Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-024-00581-x

REGULAR ARTICLE

Estimators of various *kappa* coefficients based on the unbiased estimator of the expected index of agreements

A. Martín Andrés[1] · **M. Álvarez Hernández**[2,3]

Received: 14 May 2023 / Accepted: 28 January 2024

Abstract

To measure the degree of agreement between R observers who independently classify n subjects within K categories, various *kappa*-type coefficients are often used. When $R = 2$, it is common to use the Cohen' *kappa*, Scott's *pi*, Gwet's *AC1/2*, and Krippendorf's *alpha* coefficients (weighted or not). When $R > 2$, some pairwise version based on the aforementioned coefficients is normally used; with the same order as above: Hubert's *kappa*, Fleiss's *kappa*, Gwet's *AC1/2,* and Krippendorf's *alpha*. However, all these statistics are based on biased estimators of the expected index of agreements, since they estimate the product of two population proportions through the product of their sample estimators. The aims of this article are three. First, to provide statistics based on unbiased estimators of the expected index of agreements and determine their variance based on the variance of the original statistic. Second, to make pairwise extensions of some measures. And third, to show that the old and new estimators of the Cohen's *kappa* and Hubert's *kappa* coefficients match the well-known estimators of concordance and intraclass correlation coefficients, if the former are defined by assuming quadratic weights. The article shows that the new estimators are always greater than or equal the classic ones, except for the case of Gwet where it is the other way around, although these differences are only relevant with small sample sizes (e.g. $n \leq 30$).

Keywords Agreement · Cohen's *kappa* · Concordance and intraclass correlation coefficients · Conger's *kappa* · Fleiss' *kappa* · Gwet's *AC1/2* · Hubert's *kappa* · Krippendorf's *alpha* · Pairwise multi-rater *kappa* · Scott's *pi*

✉ A. Martín Andrés
amartina@ugr.es

[1] Biostatistics, Faculty of Medicine, C8-01, University of Granada, 18071 Granada, Spain

[2] CITMAga, 15782 Santiago de Compostela, Spain

[3] Defense University Center, Spanish Naval Academy, Marín, Pontevedra, Spain

Published online: 06 March 2024

Mathematics Subject Classification 62F10 · 62F12 · 62H99 · 62P15

1 Introduction

It is often necessary to assess the degree of concordance or agreement between R raters which independently classify n subjects within $K \geq 2$ categories (Fleiss 1971; Landis and Koch 1975a, b; Warrens 2010; Schuster and Smith 2005).

Let this be the case for only two raters ($R = 2$) and nominal categories. As some of the observed agreements may be due to chance, it is most common to eliminate the effect of chance by defining a *kappa*-type coefficient of the form $\kappa = (I_o - I_e)/(1 - I_e)$. In that expression I_o is the observed index of agreements (the sum of the observed proportions of agreements),I_e is the expected index of agreements (the sum of the proportions of agreements that would happen if the two raters acted independently) and κ is the population value of the proposed agreement measure. Note that the previous indexes only consider the agreements obtained. When the categories are ordinal, the indexes defined are similar to the previous ones, but also considering the disagreements obtained, to which certain weights are assigned (see Sect. 2.1); this leads to a weighted *kappa* coefficient. From now on, κ will allude to one or the other indistinctly. According to the definition adopted for I_e, the different *kappa* coefficients are obtained: κ_S (Scott 1955), κ_C (Cohen 1960, 1968), and κ_G (Gwet 2008). The estimation of these coefficients has the general form of $\hat{\kappa} = \left(\hat{I}_o - \hat{I}_e\right) \Big/ \left(1 - \hat{I}_e\right)$, where the values $\hat{\kappa}$, $\hat{I}_o$ and $\hat{I}_e$ are the sample estimators of the previous population parameters. It can be seen that κ and $\hat{\kappa}$ are decreasing functions of I_e and $\hat{I}_e$ respectively. Additionally, Krippendorf (1970, 2004) provides an estimator $\hat{\kappa}_K$ of κ_S that differs slightly from the more classical $\hat{\kappa}_S$ because of its new definition of $\hat{I}_o$.

Let this be the case for multi-raters ($R \geq 2$). The different coefficients κ of the case $R = 2$ can be generalized for the case of multi-raters in several ways, depending on how the phrase "an agreement has occurred" is interpreted. The most common interpretation is that of Fleiss (1971) and Hubert (1977) "an agreement occurs if and only if two raters categorize an object consistently" or *pairwise* definition of agreement. This is the definition in this article. Hubert (1977) also makes the following interpretation "an agreement occurs if and only if all raters agree on the categorization of an object", or *R-wise* definition (Conger 1980). The extension *R-wise* κ_{HR} of κ_C can be seen in Conger (1980), Shuster and Smith (2005) and Martín Andrés and Álvarez Hernández (2020). The best-known *pairwise* extensions of the coefficients κ_S, κ_C and κ_G are the coefficients κ_F (Fleiss 1971), κ_H (Hubert 1977; Conger 1980) and κ_G (Gwet 2008) respectively. All of them are defined under the same format as in the case of $R = 2$. Additionally, Krippendorf (1970, 2004) provides an estimator $\hat{\kappa}_K$ of κ_F that differs slightly from the more classical $\hat{\kappa}_F$, again because of the definition of $\hat{I}_o$. An overview of all of the above can be seen in Gwet's book (2021).

However, all $\hat{\kappa}_X$ expressions are based on biased estimators (X refers to any of the letters used above), since they estimate the product of two population proportions -a term that is present in I_e- through the product of their sample estimators. The first objective of this article is to correct this bias by proposing unbiased estimators $\hat{I}_{eU}$ of

I_e − so the new estimator of κ_X will be $\hat{\kappa}_{XU} = \left(\hat{I}_o - \hat{I}_{eU}\right) \Big/ \left(1 - \hat{I}_{eU}\right)$ − , as well as to determine the variance of $\hat{\kappa}_{XU}$. This methodology is easy to apply to any other *kappa* coefficient studied. A second objective is to make *pairwise* extensions of some measures, but in a different way to traditional *pairwise* extensions.

The previous description is very general since it is necessary to specify who are the "subject population" and the "rater population". Regarding the population of subjects, the n subjects may be: (a) a random sample of an infinite population of subjects, which is what is assumed in the rest of the sections; (b) a random sample of a finite population of subjects, in which case a finite population correction (Gwet 2021a, b) must be made to the formulas of the variance; and (c) the only subjects of interest, in which case only $\hat{\kappa}_X$ makes sense, there is no κ_X parameter to estimate and it makes no sense to define $\hat{\kappa}_{XU}$.

Regarding the population of raters, the R raters may be (Shrout and Fleiss 1979): (1) different for the same subject -even with a different number- and extracted from an infinite population of raters; (2) the same for all of the subjects and extracted from an infinite population of raters; and (3) the same for all of the subjects and they are the only raters of interest, which is what is assumed in the rest of the sections. When the replies of the raters are quantitative, a traditional way of measuring the degree of agreement between them is through the intraclass correlation coefficients (ICC) ρ_{I1}, ρ_{I2}, and ρ_{I3} which are obtained from the corresponding one-way random model, two-way random model, or two-way mixed model, respectively. In the last two cases it is assumed that there is no interaction. Nevertheless, in this context of measures of agreement, Shrout and Fleiss (1979) and Carrasco and Jover (2003) point out that in case (3) it is also necessary to include the variability between raters in the total variability, so that in cases (2) and (3) we should use ρ_{I2}. Additionally, and for case (3), Lin (1989, 2000) and Barnhart et al. (2002) propose using as a measure of agreement the concordance correlation coefficient (CCC) ρ_L.

As is logical, different researchers have shown interest in searching for relations between the coefficients κ_X, ρ_{Ii}, and ρ_L, as well as between their estimators $\hat{\kappa}_X$, $\hat{\rho}_{Ii}$, and $\hat{\rho}_L$. Landis and Koch (1977) demonstrated that $\hat{\kappa}_F$ is asymptotically equivalent to ρ_{I1} when the replies are binary. Furthermore, Barnhart et al. (2002) and Carrasco and Jover (2003) demonstrated that $\rho_L = \rho_{I2}$. Since in the case of $R = 2$ Martín Andrés and Álvarez Hernández (2020) demonstrated that $\rho_L = \kappa_C$—assuming, as from now on, that the weights of the disagreements are quadratic—, then the satisfactory property $\kappa_C = \rho_L = \rho_{I2}$ is obtained when $R = 2$. The equivalences between the estimators of these parameters are more complex, since their values depend on the method of estimating their components. For example, Fleiss and Cohen (1973) demonstrated that $\hat{\kappa}_C$ is asymptotically equivalent to ρ_{I2}, King and Chinchilli (2001) and Martín Andrés and Álvarez Hernández (2020) demonstrated that $\hat{\kappa}_C = \hat{\rho}_L$ when direct (biased) estimators are used, and Davis and Fleiss (1982) verified that $\hat{\kappa}_H$ is asymptotically equivalent to ρ_{I2} when the replies are binary. The third objective of this article is to relate κ_H to ρ_L, as well as estimators $\hat{\kappa}_{CU}$ and $\hat{\kappa}_{HU}$ with estimators $\hat{\rho}_{I2}$ and $\hat{\rho}_{LU}$, which is based on the unbiased estimators of the components of ρ_{I2} and ρ_L, respectively.

From the aforementioned reasons, we can see that in this article it is assumed that n subjects, extracted randomly from an infinite population, are given a score a single time by R fixed raters (who are the only ones of interest). It is also assumed that there are no missing data, i.e. that all of the raters give a reply in all of the subjects.

2 Case of two raters

Let be two raters ($R = 2$) that independently classify n subjects within K categories. Let O_{ij} be the number of subjects whom observer 1 classifies as type i ($i = 1, 2, \ldots, K$) and observer 2 as type j ($j = 1, 2, \ldots, K$). This gives rise to a table of absolute frequencies O_{ij} like those in Tables 1 and 2, with observed proportions $\hat{p}_{ij} = O_{ij}/n$, where $\Sigma_i \Sigma_j O_{ij} = n$ and $\Sigma_i \Sigma_j \hat{p}_{ij} = 1$. The notation for row totals ($O_{i\cdot}$ and $\hat{p}_{i\cdot}$), of column ($O_{\cdot j}$ and $\hat{p}_{\cdot j}$) or general ($O_{\cdot\cdot} = n$ and $\hat{p}_{\cdot\cdot} = 1$) is the usual; for example $\hat{p}_{i\cdot} = \Sigma_j \hat{p}_{ij}$. If the subjects have been chosen randomly and both raters classify all of the subjects, then the observed dataset $\{O_{ij}\}$ comes from a multinomial distribution of parameters n and $\{p_{ij}\}$, where p_{ij} is the probability that a subject will be classified in cell (i, j). Additionally $\{p_{i\cdot}\}$ and $\{p_{\cdot j}\}$ will be the marginal distributions of the row and column observers respectively. Obviously, $\hat{p}_{ij}$, $\hat{p}_{i\cdot}$ and $\hat{p}_{\cdot j}$ are the maximum likelihood estimators of p_{ij}, $p_{i\cdot}$ and $p_{\cdot j}$ respectively. At the end of "Appendix 2", another type of sampling is mentioned in detail.

2.1 Weighted and unweighted kappa and observed index of agreements

It has already been indicated that κ depends on the indexes of agreement I_o (observed) and I_e (expected). To evaluate any of them it is necessary to previously define the weight or degree of agreement w_{ij} that is assigned to the answer (i, j), with $0 \leq w_{ij} \leq 1$, w_{ii}

Table 1 Diagnosis of $n = 100$ subjects by $R = 2$ raters in $K = 3$ categories (Fleiss et al. 2003)

Rater 1	Rater 2			*Totals* ($O_{i\cdot}$)
	Psychotic	Neurotic	Organic	
(a) Observed frequencies (O_{ij})				
Psychotic	75	1	4	80
Neurotic	5	4	1	10
Organic	0	0	10	10
Totals ($O_{\cdot j}$)	80	5	15	100 ($O_{\cdot\cdot}$)

Coefficients	Classic	New
(b) Estimated *kappa* coefficients		
Cohen's *kappa*	$\hat{\kappa}_C = 0.676$	$\hat{\kappa}_{CU} = 0.679$
Scott's *pi*	$\hat{\kappa}_S = 0.675$	$\hat{\kappa}_{SU} = 0.678$
Krippendorf's *alpha*	$\hat{\kappa}_K = 0.677$	$\hat{\kappa}_{KU} = 0.680$
Gwet's *AC1*	$\hat{\kappa}_G = 0.868$	$\hat{\kappa}_{GU} = 0.867$

Table 2 Classification of $n = 8$ subjects by $R = 2$ raters in $K = 3$ categories (Gwet 2021b, p 109)

Rater 1	Rater 2			Totals ($O_{i.}$)
	A	*B*	*C*	
(a) Observed frequencies (O_{ij})				
A	1	1	0	2
B	0	3	1	4
C	0	0	2	2
Totals ($O_{.j}$)	1	4	3	8 ($O_{..}$)

Coefficients	Classic	New
(b) Estimated *kappa* coefficients		
Cohen's *kappa*	$\hat{\kappa}_C = 0.600$	$\hat{\kappa}_{CU} = 0.632$
Scott's *pi*	$\hat{\kappa}_S = 0.595$	$\hat{\kappa}_{SU} = 0.636$
Krippendorf's *alpha*	$\hat{\kappa}_K = 0.620$	$\hat{\kappa}_{KU} = 0.659$
Gwet's *AC1*	$\hat{\kappa}_G = 0.638$	$\hat{\kappa}_{GU} = 0.619$

$= 1$, and generally $w_{ij} = w_{ji} < 1$ $(i \neq j)$. When categories are ordinal, there are many ways to assign values to w_{ij} (Schuster and Smith 2005). If we assume that categories 1, 2, …, K are ordered from the lowest to highest, it is usual that w_{ij} is related to the value of $(i - j)$. A classic definition, to which we will refer later, is the quadratic weighting $w_{ij} = 1 - [(i - j)/(K - 1)]^2$ of Fleiss and Cohen (1973). When categories are nominal, it is traditional to assign the weights $w_{ii} = 1$ and $w_{ij} = 0$ $(i \neq j)$, that is, it only considers the actual agreements. Historically, the different coefficients κ are defined first in the unweighted case, later extending it to the weighted case. However this article will be developed for the general weighted case, since the unweighted is a particular case of that: $w_{ij} = \delta_{ij}$ with δ_{ij} are the Kronecker deltas.

All coefficients κ are defined based on the same value of the index of agreements observed. Therefore, it is appropriate to indicate their definition (I_o) and their estimate ($\hat{I}_o$) as general reference for all this Sect. 2:

$$I_o = \sum_i \sum_j w_{ij} p_{ij} \quad \text{and} \quad \hat{I}_o = \sum_i \sum_j w_{ij} \hat{p}_{ij} = \sum_i \sum_j w_{ij} O_{ij} \Big/ n,$$

where $\hat{I}_o$ is an unbiased estimator of I_o.

2.2 *Cohen's kappa and the intraclass and concordance correlation coefficients*

Cohen (1960, 1968) defines the classical measure of agreement

$$k_C = (I_o - I_e)/(1 - I_e) \quad where \quad I_e = \sum_i \sum_j w_{ij} p_{i\cdot} p_{\cdot j}, \tag{1}$$

and proposes to estimate it by,

$$\hat{\kappa}_C = \left(\hat{I}_o - \hat{I}_e\right) \Big/ \left(1 - \hat{I}_e\right) \quad \text{where}$$
$$\hat{I}_e = \sum_i \sum_j w_{ij} \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \sum_i \sum_j w_{ij} O_{i\cdot} O_{\cdot j} \Big/ n^2.$$

As indicated in "Appendix 1", $\hat{p}_{i\cdot}\hat{p}_{\cdot j}$ is not an unbiased estimator of $p_{i\cdot}p_{\cdot j}$ since

$$E\left(\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right) = \frac{(n-1)p_{i\cdot}p_{\cdot j} + p_{ij}}{n}, \tag{2}$$

although it is asymptotically unbiased, as happens in the other cases that follow. Therefore $E\left(\hat{I}_e\right) = \sum_i \sum_j E\left(\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right) = \{(n-1)I_e + I_o\}/n$ and $\hat{I}_e$ is also not an unbiased estimator of I_e. From expression (2) it follows that the unbiased estimators of $p_{i\cdot}p_{\cdot j}$ and I_e are

$$\widehat{p_{i\cdot}p_{\cdot j}} = \frac{n\hat{p}_{i\cdot}\hat{p}_{\cdot j} - \hat{p}_{ij}}{n-1} \quad \text{and} \quad \hat{I}_{eU} = \sum_i w_{ij} \widehat{p_{i\cdot}p_{\cdot j}} = \frac{n\hat{I}_e - \hat{I}_o}{n-1}, \tag{3}$$

respectively. Thus, the new estimator $\hat{\kappa}_{CU}$ of κ_C will be

$$\hat{\kappa}_{CU} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{n\hat{\kappa}_C}{(n-1) + \hat{\kappa}_C}, \tag{4}$$

and its variance, which is deduced in "Appendix 2", is

$$V\left(\hat{\kappa}_{CU}\right) = \frac{(n - \kappa_C)^4}{\{n(n-1)\}^2} V\left(\hat{\kappa}_C\right), \tag{5}$$

where $V\left(\hat{\kappa}_C\right)$ refers to the formula of Fleiss et al. (1969), which can be seen in the book by Gwet (2021b); this book also contains all of the variances that are needed in what follows. This type of correction is similar to the one used by Miettinen and Nurminen (1985) for the score statistics in 2 × 2 tables. Because of expression (3), $\hat{I}_{eU} - \hat{I}_e$ is proportional to $-\left(\hat{I}_o - \hat{I}_e\right) \leq 0$ if and only if $\hat{\kappa}_C \geq 0$. As $\hat{\kappa}_C$ decreases with $\hat{I}_e$, then $\hat{\kappa}_{CU} \geq \hat{\kappa}_C$ in the case of positive agreement ($\hat{\kappa}_C \geq 0$, which is the case of greatest interest). It is easy to see that $V\left(\hat{\kappa}_{CU}\right) \leq V\left(\hat{\kappa}_{C}\right)$ if and only if $\kappa_C \geq n^{0.5}/\{n^{0.5} + (n-1)^{0.5}\}$. Something similar happens with the other variances obtained below.

Let there now be two raters with quantitative answers x_1 and x_2 with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}. Lin (1989, 2000) established the following measure of quantitative agreement ρ_L (known as CCC) and its estimation $\hat{\rho}_L$

$$\rho_L = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad \text{and} \quad \hat{\rho}_L = \frac{2S_{12}}{S_1^2 + S_2^2 + (\overline{x}_1 - \overline{x}_2)^2}, \tag{6}$$

where S_i^2 and S_{12} are the biased estimators of the variances and covariances respectively (both with denominator n) and $\overline{x}_i$ are the sample means. As mentioned in the Introduction, the quadratic weighting has the advantage of achieving that $\kappa_C = \rho_L = \rho_{I2}$ and that $\hat{\rho}_L = \hat{\kappa}_C$. On the other hand, Carrasco and Jover (2003) replaced the values of σ_i^2, σ_{12} and $(\mu_1 - \mu_2)^2$ for their unbiased estimators s_i^2, s_{12} (their sample variances and covariances with denominator $n-1$) and $\widehat{(\mu_1 - \mu_2)^2} = (\overline{x}_1 - \overline{x}_2)^2 - (s_1^2 + s_2^2 - 2s_{12})/n$ in the first expression (6), which led to the following estimator $\hat{\rho}_{LU}$ of ρ_L,

$$\begin{aligned}\hat{\rho}_{LU} &= \frac{2ns_{12}}{\left(s_1^2 + s_2^2\right)(n-1) + n\left(\overline{x}_1 - \overline{x}_2\right)^2 + 2s_{12}} \\ &= \frac{2nS_{12}}{(n-1)\left\{S_1^2 + S_2^2 + (\overline{x}_1 - \overline{x}_2)^2\right\} + 2S_{12}},\end{aligned} \tag{7}$$

Note that $\hat{\rho}_{LU} = n\,\hat{\rho}_L/\{(n-1) + \hat{\rho}_L\}$, which is the same function of expression (4) that relates $\hat{\kappa}_{CU}$ with $\hat{\kappa}_C$. Therefore, as $\hat{\kappa}_C = \hat{\rho}_L$, then $\hat{\kappa}_{CU} = \hat{\rho}_{LU}$ and the two new estimators of ρ_L and κ_C (quadratic weights) are the same. Additionally, $\hat{\rho}_{LU} \geq \hat{\rho}_L$ if $\hat{\rho}_L \geq 0$. In the "Appendix 3" it is proved that $\hat{\rho}_{LU} = \hat{\rho}_{I2}$, thus $\hat{\kappa}_{CU} = \hat{\rho}_{LU} = \hat{\rho}_{I2}$.

2.3 Scott's pi

Scott (1955) defines the following measure of agreement

$$k_S = (I_o - I_e)/(1 - I_e) \quad \text{where} \quad I_e = \sum_i \sum_j w_{ij}\pi_i\pi_j, \ \text{with } \pi_i = (p_{i\cdot} + p_{\cdot i})/2, \tag{8}$$

and proposes to estimate it by

$$\hat{\kappa}_S = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \sum_i \sum_j w_{ij}\hat{\pi}_i\hat{\pi}_j \quad \text{with} \quad \hat{\pi}_i = \frac{\hat{p}_{i\cdot} + \hat{p}_{\cdot i}}{2} = \frac{O_{i\cdot} + O_{\cdot i}}{2n}, \tag{9}$$

As indicated in “Appendix 1”, $\hat{\pi}_i\hat{\pi}_j$ is not an unbiased estimator of $\pi_i\pi_j$ since

$$E(\hat{\pi}_i\hat{\pi}_j) = \frac{(n-1)\pi_i\pi_j + \left\{\delta_{ij}\left(p_{i\cdot} + p_{\cdot j}\right) + \left(p_{ij} + p_{ji}\right)\right\}/4}{n}. \tag{10}$$

Therefore, $\mathrm{E}(\hat{I}_e) = \sum_i \sum_j E\left(\hat{\pi}_i\hat{\pi}_j\right) = \{(n-1)I_e + (1+I_o)/2\}/n$, assuming that $w_{ij} = w_{ji}$, and $\hat{I}_e$ is not an unbiased estimator of I_e. From expression (10) it is deduced that the unbiased estimators of $\pi_i\pi_j$ and I_e are

$$\widehat{\pi_i\pi_j} = \frac{n\hat{\pi}_i\hat{\pi}_j - \left\{\left(\hat{p}_{i\cdot} + \hat{p}_{\cdot j}\right)\delta_{ij} + \left(\hat{p}_{ij} + \hat{p}_{ji}\right)\right\}/4}{n-1} \quad \text{and}$$
$$\hat{I}_{eU} = \sum_i w_{ij}\widehat{\pi_i\pi_j} = \frac{n\hat{I}_e - \left(1 + \hat{I}_o\right)/2}{n-1}, \tag{11}$$

respectively. Therefore, the new estimator $\hat{\kappa}_{SU}$ of κ_S will be

$$\hat{\kappa}_{SU} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{(2n-1)\hat{\kappa}_S + 1}{(2n-1) + \hat{\kappa}_S}, \tag{12}$$

and its variance, as followed in “Appendix 2”, is

$$V\left(\hat{\kappa}_{SU}\right) = \frac{(2n-1-\kappa_S)^4}{\{4n(n-1)\}^2} V\left(\hat{\kappa}_S\right). \tag{13}$$

Because of expression (11), $\hat{I}_{eU} - \hat{I}_e$ is proportional to $-\left\{\left(1-\hat{I}_e\right) + \left(\hat{I}_o - \hat{I}_e\right)\right\}$ which is also proportional to $-\left\{1 + \hat{\kappa}_S\right\} \leq 0$ if and only if $\hat{\kappa}_S \geq -1$. As $\hat{\kappa}_S$ decreases with $\hat{I}_e$, then $\hat{\kappa}_{SU} \geq \hat{\kappa}_S$ in the case of a positive agreement.

2.4 Krippendorf’s alpha

Krippendorf (1970, 2004) proposed to estimate κ_S as in expression (9), but with a small-sample correction for $\hat{I}_o$, though Gwet (2021b, p. 65) considers that “The need for such an adjustment and its potential benefits have not been documented”. The new estimator is,

$$\hat{\kappa}_K = \frac{\hat{I}_{oC} - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_{oC} = \frac{(2n-1)\hat{I}_o + 1}{2n} \quad \text{and} \quad \hat{I}_e = \sum_i \sum_j w_{ij}\hat{\pi}_i\hat{\pi}_j, \tag{14}$$

where $\hat{I}_{oC} = \hat{I}_o + \left(1 - \hat{I}_o\right) \Big/ 2n$; therefore,

$$\begin{aligned}\hat{\kappa}_K &= \frac{(2n-1)\hat{\kappa}_S + 1}{2n} \quad \text{and} \\ \hat{\kappa}_{KU} &= \frac{(2n-1)\hat{\kappa}_{SU} + 1}{2n} = \frac{(n-1) + \{2n(n-1)+1\}\hat{\kappa}_K}{2n(n-1) + n\hat{\kappa}_K}.\end{aligned} \tag{15}$$

The first expression follows from expressions (9) and (14); the second is obtained by replacing $\hat{I}_e$ for the value of $\hat{I}_{eU}$ in expression (11). From expressions (15) it is deduces that $\hat{\kappa}_K \geq \hat{\kappa}_S$ and $\hat{\kappa}_{KU} \geq \hat{\kappa}_{SU}$. Also, as for positive degrees of agreement it occurs that $\hat{\kappa}_{SU} \geq \hat{\kappa}_S$ then, due to expressions (15), $\hat{\kappa}_{KU} \geq \hat{\kappa}_K$. Finally, if in the first expression of Eq. (15) $\hat{\kappa}_S$ is replaced by $\{(2n-1)\hat{\kappa}_{SU} - 1\}/\{(2n-1) - \hat{\kappa}_{SU}\}$ − which is deduced from expression (12) − then $\hat{\kappa}_K = 2(n-1)\,\hat{\kappa}_{SU}/\{(2n-1) - \hat{\kappa}_{SU}\}$ and $\hat{\kappa}_{SU} \geq \hat{\kappa}_K$ if $\hat{\kappa}_{SU} \geq 0$. The overall conclusion is that $\hat{\kappa}_S \leq \hat{\kappa}_K \leq \hat{\kappa}_{SU} \leq \hat{\kappa}_{KU}$ for positive degrees of agreement.

Regarding the variance, it is sufficient to use the first part of the second expression (15) and then replacing $\mathrm{V}(\hat{\kappa}_{SU})$ with the value in expression (13); thus

$$V\left(\hat{\kappa}_{KU}\right) = \left(\frac{2n-1}{2n}\right)^2 \times \frac{(2n-1-\kappa_S)^4}{\{4n(n-1)\}^2} V\left(\hat{\kappa}_S\right).$$

2.5 Gwet's AC1/2

Gwet (2008) defines the next measure regarding *AC2* (*AC1* refers to the non-weighted case),

$$k_G = (I_o - I_e)/(1 - I_e) \quad \text{where} \quad I_e = W \times \sum_i \pi_i(1-\pi_i),\ /\{K(K-1)\}$$
$$\text{and} \quad W = \sum_i \sum_j w_{ij}, \tag{16}$$

and proposes to estimate it by

$$\hat{\kappa}_G = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \frac{W}{K(K-1)} \sum_i \hat{\pi}_i\left(1 - \hat{\pi}_i\right), \tag{17}$$

where π_i and $\hat{\pi}_i$ are obtained as in expressions (8) and (9). Once again it happens that $\hat{I}_e$ is not an unbiased estimator of I_e, because $\hat{\pi}_i^2$ is not an estimator of π_i^2 either. Using the first expression (11) to estimate π_i^2 in an unbiased way, we obtain that the unbiased estimators of π_i^2 and I_e are, respectively

$$\widehat{\pi_i^2} = \frac{n\hat{\pi}_i^2 - \left\{\left(\hat{p}_{i\cdot} + \hat{p}_{\cdot i}\right) + 2\hat{p}_{ii}\right\} \Big/ 4}{n-1}$$

$$\text{and} \quad \hat{I}_{eU} = \frac{1}{n-1}\left\{n\hat{I}_e - X\right\}, \quad \text{where} \quad X = \frac{W\left(1 - \sum_i \hat{p}_{ii}\right)}{2K(K-1)}. \tag{18}$$

Therefore, the new estimator $\hat{\kappa}_{GU}$ of κ_G will be

$$\hat{\kappa}_{GU} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{(n-1)\hat{\kappa}_G + Y}{(n-1) + Y} \quad \text{where} \quad Y = \frac{X - \hat{I}_e}{1 - \hat{I}_e}. \tag{19}$$

In "Appendix 1" it is proved that $\hat{I}_{eU} - \hat{I}_e \geq 0$, so it always happens that $\hat{\kappa}_{GU} \leq \hat{\kappa}_G$. It can be observed that it is not feasible to determine V($\hat{\kappa}_{GU}$) directly from the value of V($\hat{\kappa}_G$).

3 Case of multi-raters

Let there be n subjects ($s = 1, 2, \ldots, n$) classified by R raters ($r = 1, 2, \ldots, R$) in K types ($i = 1, 2, \ldots, K$). Let $x_{sr} = 1, 2, \ldots, K$ be the answer of rater r in subject s, values that are usually presented in a two-dimensional table in which the subjects are in rows and the raters in columns. For each row (subject), let R_{is} be the number of raters that answer i in subject s; obviously $R_{i+} = \Sigma_s R_{is}$ is the total number of i answers (for every rater), $R_{+s} = \Sigma_i R_{is} = R$ and $R_{++} = \Sigma_i \Sigma_s R_{is} = nR$. For each column (rater), let n_{ir} be the number of subjects classified as i by rater r; obviously $n_{+r} = \Sigma_i n_{ir} = n$, $n_{i+} = \Sigma_r n_{ir} = R_{i+}$ is the total number of i answers and $n_{++} = \Sigma_i \Sigma_r n_{ir} = nR = R_{++}$. The results of R_{is} and n_{ir} are usually presented as in Table 3(a) and (b) respectively.

3.1 Pairwise methods and the observed index of agreement

To define and estimate the measures regarding the $R > 2$ case, the *pairwise* methods will be used. These methods in some way offer an average for what happens in the $R(R-1)$ possible pairs of raters (r, r'), with $r, r' = 1, 2, \ldots, R$ and $r \neq r'$. This obliges us to change the notation used in Sect. 2, since it is necessary to indicate for each parameter from which pair (r, r') does its value come from. Parameters p_{ij}, $p_{i\cdot}$ and $p_{\cdot j}$ of Sect. 2 will now be notated as $p_{ir,jr'}$, p_{ir} and $p_{jr'}$ respectively. Additionally, we define the new parameter $p_{i+} = \Sigma_r p_{ir} = \Sigma_{r'} p_{ir'}$, which is the proportion of i answers of every raters. A similar thing occurs with the estimated values $\hat{p}_{ij}$ and $\hat{p}_{ir,jr'}$ etc. Note that the estimators $\hat{p}_{ir}$ of p_{ir} and $\hat{p}_{i+}$ of p_{i+} are

$$\hat{p}_{ir} = \frac{n_{ir}}{n} \quad \text{and} \quad \hat{p}_{i+} = \sum\nolimits_r \hat{p}_{ir} = \frac{n_{i+}}{n} = \frac{R_{i+}}{n}, \tag{20}$$

respectively, where $\Sigma_i\ \hat{p}_{ir} = 1$ and $\Sigma_r \Sigma_i\ \hat{p}_{ir} = R$. Parameters κ, I_o and I_e of Sect. 2 will be denoted as $\kappa(r, r')$, $I_o(r, r')$ and $I_e\ (r, r')$ respectively; therefore

$$k\left(r, r'\right) = \{I_o\left(r, r'\right) - I_e\left(r, r'\right)\}/\{1 - I_e\left(r, r'\right)\}, \tag{21}$$

Table 3 Results of the classification of $n = 29$ fish by $R = 4$ raters in $K = 5$ colorations (Gwet 2021b, p 341)

Fish (s)	Coloration (i)					
	1	2	3	4	5	R_{+s}
(a) Number of raters R_{is} that classify the fish s in category i (Gwet 2021b, p. 342)						
1	0	0	0	0	4	4
2	2	0	2	0	0	4
3	0	0	0	0	4	4
4	2	0	2	0	0	4
5	0	0	0	1	3	4
6	1	1	2	0	0	4
7	3	0	1	0	0	4
8	3	0	1	0	0	4
9	0	0	2	2	0	4
10	3	0	1	0	0	4
11	0	0	0	0	4	4
12	4	0	0	0	0	4
13	4	0	0	0	0	4
14	4	0	0	0	0	4
15	0	0	3	1	0	4
16	1	0	2	1	0	4
17	0	0	0	2	2	4
18	0	0	0	0	4	4
19	0	0	3	0	1	4
20	0	1	3	0	0	4
21	0	0	1	0	3	4
22	0	0	3	1	0	4
23	4	0	0	0	0	4
24	4	0	0	0	0	4
25	2	0	2	0	0	4
26	1	0	3	0	0	4
27	2	0	2	0	0	4
28	2	0	2	0	0	4

and the same for the estimated values $\hat{\kappa}\left(r, r'\right)$ etc.

With *pairwise* methods there are several ways to average the results of every pair of raters (r, r'), but all procedures of interest define the global value of I_o as

$$I_o = \sum_{r} \sum_{r' \neq r} I_o\left(r, r'\right) / \{R(R-1)\}, \tag{22}$$

Table 3 (continued)

Fish (s)	Coloration (i)					
	1	2	3	4	5	R_{+s}
29	0	1	2	0	1	4
R_{i+}	42	3	37	8	26	$R_{++} = 116$

Rater (r)	Coloration (i)					n_{+r}
	1	2	3	4	5	
(b) Values of n_{ir} or number of replies i of the rater r (Gwet 2021b, p 75)						
1	10	0	11	1	7	29
2	10	2	11	1	5	29
3	10	1	9	3	6	29
4	12	0	6	3	8	29
n_{i+}	42	3	37	8	26	$n_{++} = 116$

Coefficients	Classic	New
(c) Estimated *kappa* coefficients		
Hubert's *kappa*	$\hat{\kappa}_H = 0.413$	$\hat{\kappa}_{HU} = 0.421$
Fleiss's *kappa*	$\hat{\kappa}_F = 0.410$	$\hat{\kappa}_{FU} = 0.422$
Fleiss's *kappa two-pairwise*	$\hat{\kappa}_{F2} = 0.408$	$\hat{\kappa}_{F2U} = 0.422$
Krippendorf's *alpha*	$\hat{\kappa}_K = 0.421$	$\hat{\kappa}_{KU} = 0.432$
Krippendorf's *alpha two-pairwise*	$\hat{\kappa}_{K2} = 0.418$	$\hat{\kappa}_{K2U} = 0.432$
Gwet's *AC1*	$\hat{\kappa}_G = 0.445$	$\hat{\kappa}_{GU} = 0.441$
Gwet's *AC1 two-pairwise*	$\hat{\kappa}_{G2} = 0.490$	$\hat{\kappa}_{G2U} = 0.487$

thus $I_o = \Sigma_r \Sigma_{r' \neq r} \Sigma_i \Sigma_j w_{ij} p_{ir,jr'} / \{R(R-1)\}$. As is traditional, the measure of global agreement will be $\kappa = (I_o - I_e)/(1 - I_e)$, where I_e is yet to be defined. If I_e is defined in a similar way to I_o

$$I_e = \sum_r \sum_{r' \neq r} I_e(r, r') / \{R(R-1)\}, \tag{23}$$

we say that the procedure that defines global κ is a "*two-pairwise*" procedure and the population coefficient thereby obtained will be,

$$k_2 = \left\{ \sum_r \sum_{r' \neq r} I_o(r, r') - \sum_r \sum_{r' \neq r} I_e(r, r') \right\} \bigg/ \left\{ R(R-1) - \sum_r \sum_{r' \neq r} I_e(r, r') \right\}.$$

It can be noticed that κ_2 is also obtained by dividing the sum of all the possible numerators ($\Sigma_r \Sigma_{r'\neq r}$) from expression (21), by the sum of all possible denominators, which indicates that κ_2 if the weighted average of $R(R-1)$ values of $\kappa(r, r')$ – the weights are the denominators – . This procedure is the one recommended by Janson and Olsson (2001), Conger (1980) and Gwet (2021b). Notice that $\Sigma_r \Sigma_{r'\neq r} I_o(r, r') = 2\Sigma_r \Sigma_{r'>r} I_o(r, r')$ and similarly with I_e. We have preferred to use the first expression because it facilitates some proofs, but regarding calculations the second expressions seems preferable. All of the above also applies to the case of estimated values.

As the base values of I_o and $\hat{I}_o$ are the same in every κ measures, it should be specified since its values are (see "Appendix 1"),

$$I_o = \frac{\sum_r \sum_{r'\neq r} \sum_i \sum_j w_{ij} p_{ir,jr'}}{R(R-1)},$$
$$\hat{I}_o = \frac{\sum_r \sum_{r'\neq r} \sum_i \sum_j w_{ij} \hat{p}_{ir,jr'}}{R(R-1)} = \frac{\sum_i \sum_j w_{ij} \sum_s R_{is} R_{js} - nR}{nR(R-1)}, \tag{24}$$

3.2 Hubert's kappa pairwise and the intraclass and concordance correlation coefficients

The κ_H coefficient of Hubert (Hubert 1977; Conger 1980) is a *two-pairwise* coefficient, and that is why the expression (23) can be applied for value $I_e(r, r')$ of Cohen. Adjusting expression (1) to the current format, $I_e(r, r') = \Sigma_i \Sigma_j w_{ij} p_{ir} p_{jr'}$ and, due to "Appendix 1"

$$k_H = (I_o - I_e)/(1 - I_e)$$
$$\text{where} \quad I_e = \sum_i \sum_j w_{ij} \left(p_{i+} p_{j+} - \sum_r p_{ir} p_{jr} \right) / \{R(R-1)\}. \tag{25}$$

Using expressions (20) the following estimation is obtained

$$\hat{\kappa}_H = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \frac{1}{n^2 R(R-1)} \sum_i \sum_j w_{ij} \left\{ n_{i+} n_{j+} - \sum_r n_{ir} n_{jr} \right\}.$$

It can be observed that for $R = 2$ it occurs that $\kappa_C = \kappa_H$ and $\hat{\kappa}_C = \hat{\kappa}_H$. In order to obtain an unbiased estimator of I_e, the second expression of (3), applied with the current notation, indicates that $\hat{I}_{eU}(r, r') = \left\{ n\hat{I}_e(r, r') - \hat{I}_o(r, r') \right\} \big/ (n-1)$; therefore $R(R-1)\hat{I}_{eU} = \sum_r \sum_{r'\neq r} \hat{I}_{eU}(r, r') = \left\{ n \sum_r \sum_{r'} \hat{I}_e(r, r') - \sum_r \sum_{r'} \hat{I}_o(r, r') \right\} \big/ (n-1)$ and so $\hat{I}_{eU} = (n\ \hat{I}_e - \hat{I}_o)/\ (n-1)$. As this expression is the same as the second expression of (3), then the conclusions in Sect. 2.2 are still valid, changing the letter C with the letter H. Thus,

$$\hat{\kappa}_{HU} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{n\hat{\kappa}_H}{(n-1) + \hat{\kappa}_H}, \tag{26}$$

and $\hat{\kappa}_{HU} \geq \hat{\kappa}_H$ in the case of positive agreement.

Generalizing the first expression of (6) in the case of two raters r and r' of answers x_r and $x_{r'}$, means μ_r and $\mu_{r'}$, variances σ_r^2 and $\sigma_{r'}^2$, and covariances $\sigma_{rr'}$, we obtain $\rho_L(r, r') = 2\sigma_{rr'} / \left\{\sigma_r^2 + \sigma_{r'}^2 + (\mu_r - \mu_{r'})^2\right\}$. If we apply to this expression the *two-pairwise* criterion which consists of adding $\Sigma_r \Sigma_{r \neq r'}$ in the numerator and in the denominator, the CCC ρ_L of Lin (1989, 2000) and Barnhart et al. (2002) is obtained for the case of multi-raters; its estimated $\hat{\rho}_L$ value is obtained in the same way as the second expression of (6). In this way,

$$\rho_L = \frac{2\sum_r \sum_{r'>r} \sigma_{rr'}}{(R-1)\sum_r \sigma_r^2 + \sum_r \sum_{r'>r} (\mu_r - \mu_{r'})^2},$$
$$\hat{\rho}_L = \frac{2\sum_r \sum_{r'>r} S_{rr'}}{(R-1)\sum_r S_r^2 + \sum_r \sum_{r'>r} (\overline{x}_r - \overline{x}_{r'})^2}. \tag{27}$$

Carrasco and Jover (2003) justified that $\hat{\rho}_L$ is based on biased estimators and they proposed the following estimator, which is based on unbiased estimators ($s_{rr'}$ and s_r^2)

$$\hat{\rho}_{LU} = \frac{2n\sum_r \sum_{r'>r} s_{rr'}}{(R-1)(n-1)\sum_r s_r^2 + n\sum_r \sum_{r'>r} (\overline{x}_r - \overline{x}_{r'})^2 + 2\sum_r \sum_{r'>r} s_{rr'}}. \tag{28}$$

It is easy to see that the same thing can be obtained applying the *two-pairwise* method to the first expression (7). As for $R = 2$ it occurred that $\kappa_C = \rho_L$ and $\hat{\kappa}_C = \hat{\rho}_L$ when the weights were quadratic, and in both cases the value for $R > 2$ is obtained in the same way – the sum of the numerators divided by the sum of the denominators – , then also $\kappa_H = \rho_L$ and $\hat{\kappa}_H = \hat{\rho}_L$ in the case of $R > 2$. Additionally, $\kappa_{HR} = \kappa_H = \rho_L = \rho_{I2}$ since $\rho_L = \rho_{I2}$ (Carrasco and Jover 2003) and $\kappa_{HR} = \rho_L$ (Martín Andrés and Álvarez Hernández 2020). Furthermore, as $\hat{\rho}_{LU} = n\hat{\rho}_L / \left\{(n-1) + \hat{\rho}_L\right\}$ -an expression which has the same form as (26)- then also

$$\hat{\kappa}_{HU} = \hat{\rho}_{LU} = \hat{\rho}_{I2} = \frac{n\sum_s x_{s\cdot}^2 + \sum_r x_{\cdot r}^2 - n\sum_s \sum_r x_{sr}^2 - x_{\cdot\cdot}^2}{\sum_s x_{s\cdot}^2 + \sum_r x_{\cdot r}^2 + (nR - n - R)\sum_s \sum_r x_{sr}^2 - x_{\cdot\cdot}^2}, \tag{29}$$

where the last two equalities are demonstrated in the "Appendix 3". In the last expression, which is simpler for the calculation, it is understood that $x_{s\cdot} = \sum_r x_{sr}$, $x_{\cdot r} = \sum_s x_{sr}$, and $x_{\cdot\cdot} = \sum_s \sum_r x_{sr}$. Something similar happens with the estimators based on the biased estimation of their components (see "Appendix 3"),

$$\hat{\kappa}_H = \hat{\rho}_L = \frac{n\sum_s x_{s\cdot}^2 + \sum_r x_{\cdot r}^2 - n\sum_s\sum_r x_{sr}^2 - x_{\cdot\cdot}^2}{\sum_r x_{\cdot r}^2 + n(R-1)\sum_s\sum_r x_{sr}^2 - x_{\cdot\cdot}^2}. \tag{30}$$

3.3 Fleiss' kappa pairwise

Fleiss (1971) extended κ_S to the case of $R > 2$ defining in the following value of I_e, which is not a *two-pairwise* type,

$$k_F = (I_o - I_e)/(1 - I_e) \quad \text{where} \quad I_e = \sum_i\sum_j w_{ij}\pi_i\pi_j$$
$$\text{and} \quad \pi_i = \sum_r p_{ir}/R = p_{i+}/R, \tag{31}$$

and proposes the following estimators

$$\hat{\kappa}_F = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \sum_i\sum_j w_{ij}\hat{\pi}_i\hat{\pi}_j = \frac{1}{n^2R^2}\sum_i\sum_j w_{ij}R_{i+}R_{j+}$$
$$\text{and} \quad \hat{\pi}_i = \frac{\hat{p}_{i+}}{R}, \tag{32}$$

since p_{i+} is estimated as the second expression of Eq. (20). As indicated in "Appendix 1", $\hat{I}_e$ is not an unbiased estimator of I_e since $n\mathrm{E}(\hat{I}_e) = (n-1)I_e + R^{-1}\{1 + (R - 1)I_o\}$. This is why the unbiased estimator $\hat{I}_{eU}$ of I_e and the new estimator $\hat{\kappa}_{FU}$ of κ_F will be

$$\hat{I}_{eU} = \frac{n\hat{I}_e - \left\{1 + (R-1)\hat{I}_o\right\}\Big/ R}{n-1}$$
$$\text{and} \quad \hat{\kappa}_{FU} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{(Rn-1)\hat{\kappa}_F + 1}{(R-1)\hat{\kappa}_F + \{R(n-1)+1\}}. \tag{33}$$

Its variance, as deduced in "Appendix 2", is

$$V\left(\hat{\kappa}_{FU}\right) = \frac{\{(nR-1) - (R-1)\kappa_F\}^4}{\left\{R^2 n(n-1)\right\}^2} V\left(\hat{\kappa}_F\right). \tag{34}$$

Through the first expression of Eq. (33), $\hat{I}_{eU} - \hat{I}_e$ is proportional to $\hat{I}_e - R^{-1}\{1 + (R-1)\,\hat{I}_o\}$, which is also proportional to $-\{1 + (R-1)\hat{\kappa}_F\} \le 0$ if and only if $\hat{\kappa}_F \ge -(R-1)^{-1}$. Therefore, $\hat{\kappa}_{FU} \ge \hat{\kappa}_F$ in the case of positive agreement.

Another way of extending κ_S is to use the *two-pairwise* method. In this case, in "Appendix 1" it is demonstrated that

$$k_{F2} = (I_o - I_e)/(1 - I_e)$$
$$\text{where} \quad I_e = \left[\sum_i \sum_j w_{ij} \left\{(R-2)\sum_r p_{ir}p_{jr} + p_{i+}p_{j+}\right\}\right] / 2R(R-1), \tag{35}$$

and therefore its estimated values in a traditional way would be

$$\hat{\kappa}_{F2} = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \frac{1}{2n^2 R(R-1)} \sum_i \sum_j w_{ij} \left\{(R-2)\sum_r n_{ir}n_{jr} + n_{i+}n_{j+}\right\}.$$

In order to obtain the unbiased estimator of I_e, the second expression of Eq. (11) is, in the current terms, $\hat{I}_{eU}(r, r') = [n\hat{I}_e(r, r') - \{1 + \hat{I}_o(r, r')/2\}]/(n-1)$. Applying expressions (22) and (23) it is obtained that the second expression of Eq. (11) is also applied to the current case, in such a way that the conclusions obtained in the case of Scott's *Pi* are valid, changing the letter *S* with *F2*. In this way

$$\hat{\kappa}_{F2U} = \frac{(2n-1)\hat{\kappa}_{F2} + 1}{(2n-1) + \hat{\kappa}_{F2}},$$
$$\text{where} \quad \hat{I}_{eU} = \frac{n\hat{I}_e - \left(1 + \hat{I}_o\right)/2}{n-1}, \; V\left(\hat{\kappa}_{F2U}\right) = \frac{(2n - 1 - \kappa_{F2})^4}{\{4n(n-1)\}^2} V\left(\hat{\kappa}_{F2}\right),$$

and $\hat{\kappa}_{F2U} \geq \hat{\kappa}_{F2}$ when $\hat{\kappa}_{F2} \geq 0$. Nevertheless, to the best of our knowledge, now the value of $\mathrm{V}(\hat{\kappa}_{F2})$ is not known.

3.4 Krippendorf's multi-rater alpha

Now the objective is similar to that of Sect. 2.4: to estimate κ_F as in expression (32), but changing the value of $\hat{I}_o$ for a value $\hat{I}_{oC}$ defined as expression (14). In this way

$$\hat{\kappa}_K = \frac{\hat{I}_{oC} - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_{oC} = \frac{(2n-1)\hat{I}_o + 1}{2n} \quad \text{and} \quad \hat{I}_e = \frac{1}{n^2 R^2} \sum\nolimits_i \sum\nolimits_j w_{ij} R_{i+} R_{j+}.$$

Given the formal equality of the expressions, all of the previous conclusions can be accepted, with the necessary changes. In particular,

$$\hat{\kappa}_K = \frac{(2n-1)\hat{\kappa}_F + 1}{2n} \quad \text{and} \quad \hat{\kappa}_{KU} = \frac{(2n-1)\hat{\kappa}_{FU} + 1}{2n} = \frac{(n-1) + \{2n(n-1) + 1\}\hat{\kappa}_K}{2n(n-1) + n\hat{\kappa}_K}, \tag{36}$$

$$\hat{\kappa}_F \leq \hat{\kappa}_K \leq \hat{\kappa}_{FU} \leq \hat{\kappa}_{KU}, \tag{37}$$

$$V\left(\hat{\kappa}_{KU}\right)=\left(\frac{2n-1}{2n}\right)^2\times\frac{(2n-1-\kappa_S)^4}{\{4n(n-1)\}^2}V\left(\hat{\kappa}_F\right). \tag{38}$$

In a similar way for the *two-pairwise* method, where now

$$\hat{\kappa}_{K2}=\frac{\hat{I}_{oC}-\hat{I}_e}{1-\hat{I}_e},\quad \text{where}\quad \hat{I}_{oC}=\frac{(2n-1)\hat{I}_o+1}{2n},$$
$$\text{and}\quad \hat{I}_e=\frac{\sum_i\sum_j w_{ij}\left\{(R-2)\sum_r n_{ir}n_{jr}+n_{i+}n_{j+}\right\}}{2n^2R(R-1)}.$$

Therefore, expressions (36) to (38) are also valid putting number “2” after the letters K or F in the sub-indexes of these expressions.

3.5 Gwet’s multi-rater AC1/2

For the case of multi-raters, Gwet (2008) defined the same measures of agreement *AC1/2* κ_G and $\hat{\kappa}_G$ of expressions (16) and (17) respectively, but with π_i and $\hat{\pi}_i$ alluding to the Fleiss values of expressions (31) and (32) respectively. Therefore, $I_e=W\left(1-\sum_i\pi_i^2\right)/\left\{K(K-1)\right\}=W\left(1-\sum_i p_{i+}^2/R^2\right)/\{K(K-1)\}$ and

$$\hat{I}_e=\frac{W}{K\left(K-1\right)}\left\{1-\sum_i\hat{\pi}_i^2\right\}=\frac{W}{K\left(K-1\right)}\left\{1-\frac{\sum_i\hat{p}_{i+}^2}{R^2}\right\}$$
$$=\frac{W}{K\left(K-1\right)}\left\{1-\frac{\sum_i R_{i+}^2}{n^2R^2}\right\}. \tag{39}$$

"Appendix 1" demonstrates that $\hat{\pi}_i^2$ is not an unbiased estimator of π_i^2 − see expression (48) − , so that $\hat{I}_e$ is also not an unbiased estimator of I_e, which is justified in this same Appendix as the unbiased estimator $\hat{I}_{eU}$ of I_e is

$$\hat{I}_{eU}=\frac{n\hat{I}_e-A}{n-1},\quad \text{where}\quad A=\frac{W(R-1)\left(1-\hat{I}_{oN}\right)}{RK(K-1)}\quad \text{and}\quad \hat{I}_{oN}=\frac{\sum_i R_{is}^2-nR}{nR(R-1)}. \tag{40}$$

Therefore, the new estimator $\hat{\kappa}_{GU}$ of κ_G will be,

$$\hat{\kappa}_{GU}=\frac{\hat{I}_o-\hat{I}_{eU}}{1-\hat{I}_{eU}}=\frac{(n-1)\hat{\kappa}_G+B}{(n-1)+B}\quad \text{where}\quad B=\frac{A-\hat{I}_e}{1-\hat{I}_e}. \tag{41}$$

It can be observed that now it is not viable to determine V($\hat{\kappa}_{GU}$) directly from the value of V($\hat{\kappa}_G$). “Appendix 1” demonstrates that $\hat{I}_{eU}-\hat{I}_e\geq 0$, so that now we also find that $\hat{\kappa}_{GU}\leq\hat{\kappa}_G$.

An alternative is to use the *two-pairwise* method. In this case, "Appendix 1" demonstrates that

$$\kappa_{G2} = \frac{I_o - I_e}{1 - I_e} \quad \text{where} \quad I_e = \frac{W}{K(K-1)}\left[1 - \frac{1}{2R(R-1)}\left\{(R-2)\sum_i \sum_r p_{ir}^2 + \sum_i p_{i+}^2\right\}\right], \tag{42}$$

and therefore its estimated (biased) values are, because of expression (20)

$$\hat{\kappa}_{G2} = \frac{\hat{I}_o - \hat{I}_e}{1 - \hat{I}_e} \quad \text{where} \quad \hat{I}_e = \frac{W}{K(K-1)}\left[1 - \frac{1}{2R(R-1)n^2}\left\{(R-2)\sum_i \sum_r n_{ir}^2 + \sum_i n_{i+}^2\right\}\right]. \tag{43}$$

To obtain unbiased estimator of I_e, expression (18) is, in current terms, $\hat{I}_{eU}\left(r, r'\right) = [n\hat{I}_e\left(r, r'\right) - W\{1 - \sum_i \hat{p}_{ir,ir'}\}/\{2K(K-1)\}]/(n-1)$. Applying expression (23) we obtain the value for the current $\hat{I}_{eU}$, which provides the value of $\hat{\kappa}_{G2U}$; i.e.

$$\hat{\kappa}_{G2U} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} \quad \text{where} \quad \hat{I}_{eU} = \frac{n\hat{I}_e - X_N}{n-1},\ X_N = \frac{W\left(1 - \hat{I}_{oN}\right)}{2K(K-1)}, \tag{44}$$

and $\hat{I}_{oN}$ as in expression (40). Note that in this expression $\hat{I}_{eU}$ has the same form as in expression (18), so that $\hat{\kappa}_{G2U}$ can be put as a function of $\hat{\kappa}_{G2}$ in a similar way to in expression (19):

$$\hat{\kappa}_{G2U} = \frac{\hat{I}_o - \hat{I}_{eU}}{1 - \hat{I}_{eU}} = \frac{(n-1)\hat{\kappa}_{G2} + Y_N}{(n-1) + Y_N} \quad \text{where} \quad Y_N = \frac{X_N - \hat{I}_e}{1 - \hat{I}_e}.$$

As in case $R = 2$ it occurred that $\hat{I}_{eU}\left(r, r'\right) \geq \hat{I}_e\left(r, r'\right)$, through expression (23) it is deduced that in the actual case $\hat{I}_{eU} \geq \hat{I}_e$; therefore $\hat{\kappa}_{G2U} \leq \hat{\kappa}_{G2}$. "Appendix 1" provides a more direct demonstration of the previous statement. To the best of our knowledge, the value of $\mathrm{V}(\hat{\kappa}_{G2})$ is not known.

4 Examples

Table 1(a) contains the data from a classic example by Fleiss et al. (2003) in which $R = 2$ raters diagnose $n = 100$ individuals in $K = 3$ categories (Psychotic, Neurotic, and Organic). Its part (b) specifies the values of the eight *kappa* coefficients mentioned in Sect. 2, all of which are calculated for the non-weighted case ($w_{ij} = \delta_{ij}$). It can be observed that the eight coefficients verify the properties mentioned in Sect. 2; for example, all of the new estimators have a value greater than or equal to that of the classic ones, except in the case of the coefficient of Gwet in which case the opposite happens. Nevertheless, the first are only slightly different from the latter. This is due to the fact that the current sample size ($n = 100$) is too large to show the differences between the estimators. When the sample size is small ($n = 8$), as occurs in the example of Gwet 2021b (p 109) in Table 2(a) ($R = 2$, $K = 3$), the differences are more evident, as shown by the results in Table 2(b).

For the case of more than two raters, Table 3(a) and (b) show the values of R_{is} and n_{ir}, respectively, values which are obtained from the data x_{sr} in an example by Gwet 2021b (p 341) related to the change in the coloring of Stickleback fish ($R = 4$, $K = 5$, $n = 50$). Table 3(c) shows the values of the fourteen *kappa* coefficients mentioned in Sect. 3, all of which are also calculated for the non-weighted case ($w_{ij} = \delta_{ij}$). It can be observed that the fourteen coefficients verify the properties mentioned in Sect. 3. It is also observed that although the values of n and $\hat{\kappa}$ are moderate, all of the new coefficients are greater than the classic ones in at least one unit of the second decimal. The exception is the case of the two coefficients of Gwet, in which the differences obtained are very small.

5 Simulation

This section has two objectives. Firstly, to assess the bias of the two estimators of κ_X ($\hat{\kappa}_X$ and $\hat{\kappa}_{XU}$) in the case of $R = 2$, where X refers to C, S, K or G. Secondly, to assess the behaviour of the estimator of the variance $\hat{V}\left(\hat{\kappa}_{CU}\right)$, in order to exemplify that the new variances act coherently in relation to the classic ones.

To assess the two estimators, the procedure is as follows. Let us consider that the observed frequencies in Table 1(a), divided by $n = 100$, are the true probabilities p_{ij} of the problem mentioned, in which $R = 2$ and $K = 3$; for example, $p_{11} = 75/100 = 0.75$. In that case the value $\hat{\kappa}_C = 0.676$ of the Table 1(b) becomes the population value $\kappa_C = 0.676$ of the Cohen *kappa* coefficient, since the values $\hat{I}_o$ and $\hat{I}_e$ of $\hat{\kappa}_C$ become the values I_o and I_e of κ_C. If we now extract $N = 10{,}000$ random samples of the multinomial distribution of parameters $\{p_{ij}, n = 100\}$, each sample will provide two estimators $\hat{\kappa}_{Ch}$ and $\hat{\kappa}_{CUh}$ of κ_C. The means $\overline{\hat{\kappa}}_C = \Sigma_h \hat{\kappa}_{Ch} \big/ N$ and $\overline{\hat{\kappa}}_{CU} = \Sigma_h \hat{\kappa}_{CUh} \big/ N$ of the values $\hat{\kappa}_{Ch}$ and $\hat{\kappa}_{CUh}$ should be approximately equal to $\kappa_C = 0.676$ if the estimators were unbiased. The results of this simulation are provided on the sixteenth line of results in Table 4. The rest of the lines, where other values of K, n, and κ_C are used, were obtained in a similar way. It can be seen that in general $\kappa_C = \overline{\hat{\kappa}}_{CU} \geq \overline{\hat{\kappa}}_C$, except in two case in which $\kappa_C > \overline{\hat{\kappa}}_C \geq \overline{\hat{\kappa}}_{CU}$. Therefore, $\hat{\kappa}_{CU}$ is less biased than $\hat{\kappa}_C$ and, for the accuracy used, is generally unbiased. Nevertheless, $\hat{\kappa}_C$ is only unbiased for values $n \geq 50$ or 100, depending on the value of K.

The same tables and previous simulations allow us to obtain the corresponding results of the other two pairs of estimators (see the rest of Table 4). In the case of Scott's *pi* coefficient, it is also observed that $\kappa_S = \overline{\hat{\kappa}}_{SU} \geq \overline{\hat{\kappa}}_S$, except in four cases in which $\kappa_S > \overline{\hat{\kappa}}_{SU} \geq \overline{\hat{\kappa}}_S$, so that $\hat{\kappa}_{SU}$ is also generally unbiased; additionally $\overline{\hat{\kappa}}_{SU} = \overline{\hat{\kappa}}_S$ only for $n = 100$. The conclusions are a little different in the case of Krippendorf's *alpha* coefficient; in general it still occurs that $\kappa_K = \overline{\hat{\kappa}}_{KU} \geq \overline{\hat{\kappa}}_K$, except in five cases in which $\kappa_K < \overline{\hat{\kappa}}_{KU}$ or $\kappa_K > \overline{\hat{\kappa}}_{KU}$, in such a way that $\hat{\kappa}_{KU}$ may also underestimate κ_K; now $\overline{\hat{\kappa}}_{KU} = \overline{\hat{\kappa}}_K$ on some occasions when $n \geq 50$. As can be seen, the three pairs of previous coefficients are either unbiased or they underestimate the value of the

Table 4 Results of the 10,000 simulations performed for the *kappa* values indicated

K	n	Cohen' kappa			Scott's pi			Krippendorf's alpha			Gwet's AC1		
		True	Estimated (mean)		True	Estimated (mean)		True	Estimated (mean)		True	Estimated (mean)	
			Classic	New		Classic	New		Classic	New		Classic	New
		κ_C	$\overline{\hat{\kappa}}_C$	$\overline{\hat{\kappa}}_{CU}$	κ_S	$\overline{\hat{\kappa}}_S$	$\overline{\hat{\kappa}}_{SU}$	κ_K	$\overline{\hat{\kappa}}_K$	$\overline{\hat{\kappa}}_{KU}$	κ_G	$\overline{\hat{\kappa}}_G$	$\overline{\hat{\kappa}}_{GU}$
2	10	0.38	0.34	0.35	0.38	0.29	0.33	0.41	0.32	0.36	0.71	0.68	0.67
		0.80	0.79	0.80	0.80	0.78	0.80	0.81	0.79	0.81	0.80	0.82	0.80
	20	0.41	0.40	0.41	0.40	0.38	0.40	0.42	0.39	0.41	0.40	0.42	0.40
		0.80	0.79	0.80	0.80	0.79	0.80	0.80	0.79	0.80	0.80	0.81	0.80
	50	0.39	0.39	0.39	0.39	0.38	0.39	0.39	0.39	0.39	0.41	0.42	0.41
		0.80	0.80	0.80	0.80	0.79	0.80	0.80	0.80	0.80	0.80	0.81	0.80
	100	0.41	0.40	0.41	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.40
		0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.84	0.84	0.84
3	10	0.41	0.39	0.41	0.39	0.35	0.39	0.42	0.38	0.42	0.40	0.42	0.40
		0.83	0.82	0.83	0.83	0.81	0.82	0.84	0.82	0.83	0.86	0.86	0.86
	20	0.42	0.41	0.42	0.39	0.37	0.39	0.41	0.39	0.41	0.40	0.41	0.41
		0.77	0.76	0.77	0.77	0.76	0.77	0.78	0.77	0.78	0.78	0.78	0.77
	50	0.40	0.40	0.40	0.40	0.39	0.40	0.41	0.40	0.41	0.44	0.44	0.44
		0.79	0.79	0.79	0.79	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79
	100	0.38	0.38	0.38	0.38	0.38	0.38	0.39	0.38	0.39	0.41	0.41	0.41
		0.68	0.67	0.67	0.68	0.67	0.67	0.68	0.67	0.68	0.87	0.87	0.87
5	10	0.44	0.42	0.44	0.80	0.80	0.80	0.80	0.80	0.80	0.87	0.87	0.87

Table 4 (continued)

K	n	Cohen' kappa			Scott's pi			Krippendorf's alpha			Gwet's AC1		
		True	Estimated (mean)		True	Estimated (mean)		True	Estimated (mean)		True	Estimated (mean)	
			Classic	New		Classic	New		Classic	New		Classic	New
		κ_C	$\overline{\hat{\kappa}}_C$	$\overline{\hat{\kappa}}_{CU}$	κ_S	$\overline{\hat{\kappa}}_S$	$\overline{\hat{\kappa}}_{SU}$	κ_K	$\overline{\hat{\kappa}}_K$	$\overline{\hat{\kappa}}_{KU}$	κ_G	$\overline{\hat{\kappa}}_G$	$\overline{\hat{\kappa}}_{GU}$
		0.85	0.84	0.85	0.43	0.40	0.44	0.46	0.43	0.47	0.51	0.52	0.51
	20	0.38	0.37	0.38	0.85	0.84	0.85	0.85	0.84	0.86	0.88	0.88	0.88
		0.81	0.80	0.81	0.37	0.35	0.37	0.39	0.37	0.39	0.38	0.38	0.37
	50	0.40	0.39	0.40	0.81	0.80	0.81	0.81	0.80	0.81	0.81	0.82	0.81
		0.80	0.79	0.80	0.39	0.38	0.39	0.40	0.39	0.40	0.40	0.41	0.40
	100	0.39	0.39	0.39	0.80	0.79	0.80	0.80	0.80	0.80	0.80	0.80	0.80
		0.80	0.80	0.80	0.39	0.38	0.39	0.39	0.39	0.39	0.45	0.45	0.45

Only two decimal places are used for an easier evaluation of the differences between the different *kappa* values populational parameter. In the case of Gwet's *AC1* coefficient, the opposite happens. In general $\kappa_G = \overline{\hat{\kappa}}_{GU} \leq \overline{\hat{\kappa}}_G$, except in four cases in which $\kappa_G < \overline{\hat{\kappa}}_{GU}$ or $\kappa_G > \overline{\hat{\kappa}}_{GU}$, so that both estimators are either unbiased or they overestimate the value of the populational parameter. Now the equality $\overline{\hat{\kappa}}_{GU} = \overline{\hat{\kappa}}_G$ generally happens when $K > 2$ and $n \geq 50$.

The general conclusion is that the estimators $\hat{\kappa}_{XU}$ are generally unbiased and, when they are biased, their bias is lower than that of the estimators $\hat{\kappa}_X$. When there is bias, it is positive in the case of the Gwet coefficient, and is negative in the other three cases.

Let us now consider the case of variance. The classic estimator $\hat{\kappa}_C$ has an unknown variance $V_E(\hat{\kappa}_C)$ which can be estimated in a quite precise way through the sample variance $\hat{V}_E(\hat{\kappa}_C)$ of the values $\hat{\kappa}_{Ch}$ of the 10,000 simulations. Moreover, each simulation provides an estimator $\hat{V}_h(\hat{\kappa}_C)$ of $V_E(\hat{\kappa}_C)$ obtained through the formula of Fleiss et al. (1969); the average value $\overline{\hat{V}}(\hat{\kappa}_C)$ of these 10,000 estimators, compared to $\hat{V}_E(\hat{\kappa}_C)$, allows us to check the bias of this estimator of the variance. The same reasoning is used in the case of the estimator $\hat{\kappa}_{CU}$, although now $\hat{V}_h(\hat{\kappa}_{CU})$ is obtained through expression (5). The results are in Table 5. It can be seen that $\hat{V}_E(\hat{\kappa}_{CU}) \approx \hat{V}_E(\hat{\kappa}_C)$ for $n \geq 20$, being in general $V_E(\hat{\kappa}_{CU}) > (<)\ V_E(\hat{\kappa}_C)$ when $\kappa_C = 0.4$ (0.8). It is also observed to that the classic variance $\overline{\hat{V}}(\hat{\kappa}_C)$ usually underestimates (overestimates) $\hat{V}_E(\hat{\kappa}_C)$ when $\kappa_C = 0.4$ (0.8), the differences being small when $n \geq 50$. However, the new variance $\overline{\hat{V}}(\hat{\kappa}_{CU})$ almost always underestimates $\hat{V}_E(\hat{\kappa}_{CU})$, the differences being small when $n \geq 50$, but somewhat higher than in the previous case. In general, $\overline{\hat{V}}(\hat{\kappa}_C)$ is closer to $\hat{V}_E(\hat{\kappa}_C)$ than $\overline{\hat{V}}(\hat{\kappa}_{CU})$ is to $\hat{V}_E(\hat{\kappa}_{CU})$.

6 Assessment of the difference between each pair of estimators

The objective of this section is to assess the difference $\Delta_{XU} = |\hat{\kappa}_{XU} - \hat{\kappa}_X|$, when $\hat{\kappa}_X$ is any of the traditional estimators. In general, these differences are only appreciable with small samples, so that it is of interest to determine from what value of n onwards is it practically indifferent to calculate $\hat{\kappa}_{XU}$ or $\hat{\kappa}_X$.

For $\hat{\kappa}_{CU}$, in which $\hat{\kappa}_{CU} \geq \hat{\kappa}_C$, through expression (4), $\Delta_{CU} = \hat{\kappa}_C(1 - \hat{\kappa}_C)/\{(n - 1) + \hat{\kappa}_C\}$. Its maximum value in $\hat{\kappa}_C \geq 0$ is reached in $\hat{\kappa}_C = (n-1)^{0.5}/\{n^{0.5} + (n-1)^{0.5}\}$ and is $\{n^{0.5} + (n-1)^{0.5}\}^{-2}$. Therefore, $\Delta_{CU} < 0.01$ (or 0.02) when $n > 50$ (or 17). The conclusion is also valid for Δ_{HU} and $\Delta_{LU} = |\hat{\rho}_{LU} - \hat{\rho}_L|$, since $\hat{\kappa}_{HU}$ and $\hat{\rho}_{LU}$ have the same form as $\hat{\kappa}_{CU}$.

For $\hat{\kappa}_{SU}$, in which $\hat{\kappa}_{SU} \geq \hat{\kappa}_S$, $\Delta_{SU} = (1 - \hat{\kappa}_S^2)/\{(2n-1) + \hat{\kappa}_S\}$ through expression (12). Its maximum value in $\hat{\kappa}_S \geq 0$ is reached in $\hat{\kappa}_S = 0$ and is $1/(2n-1)$. Therefore, $\Delta_{SU} < 0.01$ (or 0.02) when $n > 100$ (or 33). The conclusion is also valid for Δ_{F2U}, since $\hat{\kappa}_{F2U}$ has the same form as $\hat{\kappa}_{FU}$. The case of $\hat{\kappa}_{KU}$ for $R = 2$ – last expression of Eq. (15) – provides a maximum for Δ_{KU} of $1/2n$ and leads to the same conclusion as above. The conclusion is also maintained for $\hat{\kappa}_{KU}$ in $R > 2$ and $\hat{\kappa}_{K2U}$, since they have the same form as $\hat{\kappa}_{KU}$ for $R = 2$.

The case of $\hat{\kappa}_{FU}$, in which $\hat{\kappa}_{FU} \geq \hat{\kappa}_F$, is somewhat more complex. Through expression (33), $\Delta_{FU} = (1 - \hat{\kappa}_F)\{R - (R-1)(1 - \hat{\kappa}_F)\}/\{Rn - (R-1)(1 - \hat{\kappa}_F)\}$. Its

Table 5 Results of the 10,000 simulations performed for the variances of two estimators of the Cohen *kappa* coefficient

K	n	κ_C	Classic $\hat{\kappa}_C$ estimator		New $\hat{\kappa}_{CU}$ estimator	
			$\hat{V}_E(\hat{\kappa}_C)$	$\overline{\hat{V}}(\hat{\kappa}_C)$	$\hat{V}_E(\hat{\kappa}_{CU})$	$\overline{\hat{V}}(\hat{\kappa}_{CU})$
2	10	0.38	0.0680	0.0711	0.0727	0.0633
		0.80	0.0234	0.0482	0.0226	0.0325
	20	0.41	0.0394	0.0367	0.0403	0.0357
		0.80	0.0153	0.0209	0.0147	0.0167
	50	0.39	0.0162	0.0158	0.0164	0.0157
		0.80	0.0070	0.0072	0.0069	0.0065
	100	0.41	0.0079	0.0078	0.0079	0.0078
		0.79	0.0043	0.0044	0.0043	0.0042
3	10	0.41	0.0451	0.0383	0.0474	0.0344
		0.83	0.0256	0.0347	0.0241	0.0221
	20	0.42	0.0230	0.0214	0.0234	0.0202
		0.77	0.0135	0.0144	0.0129	0.0113
	50	0.40	0.0117	0.0113	0.0118	0.0111
		0.79	0.0055	0.0054	0.0053	0.0049
	100	0.38	0.0057	0.0057	0.0058	0.0056
		0.68	0.0081	0.0078	0.0080	0.0075
5	10	0.44	0.0320	0.0273	0.0329	0.0226
		0.85	0.0155	0.0208	0.0138	0.0112
	20	0.38	0.0181	0.0164	0.0185	0.0155
		0.81	0.0094	0.0103	0.0089	0.0077
	50	0.40	0.0077	0.0074	0.0077	0.0072
		0.80	0.0045	0.0042	0.0044	0.0038
	100	0.39	0.0036	0.0035	0.0036	0.0035
		0.80	0.0025	0.0024	0.0024	0.0023

(1) $\hat{V}_E(\hat{\kappa}_C)$ and $\hat{V}_E(\hat{\kappa}_{CU})$ are the "exact" variances, or sample variances of the 10,000 values obtained of $\hat{\kappa}_C$ or $\hat{\kappa}_{CU}$, respectively. (2) $\overline{\hat{V}}(\hat{\kappa}_C)$ and $\overline{\hat{V}}(\hat{\kappa}_{CU})$ are the averages of the 10,000 estimated variances $\hat{V}(\hat{\kappa}_C)$ or $\hat{V}(\hat{\kappa}_{CU})$, respectively.

maximum value in $\hat{\kappa}_F \geq 0$ is reached in $\hat{\kappa}_F = \{(R-1)(n-1)^{0.5} - n^{0.5}\}/[(R-1)\{n^{0.5} + (n-1)^{0.5}\}]$ and is $\{R/(R-1)\} \times \{n^{0.5} + (n-1)^{0.5}\}^{-2}$. Note that for $R = 2$ this value is double that which is obtained for $\hat{\kappa}_{CU}$. Therefore, if we require that $\Delta_{FU} < 0.01$ (or 0.02), the value of n depends on the value of R. For example: $n > 100$ (or 33) for $R = 2$, $n > 75$ (or 25) for $R = 3$, $n > 63$ (or 21) for $R = 5$, and $n > 56$ (or 19) for $R = 10$. Moreover, Δ_{FU} is a decreasing function in R, taking its extreme values $\hat{\kappa}_F(1-\hat{\kappa}_F)/\{(n-1)+\hat{\kappa}_F\}$ in $R = \infty$, and $(1-\hat{\kappa}_F^2)/\{(2n-1)+\hat{\kappa}_F\}$ in $R = 2$. As those expressions have the same form as Δ_{CU} and Δ_{SU} respectively, then the precise minimum values of n for this case are an intermediate value from among the pairs of values indicated for those two cases. This is compatible with the numerical results above.

The case of $\hat{\kappa}_{GU}$, in which $\hat{\kappa}_{GU} \le \hat{\kappa}_G$, is much more complex since its values Δ_{GU} also depend on $\hat{I}_e$ because of expression (41). In the most simple situation -the unweighted case-, it can be demonstrated that $\Delta_{GU} \le \{R/(R-1)\}/\{m^{0.5} + (m-1)^{0.5}\}^{-2}$, with $m = (n-1)(K-1)$, an expression that depends on n, R and K; the level is also valid for the weighted case, although it is conservative. Therefore, if $R = 2$ and we require that $\Delta_{GU} < 0.01$ (or 0.02), the value of n depends on the value of K. For example: $n > 101$ (or 34) for $K = 2$, $n > 51$ (or 17) for $K = 3$, and $n > 26$ (or 9) for $K = 5$. The conclusion is also valid for Δ_{G2U}, since $\hat{\kappa}_{G2U}$ has the same form as $\hat{\kappa}_{GU}$.

The previous formulas provide values which are compatible with the results of Tables 1, 2, 3 and 4. Excluding the Gwet estimators and adopting the criterion that we want to guarantee that $\Delta_{XU} < 0.02$ (0.01), the overall conclusion is that we should use the current estimators at least when $n \le 17$ (50) in the case of $\hat{\kappa}_{CU}$ and $\hat{\kappa}_{HU}$, or when $n \le 33$ (100) in the rest of the cases.

7 Conclusions

There are different types of *kappa* coefficients which measure the experimental degree of agreement between R raters. In this article, we have focused on Cohen's *kappa* (Cohen 1960, 1968), Scott's *pi* (Scott 1955), Gwet's *AC1/2* (Gwet 2008) and Krippendorf's *alpha* coefficients (Krippendorf 1970, 2004), whether weighted or not, for $R = 2$, and in its *pairwise* type extensions, Hubert's *kappa* (Hubert 1977; Conger 1980), Fleiss's *kappa* (Fleiss 1971), Gwet's *AC1/2* and Krippendorf's *alpha* coefficients, for $R > 2$. In this last case ($R > 2$), the four measures of agreement use the *pairwise* method to determine the observed index of agreements I_o, but only the measure of Hubert's *kappa* also uses the *pairwise* method to determine the expected index of agreements I_e. We have called the measures obtained in this last way as *two-pairwise* measures. We have also defined the other three coefficients (Fleiss's *kappa*, Gwet's *AC1/2* and Krippendorf's *alpha*) from the *two-pairwise* point of view, thus obtaining the three Fleiss's *kappas two-pairwise*, etc. That is why the number of agreement coefficients that have been defined is eleven.

The article demonstrates that all of the traditional estimators of the eleven coefficients are based on biased estimators of I_e. The alternative is to use the eleven new proposed coefficients, which are based on unbiased estimators of I_e. In all cases, the traditional estimators are smaller than or equal to the new ones, except for the case of Gwet, where it is the other way around. The simulations carried out for the case of $R = 2$ show that the classic estimators $\hat{\kappa}_X$ usually underestimate κ_X (or overestimate, in the case of $X = G$), while the new estimators $\hat{\kappa}_{XU}$ are usually approximately unbiased. Additionally, it is verified that the new estimators $\hat{\kappa}_{XU}$ may be unnecessary when the sample size n is sufficiently large (e.g. $n > 30$). The article also provides the variances of the new estimators as a function of the variances of the classic estimators, except in the case of the Gwet estimators.

One question of interest is the relation between the coefficients and estimators of Hubert's *kappa* (Hubert 1977; Conger 1980), the CCC (Lin 1989, 2000), and the ICC (Shrout and Fleiss 1979; Carrasco and Jover 2003), when in the first case quadratic weights are used. In the article it has been justified that: (1) $\kappa_H = \rho_L = \rho_{I2}$, with respect to the coefficients; (2) $\hat{\kappa}_H = \hat{\rho}_L$, with respect to classical estimators based on biased estimators of the components of the coefficients; and (3) $\hat{\kappa}_{HU} = \hat{\rho}_{LU} = \hat{\rho}_{I2}$, with respect to classical ($\hat{\rho}_{LU}$ and $\hat{\rho}_{I2}$) or new ($\hat{\kappa}_{HU}$) estimators based on unbiased estimators of all components of the coefficients. These statements are true for $R \geq 2$, so that for $R = 2$ it is obtained that: $\kappa_C = \rho_L = \rho_{I2}$, $\hat{\kappa}_C = \hat{\rho}_L$, and $\hat{\kappa}_{CU} = \hat{\rho}_{LU} = \hat{\rho}_{I2}$.

Finally, the entire article has been developed for the general case in which the measures are defined based on any w_{ij} weights, thus avoiding a repetition of expressions and demonstrations. Nevertheless the non-weighted case ($w_{ij} = \delta_{ij}$) is very common. To make the text more reader friendly "Appendix 4" includes the eleven non-weighted coefficients mentioned in this article.

Acknowledgements This research was supported by the Ministry of Science and Innovation (Spain), Grant PID2021-126095NB-I00 funded by MCIN/AEI/https://doi.org/10.13039/501100011033 and by "ERDF A way of making Europe".

Funding Funding for open access publishing: Universidad de Granada/CBUA. Funding for open access charge: Universidad de Granada / CBUA. The authors are grateful to the reviewers for their invaluable comments.

Appendices

Appendix 1: Average values of some functions of parameters of the multinomial distribution and simplification of some expressions

In a multinomial distribution M$\{n; p_i\}$, it occurs that E$(\hat{p}_i) = p_i$, V$(\hat{p}_i) =$ E$(\hat{p}_i^2) -$ E$^2(\hat{p}_i) = p_i(1 - p_i)/n$ and Cov$(\hat{p}_i, \hat{p}_j) =$ E$(\hat{p}_i \hat{p}_j) -$ E$(\hat{p}_i) \times$ E$(\hat{p}_j) = -p_i p_j/n$ (if $i \neq j$). Therefore

$$E\left(\hat{p}_i \hat{p}_j\right) = \frac{(n-1) p_i p_j + \delta_{ij} p_i}{n}. \tag{45}$$

In the case of Sect. 2, applying the previous point to the distribution M$\{n; p_{ij}\}$ it is deduced that E$(\hat{p}_{i\cdot} \hat{p}_{\cdot j}) = E\left[\left(\sum_h \hat{p}_{ih}\right)\left(\sum_t \hat{p}_{tj}\right)\right] = \sum_h \sum_t E\left(\hat{p}_{ih} \hat{p}_{tj}\right) = \sum_h \sum_t \left\{(n-1) p_{ih} p_{tj} + \delta_{ti} \delta_{hj} p_{ij}\right\} / n$, where the last equality is due to expression (45), and h, $t = 1, 2, \ldots, K$. Operating it is obtained that E$(\hat{p}_{i\cdot} \hat{p}_{\cdot j}) = \{(n-1) p_{i\cdot} p_{\cdot j}$

$+ p_{ij}\}/n$, as in expression (2). In the same way, $\mathrm{E}(\hat{p}_{i\cdot}\hat{p}_{j\cdot}) = \sum_h \sum_t E\left(\hat{p}_{ih}\hat{p}_{jt}\right) = \sum_h \sum_t \left\{(n-1)p_{ih}p_{jt} + \delta_{ij}\delta_{ht}p_{ih}\right\}/n = \{(n-1)p_{i\cdot}p_{j\cdot} + \delta_{ij}p_{i\cdot}\}/n$ so that,

$$E\left(\hat{p}_{i\cdot}\hat{p}_{j\cdot}\right) = \frac{(n-1)p_{i\cdot}p_{j\cdot} + \delta_{ij}p_{i\cdot}}{n} \text{ and } \widehat{p_{i\cdot}p_{j\cdot}} = \frac{n\hat{p}_{i\cdot}\hat{p}_{j\cdot} - \delta_{ij}\hat{p}_{i\cdot}}{n-1}. \tag{46}$$

In a similar way, for $\hat{p}_{\cdot i}\hat{p}_{\cdot j}$. As $\hat{\pi}_i\hat{\pi}_j = \left(\hat{p}_{i\cdot}\hat{p}_{j\cdot} + \hat{p}_{\cdot i}\hat{p}_{\cdot j} + \hat{p}_{i\cdot}\hat{p}_{\cdot j} + \hat{p}_{\cdot i}\hat{p}_{j\cdot}\right)/4$ because of the expression (9) then, having applied the previous equalities, expression (10) is obtained. Finally, regarding the end of Sect. 2.5, through expression (18) it is deduced that $\hat{I}_{eU} - \hat{I}_e$ is proportional to $n\,\hat{I}_e - W(1 - \sum_i \hat{p}_{ii})/\{2(K-1)\} - (n-1)\hat{I}_e = \hat{I}_e - W(1 - \sum_i \hat{p}_{ii})/\{2(K-1)\}$ which, through expression (17), is also proportional to $1 + \sum_i \hat{p}_{ii} - 2\sum_i \hat{\pi}_i^2 = \sum_i \left\{\hat{\pi}_i + \hat{p}_{ii} - 2\hat{\pi}_i^2\right\}$. Taking into account the value of $\hat{\pi}_i$ expression (9) and operating, it is deduced that each term i of the previous expression is also proportional to $S_i(1 - S_i) + \hat{p}_{ii}\left(1 - \hat{p}_{ii}\right) + 2\,\hat{p}_{ii}(1 + S_i) \geq 0$, where $S_i = \hat{p}_{i\cdot} + \hat{p}_{\cdot i} - \hat{p}_{ii} \geq 0$. The conclusion is always that $\hat{I}_{eU} - \hat{I}_e \geq 0$.

In the case of Sect. 3, expression (46) adopts the form,

$$E\left(\hat{p}_{ir}\hat{p}_{jr}\right) = \frac{(n-1)p_{ir}p_{jr} + \delta_{ij}p_{ir}}{n} \text{ and } \widehat{p_{ir}p_{jr}} = \frac{n\hat{p}_{ir}\hat{p}_{jr} - \delta_{ij}\hat{p}_{ir}}{n-1}.$$

Let the value $I_o = \Sigma_r\Sigma_{r'\neq r}\Sigma_i\Sigma_j w_{ij}p_{ir,jr'}/\{R(R-1)\} = \Sigma_i\Sigma_j w_{ij}\Sigma_r\Sigma_{r'\neq r}p_{ir,jr'}$ defined in Sect. 3.1, the one we are trying to estimate. For a given subject s, the possible pairs of replies (i, j), with $i \neq j$, are $R_{is}R_{js}$, and the possible pairs of replies (i, i) are $R_{is}(R_{is} - 1)$, since the two raters must be different. Adding in s and dividing by n we obtain the estimations $\Sigma_r\Sigma_{r'\neq r}\ \hat{p}_{ir,\,jr'}$ and $\Sigma_r\Sigma_{r'\neq r}\ \hat{p}_{ir,\,ir'}$ of $\Sigma_r\Sigma_{r'\neq r}p_{ir,jr'}$ and $\Sigma_r\Sigma_{r'\neq r}p_{ir,ir'}$ respectively. Therefore, the estimation $\hat{I}_o$ of the value I_o of the second expression of the beginning of this paragraph will verify that $nR(R-1)\hat{I}_o = \Sigma_i\Sigma_{j\neq i}w_{ij}\Sigma_s R_{is}R_{js} + \Sigma_i w_{ii}\Sigma_s R_{is}(R_{is} - 1) = \Sigma_i\Sigma_j w_{ij}\Sigma_s R_{is}R_{js} - nR$, since $\Sigma_i\Sigma_s w_{ii}R_{is} = nR$ as $w_{ii} = 1$. This leads to the second expression of Eq. (24).

The value of I_e of Sect. 3.2 is given by $I_e = \Sigma_r\Sigma_{r'\neq r}I_e(r, r') = \Sigma_r\Sigma_{r'\neq r}\Sigma_i\Sigma_j w_{ij}p_{ir}p_{jr'} = \Sigma_i\Sigma_j w_{ij}\Sigma_r\Sigma_{r'\neq r}p_{ir}p_{jr'} = \Sigma_i\Sigma_j w_{ij}(p_{i+}p_{j+} - \Sigma_r p_{ir}p_{jr})$ since $\Sigma_r\Sigma_{r'\neq r}p_{ir}p_{jr'} = \Sigma_r\Sigma_{r'}p_{ir}p_{jr'} - \Sigma_r p_{ir}p_{jr} = \Sigma_r p_{ir}\Sigma_{r'}p_{jr'} - \Sigma_r p_{ir}p_{jr} = p_{i+}p_{j+} - \Sigma_r p_{ir}p_{jr}$. This leads to the second expression (25).

Regarding what is highlighted in the first paragraph of Sect. 3.3, $R^2E\left(\hat{\pi}_i\hat{\pi}_j\right) = E\left\{\sum_r\sum_{r'}\hat{p}_{ir}\hat{p}_{jr'}\right\} = E\left\{\sum_r\sum_{r'\neq r}\hat{p}_{ir}\hat{p}_{jr'} + \sum_r\hat{p}_{ir}\hat{p}_{jr}\right\}$. Through expressions (46) and (2) which are placed in the format of Sect. 3, $nR^2E\left(\hat{\pi}_i\hat{\pi}_j\right) = (n-1)\sum_r\sum_{r'\neq r}p_{ir}p_{jr'} + (n-1)\sum_r p_{ir}p_{jr} + \sum_r\sum_{r'\neq r}p_{ir,\,jr'} + \delta_{ij}\sum_r p_{ir}$ where the sum of the two terms is $(n-1)\sum_r\sum_{r'}p_{ir}p_{jr'} = (n-1)p_{i+}p_{j+} = (n-1)R^2\pi_i\pi_j$. Therefore, $\hat{\pi}_i\hat{\pi}_j$ is not an unbiased estimator of $\pi_i\pi_j$ since,

$$E\left(\hat{\pi}_i\hat{\pi}_j\right) = \frac{1}{n}\left[(n-1)\pi_i\pi_j + \frac{1}{R^2}\left\{\sum_r\sum_{r'\neq r}p_{ir,\,jr'} + \delta_{ij}\sum_r p_{ir}\right\}\right]. \tag{47}$$

As $E\left(\hat{I}_e\right) = \sum_i \sum_j w_{ij} E\left(\hat{\pi}_i \hat{\pi}_j\right) = n^{-1}[(n-1)I_e + R^{-2}\{\Sigma_i \Sigma_j w_{ij} \Sigma_r \Sigma_{r \neq r'} p_{ir,jr'}$ $+ \Sigma_i \Sigma_j w_{ij} \delta_{ij} \Sigma_r p_{ir}\}] = n^{-1}[(n-1)I_e + R^{-2}\{R(R-1)I_o + R\}]$, from here we deduce the first expression of (33).

Regarding what is highlighted in the second paragraph of Sect. 3.3, through expression (8) $4I_e(r, r') = \Sigma_i \Sigma_j w_{ij}(p_{ir} + p_{ir'})(p_{jr} + p_{jr'}) = \Sigma_i \Sigma_j w_{ij}(p_{ir}p_{jr} + p_{ir'}p_{jr'} + p_{ir}p_{jr'} + p_{ir'}p_{jr})$ and, through expression (23), $4R(R-1)I_e = \Sigma_i \Sigma_j w_{ij}[\Sigma_r \Sigma_{r' \neq r} p_{ir}p_{jr} + \Sigma_r \Sigma_{r' \neq r} p_{ir'}p_{jr'} + \Sigma_r \Sigma_{r' \neq r} p_{ir}p_{jr'} + \Sigma_r \Sigma_{r' \neq r} p_{ir'}p_{jr}]$. As $\Sigma_r \Sigma_{r' \neq r} p_{ir}p_{jr} + \Sigma_r \Sigma_{r' \neq r} p_{ir'}p_{jr'} = 2(R-1)\Sigma_r p_{ir}p_{jr}$ and $\Sigma_r \Sigma_{r' \neq r} p_{ir}p_{jr'} + \Sigma_r \Sigma_{r' \neq r} p_{ir'}p_{jr} = 2p_{i+}p_{j+} - 2\Sigma_r p_{ir}p_{jr}$, then expression (35) is deduced.

Regarding the first paragraph of Sect. 3.5, expression (47) for $i = j$ is,

$$E\left(\hat{\pi}_i^2\right) = \frac{1}{n}\left[(n-1)\pi_i^2 + \frac{1}{R^2}\left\{\sum_r \sum_{r' \neq r} p_{ir,ir'} + \sum_r p_{ir}\right\}\right]. \tag{48}$$

Therefore, the unbiased estimator of π_i^2 is $\widehat{\pi_i^2} = (n-1)^{-1}[n\ \hat{\pi}_i^2 - \left\{\sum_r \sum_{r' \neq r} \hat{p}_{ir,ir'} + \sum_r \hat{p}_{ir}\right\} \Big/ R^2]$ and that of $\sum_i \pi_i^2$ will be $\sum_i \widehat{\pi_i^2} = (n-1)^{-1}[n \sum_i \hat{\pi}_i^2 - \left\{\sum_i \sum_r \sum_{r' \neq r} \hat{p}_{ir,ir'} + \sum_i \sum_r \hat{p}_{ir}\right\} \Big/ R^2]$. In this last expression, $\sum_i \hat{\pi}_i^2 = R^{-2}\{1 - K(K-1)\hat{I}_e/W\}$ through expression (39), $\sum_i \sum_r \hat{p}_{ir} = R$ since $\sum_i \hat{p}_{ir} = 1$, and $\sum_i \sum_r \sum_{r' \neq r} \hat{p}_{ir,ir'} = R(R-1)\ \hat{I}_{oN}$, where $\hat{I}_{oN}$ is obtained from the second expression of Eq. (22) applied to the non-weighted case of $\omega_{ij} = \delta_{ij}$. Substituting all of these values in $W(1 - \sum_i \widehat{\pi_i^2})/\{K(K-1)\}$ we obtain the value of $\hat{I}_{eU}$ of expression (40). Regarding the statement that $\hat{I}_{eU} - \hat{I}_e \geq 0$ one must take into account that $\hat{I}_{eU} - \hat{I}_e$ is proportional to $\hat{I}_e - A = \hat{I}_e - W(R-1)(1 - \hat{I}_{oN})/\{RK(K-1)\}$; substituting in this expression the estimators $\hat{I}_e$ and $\hat{I}_{oN}$ with their values from the last expressions of Eq. (39) and Eq. (40) respectively, it is obtained that $\hat{I}_{eU} - \hat{I}_e$ is proportional to $\sum_i \sum_s R_{is}^2 - \sum_i R_{i+}^2 \big/ n = \Sigma_i \Sigma_s (R_{is} - \overline{R}_i)^2 \geq 0$, where $\overline{R}_i = \Sigma_s R_{is}/n$.

Regarding what is highlighted in the second paragraph of Sect. 3.5, through expression (16) $\{K(K-1)/W\}I_e(r, r') = 1 - \Sigma_i (p_{ir} + p_{ir'})^2/4 = 1 - \Sigma_i (p_{ir}^2 + p_{ir'}^2 + 2p_{ir}p_{ir'})/4$. But through expression (23), $\{K(K-1)/W\}I_e = 1 - \Sigma_i[2(R-1)\Sigma_r p_{ir}^2 + 2p_{i+} - 2\Sigma_r\ p_{ir}^2]/\{4R(R-1)\}$; this leads to the expression (42). Finally, to demonstrate that in the *two-pairwise* case it also occurs that $\hat{I}_{eU} - \hat{I}_e \geq 0$, one must take into that through expression (44) $\hat{I}_{eU} - \hat{I}_e$ is proportional to $\hat{I}_e - X_N = \hat{I}_e - W(1 - \hat{I}_{oN})/\{2\ K(K-1)\}$. Substituting in this expression the estimators $\hat{I}_e$ and $\hat{I}_{oN}$ through its values of the last expressions of expressions (43) and (40) respectively, it is obtained that $\hat{I}_{eU} - \hat{I}_e$ is proportional to $nR(R-2) + \Sigma_i \Sigma_s\ R_{is}^2 - \Sigma_i\ n_{i+}^2/n - (R-2)\Sigma_i \Sigma_r\ n_{ir}^2/2 = \Sigma_i \Sigma_s (R_{is} - \overline{R}_i)^2 + (R-2)\Sigma_i \Sigma_r n_{ir}(n - n_{ir})/n \geq 0$.

As stated previously, all of the above is valid if there is only one multinomial sample. Let us suppose that $R = 2$, that the rater in the rows is a standard one and that the frequencies O_{ij} are obtained from K multinomial distributions $\{O_{i\cdot}; p_1, p_2, \ldots, p_K\}$, with $\Sigma p_i = 1$. Now $\hat{I}_e = \sum_i \sum_j w_{ij} O_{i\cdot} \hat{p}_j \Big/ n = \sum_i \sum_j w_{ij} O_{i\cdot} O_{\cdot j} \Big/ n^2$ is an unbiased estimator of $I_e = \Sigma_i \Sigma_j w_{ij} O_{i\cdot} p_j / n$, since $\mathrm{E}(\hat{p}_j) = p_j$.

Appendix 2: Variances of the new estimators of *kappa*

From hereon it is assumed that new estimators of *kappa* are approximately unbiased, since they are based on unbiased estimators of I_o and I_e. In the case of Sect. 2, from expression (4) it is deduced that $\hat{\kappa}_C = (n-1)\hat{\kappa}_{CU}/(n-\hat{\kappa}_{CU})$. Therefore $\mathrm{d}\,\hat{\kappa}_C/\mathrm{d}\,\hat{\kappa}_{CU} = n(n-1)/(n-\hat{\kappa}_{CU})^2$, whose value in $\mathrm{E}(\hat{\kappa}_{CU}) \approx \kappa_C$ is $n(n-1)/(n-\kappa_C)^2$ and, through the delta method, $\mathrm{V}(\hat{\kappa}_C) = n^2(n-1)^2\ \mathrm{V}(\hat{\kappa}_{CU})/(n-\kappa_C)^4$. This leads to expression (5). In a similar way, from expression (12) it is deduced that $\hat{\kappa}_S = \{(2n-1)\ \hat{\kappa}_{SU} - 1\}/(2n-1-\hat{\kappa}_{SU})$. Therefore, $\mathrm{d}\,\hat{\kappa}_S/\mathrm{d}\,\hat{\kappa}_{SU} = 4n(n-1)/(2n-1-\hat{\kappa}_{SU})^2$, whose value in $\mathrm{E}(\hat{\kappa}_{SU}) \approx \kappa_S$ is $4n(n-1)/(2n-1-\kappa_S)^2$ and $\mathrm{V}(\hat{\kappa}_S) = 16n^2(n-1)^2\ \mathrm{V}(\hat{\kappa}_{SU})/(2n-1-\kappa_S)^4$. This leads to expression (13). In the case of Sect. 3.3, from the second expression of Eq (33) it is deduced that $\hat{\kappa}_F = \{(nR-R+1)\hat{\kappa}_{FU} - 1\}/\{(nR-1)-(R-1)\hat{\kappa}_{FU})\}$. Therefore $\mathrm{d}\,\hat{\kappa}_F/\mathrm{d}\,\hat{\kappa}_{FU} = R^2 n(n-1)/\{(Rn-1)-(R-1)\hat{\kappa}_{FU}\}^2$, whose value in $\mathrm{E}(\hat{\kappa}_{FU}) \approx \kappa_F$ is $R^2 n(n-1)/\{(nR-1)-(R-1)\kappa_F\}^2$ and $\mathrm{V}(\hat{\kappa}_F) = R^4 n^2(n-1)^2\ \mathrm{V}(\hat{\kappa}_{FU})/\{(nR-1)-(R-1)\kappa_F\}^4$. This leads to expression (34). In a similar way with $\mathrm{V}(\hat{\kappa}_{KU})$ and $\mathrm{V}(\hat{\kappa}_{F2U})$.

Appendix 3: Justification of the equality $\hat{\rho}_{LU} = \hat{\rho}_{I2}$ $\hat{\rho}_L = \hat{\rho}_{I2S}$ and its simplified formula

Using the notation of the end of Sect. 3.2, the expression $\hat{\rho}_{LU}$ of (28) is equivalent to this one, where $\overline{x}_{\cdot r} = x_{\cdot r}/n$,

$$\hat{\rho}_{LU} = \frac{2n\sum_r \sum_{r'\neq r} s_{rr'}}{2(R-1)(n-1)\sum_r s_r^2 + n\sum_r \sum_{r'} (\overline{x}_{\cdot r} - \overline{x}_{\cdot r'})^2 + 2\sum_r \sum_{r'\neq r} s_{rr'}}. \quad (49)$$

As $s_{rr'} = \Sigma_s(x_{sr} - \overline{x}_{\cdot r})(x_{sr'} - \overline{x}_{\cdot r'})/(n-1) = \{\Sigma_s x_{sr}x_{sr'} - x_{\cdot r}x_{\cdot r'}/n\}/(n-1)$, $s_r^2 = \Sigma_s(x_{sr} - \overline{x}_{\cdot r})^2/(n-1) = (\Sigma_s\, x_{sr}^2 - x_{\cdot r}^2/n)/(n-1)$, and $(\overline{x}_{\cdot r} - \overline{x}_{\cdot r'})^2 = (x_{\cdot r} - x_{\cdot r'})^2/n^2$, then.

$$\begin{aligned}
&\sum_r \sum_{r'\neq r} s_{rr'} = \left(n\sum_s x_{s\cdot}^2 + \sum_r x_{\cdot r}^2 - n\sum_s \sum_r x_{sr}^2 - x_{\cdot\cdot}^2\right)\Big/\{n(n-1)\},\\
&\sum_r s_r^2 = \left(n\sum_s \sum_r x_{sr}^2 - \sum_r x_{\cdot r}^2\right)\Big/\{n(n-1)\}, \quad \text{and}\\
&\sum_r \sum_{r'} (\overline{x}_{\cdot r} - \overline{x}_{\cdot r'})^2 = 2\left(R\sum_r x_{\cdot r}^2 - x_{\cdot\cdot}^2\right)\Big/ n^2.
\end{aligned}$$

Substituting the expression (49) it is obtained the last expression of Eq. (29). Similarly the expression (27) of $\hat{\rho}_L$ leads to the last expression of Eq. (30).

On the other hand, the estimator $\hat{\rho}_{I2}$ of ρ_{I2} − which is based on the unbiased estimators of its components − is the ICC(2, 1) of Shrout and Fleiss (1979)

$$\hat{\rho}_{I2} = \frac{n \times (MSS - MSE)}{n \times MSS + R \times MSR + (nR - n - R) \times MSE}, \tag{50}$$

where $MSS = SSS/(n-1)$, $MSR = SSR/(R-1)$, and $MSE = SSE/\{(n-1)(R-1)\}$ (or SSS, SSE, and SSD) denote the mean squares (or sum of squares) for subjects, raters, and error (residual) in the analysis of variance, respectively. In addition, $SSE = SST - SSS - SSR$, with SST the sum of squares total. As,

$$SSS = R\sum_s (\overline{x}_{s\cdot} - \overline{x}_{\cdot\cdot})^2 = \frac{1}{R}\left\{\sum_s x_{s\cdot}^2 - \frac{x_{\cdot\cdot}^2}{n}\right\},$$

$$SSR = n\sum_r (\overline{x}_{\cdot r} - \overline{x}_{\cdot\cdot})^2 = \frac{1}{n}\left\{\sum_r x_{\cdot r}^2 - \frac{x_{\cdot\cdot}^2}{R}\right\}, \text{ and}$$

$$SST = \sum_s\sum_r (x_{sr} - \overline{x}_{\cdot\cdot})^2 = \sum_s\sum_r x_{sr}^2 - \frac{x_{\cdot\cdot}^2}{nR}$$

then, substituting in the expression (50) it is obtained again the expression (29). Therefore $\hat{\rho}_{LU} = \hat{\rho}_{I2}$.

Appendix 4: Classic non-weighted *kappa* coefficients

We will now provide the values necessary to define any non-weighted coefficient $\kappa = (I_o - I_e)/(1 - I_e)$, and calculate the value of its classic estimator $\hat{\kappa} = \left(\hat{I}_o - \hat{I}_e\right)\Big/\left(1 - \hat{I}_e\right)$. The new estimator $\hat{\kappa}_U$ is obtained with the same formulas from the text of the article.

When $R = 2$ all of the *kappa* coefficients are based on $I_o = \Sigma p_{ii}$ and $\hat{I}_o = \sum_i \hat{p}_{ii} = \sum_i O_{ii}\big/n$. The actual and estimated values of I_e in each coefficient are:

(a) κ_C and $\hat{\kappa}_C$ (Cohen's *kappa*): $I_e = \Sigma_i p_{i\cdot}p_{\cdot i}$ and $\hat{I}_e = \sum_i \hat{p}_{i\cdot}\hat{p}_{\cdot i} = \sum_i O_{i\cdot}O_{\cdot i}\big/n^2$.

(b) κ_S and $\hat{\kappa}_S$ (Scott's *pi*): $I_e = \sum_i \pi_i^2$ where $\pi_i = (p_{i\cdot} + p_{\cdot i})/2$ and $\hat{I}_e = \sum_i \hat{\pi}_i^2$ where $\hat{\pi}_i = \left(\hat{p}_{i\cdot} + \hat{p}_{\cdot i}\right)\big/2 = (O_{i\cdot} + O_{\cdot i})\big/2n$.

(c) $\hat{\kappa}_K$ (Krippendorf's *alpha*) which estimates κ_S: $\hat{I}_e = \sum_i \hat{\pi}_i^2$, with $\hat{\pi}_i$ as in (b), but $\hat{I}_o$ is special: $\hat{I}_o = \left\{(2n-1)\sum_i \hat{p}_{ii} + 1\right\}\big/2n = \left\{(2n-1)\sum_i O_{ii} + n\right\}\big/2n^2$.

(d) κ_G and $\hat{\kappa}_G$ (Gwet's *AC1*): $I_e = \Sigma_i \pi_i(1 - \pi_i)/(K - 1)$ and $\hat{I}_e = \sum_i \hat{\pi}_i\left(1 - \hat{\pi}_i\right)\big/(K-1)$, with $\hat{\pi}_i$ as in case (b). Note that $\hat{I}_e = \left\{1 - \sum_i \hat{\pi}_i^2\right\}\big/(K-1)$, where $\sum_i \hat{\pi}_i^2$ is the value of $\hat{I}_e$ in (b). In this case, the formula of $\hat{\kappa}_{GU}$ does have a particular expression:

$$\hat{\kappa}_{GU} = \frac{(n-1)\hat{\kappa}_G + Y_N}{(n-1) + Y_N} \quad \text{where} \quad Y_N = \frac{1-\hat{\kappa}_G}{2(K-1)} - \frac{\hat{I}_e}{1-\hat{I}_e}$$

When $R \geq 2$ all of the *kappa* non-weighted coefficients are based on $I_o = \Sigma_r \Sigma_{r' \neq r} \Sigma_i p_{ir,ir'} / \{R(R-1)\}$ and $\hat{I}_o = \{\sum_i \sum_s R_{is}^2 - nR\}/\{nR(R-1)\}$. The actual and estimated values of I_e are:

(A) κ_H and $\hat{\kappa}_H$ (Hubert's *kappa*): $I_e = \sum_i \left\{p_{i+}^2 - \sum_r p_{ir}^2\right\}/\{R(R-1)\}$ and $\hat{I}_e = \sum_i \left\{n_{i+}^2 - \sum_r n_{ir}^2\right\}/\ \{n^2 R(R-1)\}$.

(B) κ_F and $\hat{\kappa}_F$ (Fleiss's *kappa*): $I_e = \sum_i p_{i+}^2/R^2$ and $\hat{I}_e = \sum_i R_{i+}^2/(nR)^2$.

(C) κ_{F2} and $\hat{\kappa}_{F2}$ (Fleiss's *kappa two-pairwise*): $I_e = [(R-2)\sum_i \sum_r p_{ir}^2 + \sum_i p_{i+}^2]/[2R(R-1)]$ and $\hat{I}_e = [(R-2)\sum_i \sum_r n_{ir}^2 + \sum_i n_{i+}^2]/[2n^2 R(R-1)]$.

(D) $\hat{\kappa}_K$ (Krippendorf's *alpha*) which estimates κ_F: $\hat{I}_e = \sum_i R_{i+}^2 \big/ (nR)^2$, but $\hat{I}_o$ is special: $\hat{I}_o = \{(2n-1)T + 1\}/2n$ where $T = \{\sum_i \sum_s R_{is}^2 - nR\}/\{nR(R-1)\}$.

(E) $\hat{\kappa}_{K2}$ (Krippendorf's *alpha two-pairwise*) which estimates κ_{F2}: $\hat{I}_o$ is the same as in paragraph (D) and $\hat{I}_e = \left\{(R-2)\sum_i \sum_r n_{ir}^2 + \sum_i n_{i+}^2\right\} \big/ \left\{2n^2 R(R-1)\right\}$.

(F) κ_G and $\hat{\kappa}_G$ (Gwet's *AC1*): $I_e = \left(1 - \sum_i p_{i+}^2 \big/ R^2\right)/(K-1)$ and $\hat{I}_e = \left\{1 - \sum_i R_{i+}^2 \big/ (nR)^2\right\}/(K-1)$. It can be observed that $\hat{\kappa}_G \geq \hat{\kappa}_F$, since $\hat{I}_e(\text{Gwet}) - \hat{I}_e(\text{Fleiss})$ is proportional to $K^{-1} - \sum_i \hat{\pi}_i^2 \leq 0$; the first statement because of expressions (39) and (32) respectively; the second one because $\sum_i \hat{\pi}_i^2$ reaches a minimum value of $1/K$ when $\hat{\pi}_i = 1/K$. In this case, the formula of $\hat{\kappa}_{GU}$ does have a particular expression:

$$\hat{\kappa}_{GU} = \frac{(n-1)\hat{\kappa}_G + B_N}{(n-1) + B_N} \quad \text{where} \quad B_N = \frac{(R-1)\left(1-\hat{\kappa}_G\right)}{R(K-1)} - \frac{\hat{I}_e}{1-\hat{I}_e}.$$

(G) κ_{G2} and $\hat{\kappa}_{G2}$ (Gwet's *AC1 two-pairwise*):

$$I_e = \frac{1}{K-1}\left[1 - \frac{1}{2R(R-1)}\left\{(R-2)\sum_i \sum_r p_{ir}^2 + \sum_i p_{i+}^2\right\}\right], \text{ and}$$

$$\hat{I}_e = \frac{1}{K-1}\left[1 - \frac{1}{2n^2 R(R-1)}\left\{(R-2)\sum_i \sum_r n_{ir}^2 + \sum_i n_{i+}^2\right\}\right].$$

In this case, the formula of $\hat{\kappa}_{G2U}$ does have a particular expression:

$$\hat{\kappa}_{G2U} = \frac{(n-1)\hat{\kappa}_{G2} + C_N}{(n-1) + C_N} \quad \text{where} \quad C_N = \frac{1-\hat{\kappa}_{G2}}{2(K-1)} - \frac{\hat{I}_e}{1-\hat{I}_e}.$$

References

Barnhart HX, Haber M, Song J (2002) Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 58:1020–1027. https://doi.org/10.1111/j.0006-341X.2002.01020.x

Carrasco JL, Jover LL (2003) Estimating the generalized concordance correlation coefficient through variance components. Biometrics 59:849–858

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20:37–46

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreemet or parcial credit. Psychol Bull 70:213–220

Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull 88:322–328. https://doi.org/10.1037/0033-2909.88.2.322

Davies M, Fleiss JL (1982) Measuring agreement for multinomial data. Biometrics 38:1047–1051

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Measur 33(3):613–619

Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. Psychol Bull 72:323–327. https://doi.org/10.1037/h0028106

Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, 3rd edn. Wiley, New York

Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 61(1):29–68

Gwet KL (2021a) Large-sample variance of Fleiss generalized kappa. Educ Psychol Measur 81(4):781–790. https://doi.org/10.1177/0013164420973080

Gwet KL (2021b) Handbook of inter-rater reliability. Volume 1: analysis of categorical ratings, 5th edn. Gaithersburg, MD, USA

Hubert L (1977) Kappa revisited. Psychol Bull 48(2):289–297

Janson S, Olsson U (2001) A measure of agreement for interval or nominal multivariate observations. Educ Psychol Measur 61(2):277–289

King TS, Chinchilli VM (2001) A generalized concordance correlation coefficient for continuous and categorical data. Statist Med 20(14):2131–2147. https://doi.org/10.1002/sim.845

Krippendorff K (1970) Estimating the reliability, systematic error, and random error of interval data. Educ Psychol Measur 30:61–70

Krippendorff K (2004) Measuring the reliability of qualitative text analysis data. Qual Quant 38:787–800

Landis JR, Koch GG (1975a) A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). Stat Neerl 29:101–123

Landis JR, Koch GG (1975b) A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). Stat Neerl 29:151–161

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Lin LI-K (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268

Lin LI-K (2000) A note of the concordance correlation coefficient. Letter to the Editor (Corrections). Biometrics 56:324–325

Martín Andrés A, Álvarez Hernández M (2020) Hubert's multi-rater kappa revisited. Br J Math Stat Psychol 73(1):1–22

Miettinen O, Nurminen M (1985) Comparative analysis of two rates. Stat Med 4:213–226. https://doi.org/10.1002/sim.4780040211

Schuster C, Smith DA (2005) Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. Psychometrika 70(1):135–146

Scott WA (1955) Reliability of content analysis: the case of nominal scale coding. Public Opin Q 19:321–325

Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428

Warrens MJ (2010) Inequalities between multi-rater kappas. Adv Data Anal Classif 4:271–286