequencies into a single vector:

$$\sqrt{n}\left[\begin{pmatrix} q_{111} \\ \vdots \\ q_{11K} \\ \vdots \\ q_{1K1} \\ \vdots \\ q_{1KK} \\ \vdots \\ q_{N11} \\ \vdots \\ q_{N1K} \\ \vdots \\ q_{NK1} \\ \vdots \\ q_{NKK} \end{pmatrix} - \begin{pmatrix} \theta_{111} \\ \vdots \\ \theta_{11K} \\ \vdots \\ \theta_{1K1} \\ \vdots \\ \theta_{1KK} \\ \vdots \\ \theta_{N11} \\ \vdots \\ \theta_{N1K} \\ \vdots \\ \theta_{NK1} \\ \vdots \\ \theta_{NKK} \end{pmatrix} \right] \xrightarrow{\mathcal{L}} \mathcal{N}_{N\times K\times K}\,(0, \Lambda) \qquad \text{for } n \to \infty$$

where Λ is a $(N \times K \times K) \times (N \times K \times K)$ block-diagonal matrix with the N matrices Λ_i as diagonal blocks of size $K^2 \times K^2$.

The delta method leads to the proof of the asymptotic normality

$$\sqrt{n}\left[\begin{pmatrix} \hat{p}_{oA,n} \\ \hat{p}_{eA,n} \\ \hat{p}_{oB,n} \\ \hat{p}eB,n \end{pmatrix} - \begin{pmatrix} p_{oA} \\ p_{eA} \\ p_{oB} \\ p_{eB} \end{pmatrix} \right] \xrightarrow{\mathcal{L}} \mathcal{N}_4 \left(0, \left(\begin{array}{cc:cc} \sigma_{oA,oA} & \sigma_{oA,eA} & \sigma_{oA,oB} & \sigma_{oA,eB} \\ \sigma_{eA,oA} & \sigma_{eA,eA} & \sigma_{eA,oB} & \sigma_{eA,eB} \\ \hdashline \sigma_{oB,oA} & \sigma_{oB,eA} & \sigma_{oB,oB} & \sigma_{oB,eB} \\ \sigma_{eB,oA} & \sigma_{eB,eA} & \sigma_{eB,oB} & \sigma_{eB,eB} \end{array} \right) \right) \qquad \text{for } n \to \infty \tag{5}$$

