Biometrics And Biostatistics Journal

Research Article

Sample Size Calculation for Agreement Studies on Qualitative Variables Using Cohen's Kappa: A Review and New Proposals

Bruno Mario Cesana[1]* and Paolo Antonelli[2]

[1]Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics, University of Milan, Italy

[2]Formerly, Contract Professor, Department of Molecular and Translational Medicine, University of Brescia, Italy

ARTICLE INFO

Received Date: July 07, 2021
Accepted Date: September 01, 2021
Published Date: September 04, 2021

KEYWORDS

Raters' agreement
Cohen's kappa statistic
Sample sizes calculation

Citation for this article: Bruno Mario Cesana and Paolo Antonelli. Sample Size Calculation for Agreement Studies on Qualitative Variables Using Cohen's Kappa: A Review and New Proposals. Biometrics And Biostatistics Journal. 2021; 3(1):115

Corresponding author:

Bruno Mario Cesana, Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics "Giulio A. Maccacaro", Faculty of Medicine and Surgery, University of Milan, Via Vanzetti 5, 20133 Milan, Italy, Tel: + 39-02503 20854;
Email: brnmrcesana@gmail.com; bruno.cesana@guest.unimi.it

ABSTRACT

Introduction: We have revised the steps of the sample sizes calculation for the biomedical research.

Then, focusing on the agreement studies on qualitative variables we have shown the derivation of the Cohen's kappa together with the factors influencing its value, leading to have low kappa values even in presence of a relevant observed agreement. We have revised the sample size calculation approaches for the Cohen's kappa proposed in the statistical literature.

Methods: We have calculated the sample sizes for agreement studies based on the Cohen's kappa statistics according to the main relevant literature proposals, such as those of Flack et al. [74] and of Donner et al. [92,94] Then, we have proposed a partial extension of the common correlation model (PCCM) for 2x2 contingency table to cxc square contingency tables with a common correlation coefficient of the cells on the principal diagonal, considered pair wise. Finally, we have conceived a full generalization with a common correlation coefficient model (FCCM) of all cells of the square contingency table.

Results: From our PCCM, we have obtained very similar maximum and minimum values of the kappa variance, leading to have sample sizes slightly different. Otherwise, from FCCM, it is obtained a unique contingency table and, consequently, a unique value of the kappa variance to be used under the null and the alternative hypothesis for sample sizes calculation leading to have only one value of the sample size. More relevantly, this latter sample size is within or equal to the sample sizes calculated by using the maximum and minimum value of the kappa variance under PCCM. In the case of 2x2 contingency tables, all our sample sizes proposed methods (SS-A&C-max, SS-A&C-min, and SS-A&C-full) gave equal sample sizes which are, in addition, equal to those calculated according to Flack et al. [74], being the sample sizes from Donner et al. [92,94] generally greater.

For 3x3 contingency tables, the sample sizes calculated according to our proposed models are lower than or equal to those calculated according to Flack et al. [74] and, also almost always, lower than those from Donner et al. [92,94] The same occurs for the 4x4 tables.

Discussion / conclusions: Sample sizes from our proposed models can be generally recommended since they are lower than or at maximum equal to those obtained by

the Flack et al.'s approach [74]. In addition, our proposed procedures give very similar maximum and minimum sample size values under PCCM and only one sample size value under FCCM. Furthermore, sample size proposed by Donner et al. [92,94] can be considered only in the case in which they are lower and the primary objective of the study is the "agreement or not" instead of a diversified level of agreement assessed through the weighted kappa. Finally, the fact that there is a unique kappa variance value under the FCCM, leading to sample sizes always within those calculated from the maximum and minimum values of the kappa variance under PCCM, allows to argue that this model actually refers to the population probability table.

INTRODUCTION

General considerations on sample size calculation

It is very well known that the question most frequently asked to a biostatistician is: "how many patients/subjects/experimental or observational units I have to enroll in a particular trial for demonstrating…?". This question could have a not so difficult answer, after having obtained the necessary information that the researcher is often surprised to have to add. More disarming for the biostatisticians is the question: "I have only X patients/subjects/experimental or observational units. Can I demonstrate X differences and X associations and so on, and so on…?".Actually, for obtaining a sensible response the biostatistician and the researcher need to arm themselves with a lot of patience and willingness to interact.

Indeed, it is widely recognized that biomedical research has to be adequately powered in order to have a high probability to achieve their goals. Particularly, in experimental research the sample size calculation is usually power-based on the statistical significance test of the primary endpoint. Of course, the need of having a high probability of rejecting a statistical null hypothesis of no difference, superiority (difference given by a superiority margin), non-inferiority (difference given by a non-inferiority margin), or equivalence (difference given by two, usually symmetric, equivalence margins) is well recognized and pursued by researchers.

Also the sample sizes calculation for observational studies (cross-sectional, case-control and cohorts) with two groups are traditionally based on the power of statistical tests for having a satisfactory enough probability of demonstrating a clinically relevant difference from the null value of 1 of the odds ratio (OR) or of the relative risk (RR) or of the hazard ratio (HR), for example.

Furthermore, cross-sectional studies can have sample sizes power-based calculated on a statistical test for demonstrating relationships between two qualitative variables such as the presence of a disease (yes or not) and the presence of a risk factor (yes or not). Of course, this association can be quantified, in addition to the difference between two proportions, by means of OR or of RR. Of course, the correlation coefficient (parametric or not) has to be used for assessing the relationship between quantitative variables. In addition, it has to be noted that if multivariable statistical procedures (linear regression, logistic regression, Cox's proportional regression, etc.) are used for obtaining the most parsimonious set of the "independent variables independently" associated with the dependent variable, the sample size has to be calculated accordingly.

However, sample sizes for observational research enrolling only one group of patients are frequently based on the precision of the estimate of the primary outcome. This occurs, particularly, in the case of observational research on administrative data with very huge mass of data.

Nonetheless, it has to be pointed out that also in one-group observational studies, the conclusion of rejecting a null hypothesis formulated on an "expected value", obtained from a careful search of the pertinent literature, makes the research more useful and scientifically more appealing instead (or in addition) of a statement about the precision of an estimate. An "expected value" can be the proportion of success to be tested by means of the binomial test, for example, or the mean of a quantitative variable to be tested by means of the paired Student's t test, for example.

Indeed, a sentence on the precision of an estimate, given by the confidence interval width, is a generic statement much less interesting to the clinicians/researchers than a sentence about a research hypothesis such as "the proportion of survival is more (lower) than an expected value" or such as "the mean blood pressure decrease is more (lower) than an expected value". It has to be stressed that the "expected value" has to be obtained from the mostly relevant pertinent literature (meta-analysis, particularly) and, consequently, is a relevant piece of scientific knowledge that can rightly to be taken as the parameter of statistical hypotheses.

Furthermore, it has to be pointed out that the required precision of an estimate, say the width of a 95% Confidence Interval (CI), is obtained in only about the 50% of the cases (as the probability of getting heads or tails by flipping a fair coin), unless appropriate statistical procedures are activated which take into account the "Confidence Interval Power", according, among others, to Beal [1].

A different approach of increasing the power-based sample size for having an adequate probability of obtaining a sample 95%CI width less than a "required width", in the case of testing a proportion against an expected value by means of the binomial test, has been proposed by Cesana et al. [2] and, in the case of testing two paired and unpaired proportions, by Cesana [3]. Jiroutek et al. [4] suggested, in the context of quantitative variables assumed Gaussianly distributed, a further extension of the sample size calculation based on considering jointly the power of the statistical test and the probability of obtaining a required 95% CI width, given the coverage. However, Jiroutek et al.'s approach [4],primarily based on the precision and, consequently, on very narrow 95% CI widths, requires too high sample sizes, not acceptable for ethical reasons as it has been pointed out by Cesana and Antonelli [5]. Indeed, for comparing two groups at a significance value of 0.05 (two-tailed) with a probability equal to 0.80 of jointly obtaining a required effect size of 1 and a required standardized 95% CI width of 1, given the coverage, is necessary a total of 74 subjects; otherwise, with the same significance and power, a total of 34 subjects are necessary for demonstrating an effect size of 1 for the statistical test which is the main objective in a controlled clinical trial (Jiroutek et al.'s Table 2, page 584 [4]). Furthermore, the above total of 74 becomes 268 if the required standardized CI width is of 0.5. Finally, an approach based on the two joint powers of the statistical test and of the 95%CI, given the coverage, and requiring lower and ethically acceptable sample sizes, has been proposed by Cesana and Antonelli [6] Indeed, Cesana and Antonelli's approach [6] starts from considering the 95% CI width obtained with the calculated power-based sample size. Then, if this 95% CI width is considered to be an adequate precision of the estimate, as often happens, the power-based sample size is iteratively increased for jointly obtaining an adequate test power (at least more than 0.80) and a 95% CI power (at least more than 0.75), given the coverage. Very importantly, this composite objective is usually obtained by increasing the power-based sample size of only a few units.

Finally, it has to be stressed that underpowered scientific research have to be avoided since it is unethical enrolling human beings (or animals, too) into a research without a satisfactory enough probability of obtaining some scientifically relevant results. However, also overpowered scientific research are, again, unethical since too many human beings (or animals, too) are exposed to experimental risks and too many resources are wasted. The wasting of resources is mainly relevant in the case of observational studies carried out, particularly, on administrative database with thousands of records. Let's remember the essential ingredients of the sample size calculation, already proposed, for example, by Cesana and Cavaliere [7].

Apart, the kind of the variable that is the main expression of the investigated phenomenon and, consequently, represents the primary outcome of the study on which procedure of the statistical testis determined, it has to be remembered: (i) the significance threshold (usually fixedat α= 0.05, two tailed); (ii) the power (usually fixed at 1 - β = 0.80, at least); (iii) the minimal clinically relevant effect to demonstrate in the context of a superiority trial (null hypothesis of no difference). It has to be noted that the "minimal clinically relevant effect" becomes the "maximal clinically not relevant effect" in the context of non-inferiority or equivalence studies.

Then, it is worthwhile to specify that "the effect" can be the difference between two proportions (complete response, survival, relapse, etc.), the difference between two changes of a quantitative variable (blood pressure, cardiac index, left ejection ventricular fraction – LEVF -, etc.) or the OR, or the RR, or the HR. Sometimes, it has to be remembered, the effect consists in a difference between two (or more) regression or correlation coefficients. An optimal estimate of the effect to be considered in the sample sizes calculation has to be appropriately obtained from an exhaustive search of the pertinent literature. Having retrieved some values, as usual, they have to be combined through a weighted mean, for example [7].

In the case of quantitative variables, an adequate estimate of the phenomenon variability must also be obtained. To this regards, it is quite easy to obtain cross-sectionally (at baseline and at the end of a trial, for example) some values of the standard deviation together with their pertinent means. However, if the researcher is interested on a difference between two (or more) means of the change (decrease / increase) from the start and the end of a treatment obtained in two groups (one treated with a new drug and the other treated with the standard drug, for example), the pertinent variability has to be expressed by the standard deviation of the change recorded on each patient. If the standard deviation of the change has not been reported in the paper, it has to be estimated by taking into account the correlation coefficient (generally, ranging from -1 to 1, but in this case, sensibly only from 0.3 to 0.8) between baseline and final values [7].

In some cases, a more refined pooling can be done, taking into account also the variability; indeed, sample sizes are not optimal weights for a mean calculation when the effect size index is a standardized difference mean [8].

It is warmly recommended to calculate the sample sizes for some possible scenarios leading to the construction of the so-called “power curves” with the power on the vertical axis and the sample size values on the horizontal axis for a fixed effect size (or the effect size values on the horizontal axis for a fixed sample size) in order to more effectively grasp the increasing power pattern as a function of the sample size (or effect size). It has to be noted that power curves can be obtained, among the free sample size calculation software, also by G*Power [9].

Effect size

According to Wilkinson [10], the effect sizes have to be always presented for the primary outcomes of a research. Furtherly, for variables with measurement unit practically meaningful, it is recommended to report the absolute measure of the difference together with the measure of the phenomenon variability without providing only the standardized effect size index such as the Cohen’s d [11], obtained by the ratio between the true difference and the true phenomenon variability (δ/σ). Indeed, even if only one number might seem more simpler, it is better that researchers get used to consider and to communicate the phenomenon variability, since the same effect size value is obtained by very different combination of the difference (δ) and of the variability (σ) or by quoting Length [12] “you'll choose the same sample size (n) regardless of the accuracy or reliability of your instrument, or the narrowness or diversity of your subjects.” However, it has also to be said that in the same biomedical research field, the variability can be within narrow limits that are very well known by expert researchers, leading to have a very precise idea of the suitable difference given by the effect size value, without the need of specifying both.

It is absolutely to be avoided, as Lipsey et al. [13] pointed out: “The widespread indiscriminate use of Cohen’s generic small, medium, and large effect size values to characterize effect sizes in domains to which his normative values do not apply is thus likewise inappropriate and misleading."However, also Cohen [11] cautioned to refer generally to the classes of “small”, “medium”, or “big” effect size by stating: "The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation.”

Apart from the inclusion of some other effect sizes classes such as “very small”, “very large”, and “huge” proposed by Sawilowsky [14], it has to be reported the absolutely shareable conclusion from Ellis [15]: “Researchers should interpret the substantive significance of their results by grounding them in a meaningful context or by quantifying their contribution to knowledge, and Cohen's effect size descriptions can be helpful as a starting point."

Finally, for sake of completeness, it has to be remembered that there are many kind of “effect size” as the Cohen’s books [11] report; particularly, “d” for the comparison of two population means through the unpaired Student’s t test, “f” for the comparison of more than two population means through the Analysis of Variance, “h” for differences between proportions trough the arcsine transformation, “f^2” for multiple regression and correlation analysis, etc. For a general and simple overview, readers are referred to the Wikipedia Web site of the effect size [16]. For further details the Ellis’ book [15] and Cohen’s books [11] are strongly recommended.

Power calculation "a priori" or "a posteriori"

A further debated point is whether the sample size calculation can be done also "a posteriori", after the data analysis of a study for justifying a not statistical significant result, for example. However, it has to be said that this it is a useless procedure as a result is not statistically significant because the sample size is not enough, given the effect size shown by the sample data or, in other words, the power of the statistical test is insufficient. Otherwise, a power analysis after a statistically significant result will always result in a high satisfactory power value. So, this approach has to be strongly discouraged since, in addition to be meaningless, it is not acceptable, for the statistical point of view, to assume the sample statistics as the population parameters under the alternative hypothesis, being the null hypothesis of no difference automatically formulated. A recent Editorial about this point, easily readable also for non-biostatisticians, is from Jiroutek and Turner [17].

A completely different situation is if we ask ourselves from a theoretical point of view which is the effect sizes that are possible to demonstrate at a satisfactory probability (power) by means of some pertinent statistical tests, given the sample size actually enrolled in a trial without an "a priori" power analysis or prematurely terminated. Indeed, the knowledge of the effect sizes allows readers to form their own opinion about the clinical relevance of the statistical significant (or not) results. It has also to be pointed out that only the statistical tests carried out on randomized groups have their nominal values of Type I error (a true null hypothesis is rejected) and of Type II error (a false null hypothesis is not rejected) from whose complement to 1 the statistical test power is calculated. Of course, the two above errors (probabilities) have to be considered in the context of the frequentist paradigm of the "infinite repetition" of the same trial under absolutely identical conditions.

Of course, also the agreement studies between two (or more) raters on qualitative variables (nominal or ordinal) and between two (or more) measurement methods of quantitative variables need to be adequately powered.

Considering the agreement between two or more raters on qualitative variables, it is immediate to refer to Cohen's kappa statistics, "in spite of its shortcomings and perceived paradoxes" as reported by Choudhary and Nagaraja [18] in Chapter 12 on "Categorical Data" of their book. However, it has to be immediately said that the above paradoxes can be explained by using the mathematical properties of the Cohen's kappa statistics.

Models for exploring agreement

It is very well know that it is often recommended to model the agreement and disagreement by means of "Conditional Logistic Regression Models" introduced firstly by Barlow [19], "Log-Linear Models", useful for exploring the agreement in the case of nominal and ordinal variables, firstly proposed by Tanner and Young [20], and "Generalized Linear Mixed-Effects Models" with their usual probit link function, adapted by Choudhary and Nagaraja [18] from Nelson and Edwards [21]. An excellent overview of modeling the agreement and disagreement between raters using categorical rating scales is in Agresti [22] and in both editions of Agresti's book [23]. Indeed, Agresti [23] (Chapter 11 Models for Matched Pairs (pages 413-454), paragraph 11.5.4 Kappa: A Summary Measure of Agreement, page 434- and 11.5.5 Weighted Kappa: Quantifying Disagreement) wrote as the last sentence: "models can provide more detailed description of the agreement and disagreement structure".

In addition, Von Eye and Mun [24], considered in Chapter 3 "Exploring Rater Agreement" of their book "The Configural Frequency Analysis (CFA)" proposed by Lienert and Krauth [25] and reconsidered particularly by von Eye [26,27] as "a method for exploring cross-classifications in order to answer the question if the absolute frequency of a cell is more than or less than expected according to a particular chance model".

Indeed, summarizing the characteristics of a probability table by a single measure could lead to a loss of information and, apart from the case in which kappa is close to 1, the joint distribution of raters' judgments is not adequately described. Furthermore, kappa values from two or more tables cannot be compared, leading to a further limitation of an approach based on Cohen's kappa. However, it has to be said that in the biomedical literature, the Cohen's kappa approach is the most used, and, therefore, we will focus on its sample size calculation.

Historical digression on agreement indices

Without any claim of being exhaustive, it has to be remembered an early paper from Guttman [28], the Goodman and Kruskal's λ [29], the Bennett et al.'s S [30], the Scott's π [31] as some early proposals of agreement indices, and, finally, the Cohen's κ [32]. Generally, these indices are used to summarize the cross classification of two qualitative variables with identical categories, as it has been pointed out, among others, by Brennan and Prediger [33], Zwick [34], Krippendorff [35], and DeMast [36]. Warrens [37] showed some inequalities and relationships of the above reported indices such as the fact that Bennett et al.'s "S" [30] is an upper bound of Cohen's κ [32] "if for the k×k table the permutation that orders the marginal probabilities from lowest to highest is the same for the rows and the columns (weak symmetry)".

In addition, Warrens [38] demonstrated the identity or the similarity with Cohen's kappa "of all the three agreement coefficients, originally introduced in 1914-1915 by the Italian statistician Corrado Gini" [39,40]. Moreover, it has been shown that the point estimates of Cohen's kappa and the two smaller Gini coefficients are very similar from the real data, leading, in practice, to the same conclusions about the degree of inter-rater agreement. Moreover, Gini's coefficients [39,40] were also considered by other authors such as Brennan and Prediger [33], Cohen [32], Janson and Vegelius [41], and Popping [42].

Furthermore, Ato et al. [43] showed a recent comparison among six rater agreement measures, including three historical indices such as the coefficient σ(S) suggested by Bennet's [30], Scott's π [31], and Cohen's κ [32]. In addition, Ato et al. [43] considered Gwet's γ [44] "as an expression of the classical descriptive approach", the Aickin's α [45] "as an expression of log-linear and mixture model approaches", and Martín and Femia's Δ measure [46] "representing a multiple-choice test". Accordingly to Ato et al. [43]: "π and κ descriptive measures present high levels of mean bias in presence of extreme values of prevalence and rater bias but small to null levels with moderate values". The best behavior was observed with Bennet [30] and Martín and Femia [46] agreement measures for all levels of prevalence."

Cohen's kappa model

Let's consider a square 3 by 3 (c by c or cxc) contingency table as the following with the judgments of the Rater A and Rater B about the membership or not to three categories simply coded as "1", "2", and "3". The judgments of the Rater A are on the columns and those of Rater B are on the rows.

	Rater A			Total
Rater B	"1"	"2"	"3"	
"1"	π_{11} ($\pi_{\iota\varphi}$)	π_{12}	π_{13}	$\pi_{1\bullet}$
"2"	π_{21}	π_{22}	π_{23}	$\pi_{2\bullet}$
"3"	π_{31}	π_{32}	π_{33}	$\pi_{3\bullet}$
Total	$\pi_{\bullet 1}$	$\pi_{\bullet 2}$	$\pi_{\bullet 3}$	$\pi_{\bullet\bullet}$

In the cells there are the true probabilities of the joint judgment of the two raters with the subscript "i" for the rows and the subscript "j" for the columns; in the marginal row and column the "dot" replaces the pertinent subscript on which the sum has been done. Of course, the Latin "p" letter replaces the Greek one(π) in the case of the observed proportions, as it has been shown, between brackets, in the cell$_{(1,1)}$. Obviously, the numeration of the rows starts from the top and the numeration of the columns starts from the left.

Considering that the true proportions (π_{ii}) on the diagonal of the above square contingency table represent the proportion of subjects in each category for which the two raters agreed on their assignment, the overall proportion of agreement is given by:

$\pi_0 = \sum_i^c \pi_{ii}$, being the observed agreement given by:

$$p_0 = \sum_i^c p_{ii} \,.$$

The true proportion of the agreement expected by chance or under the assumption of raters independence (π_e) is given by the sum of the products of the corresponding rows ($\pi_{i.}$) and columns ($\pi_{.i}$) marginal probabilities shown in the cells of the principal diagonal; so, we obtain:

$\pi_e = \sum_{i}^{c} \pi_{\bullet i} \bullet \pi_{i\bullet}$ with $\pi_{\bullet i}$ ($\pi_{i\bullet}$) the column (row) marginal and $p_e = \sum_{i}^{c} p_{\bullet i} \bullet p_{i\bullet}$ for the corresponding observed agreement by chance.

In addition, it has to be reported that when row and column marginal are equal (first row marginal equal to the first column marginal, etc.), the contingency table is defined as "symmetrical". Furthermore, we defined a symmetrical table as "uniform" when the rows and columns marginal are the same; then, a "uniform table" is by definition also symmetrical.

In 1960, Cohen [32] critiqued the use of the observed proportion of agreement due to its inability to account for the chance agreement "resulting from a random choice of the categories by two raters".

To adjust for this, Cohen [32] proposed an agreement measure "corrected for chance" or, otherwise defined, "the degree of raters' agreement in excess of chance", given by:

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_e} \quad (1)$$

The kappa estimate from the observed data is:

$$k_0 = \frac{p_0 - p_e}{1 - p_e}$$

The maximum value of kappa is 1, leading to a perfect agreement; of course, this case occurs in a 2x2 contingency table only when $\pi_{12} = \pi_{21} = 0$.

It has to be immediately said that for an asymmetrical contingency table, the maximum kappa value cannot be equal to 1. Indeed, the maximum proportion of the observed agreement is equal to the sum of the minimum values between the two corresponding row and column marginal; namely:

$$\pi_{MAX} = \sum_{i=1}^{c} \min(\pi_{\bullet i}, \pi_{i\bullet})$$

Consequently, the maximum kappa value reached depends on the marginal probabilities and it can be calculated as:

$$k_{MAX} = \frac{\pi_{max} - \pi_e}{1 - \pi_e}$$

Of course, the p Latin letter replaces the Greek letter in the case of the observed values.

Feinstein and Cicchetti [47] proposed to calculate Cohen's kappa by using the maximum possible value of the observed agreement (p_{max}) rather than 1, in the denominator of the k formula. Thus, it is possible to determine what proportion of the maximum observable agreement is actually achieved by the calculated k value in the context of an actual agreement study.

Furthermore, a kappa value of 0 indicates that the observed agreement is the same as that expected by chance, and the minimum value of kappa falls between -1.0 and 0.0. Landis and Koch [48] suggested that values of kappa above 0.60 show "good to excellent agreement between the two raters' scores", and values of 0.40 or less "show fair to poor agreement". Particularly, Landis and Koch's interpretation scale [48] of the kappa value is: <0.0 "No agreement"; 0.0 - 0.20 "Slight agreement"; 0.21 - 0.40 "Fair agreement"; 0.41 - 0.60 "Moderate agreement"; 0.61 - 0.80 "Substantial agreement"; 0.81 - 1.0 "Perfect agreement". So, since each class includes its upper limit, a "Substantial agreement" is for k values lower than 0.81.

It has to be noted that the above Landis and Koch's [48] definition of "Moderate agreement" for the class of 0.41 - 0.60, has been reported as "moderately strong agreement" on page 435 of both editions of Agresti's book [23], leading to an improved interpretation of the study results.

It is worthwhile to be remembered also that the Svanholm's scale [49], practically ignored in the scientific literature, set a higher threshold of 0.5 for the "Good agreement" (anything less being poor); this threshold is, in our opinion, more suitable, being in the middle between the no agreement (k = 0) and the perfect agreement (k = 1, a value reached only for symmetrical contingency tables).

Interestingly, the parameter π_e is also between 0 and 1; particularly, it is equal to 0 when, in the case of a 2 by 2 contingency table, π_{12}(or π_{21}) is equal to 1 with a perfect disagreement ($\pi_0 = 0$ and $k = 0$), and it is equal to 1 when π_{11} (or π_{22}) is equal to 1 with a perfect agreement ($\pi_0 = 1$, leading to obtain an indeterminate form for the value of k equal to 0/0).

In addition, Fleiss et al. [50] have stressed the fact that kappa values may be significantly different from zero but not large enough to satisfy the investigator's desire for a satisfactory agreement, as it has been also reported by Bakeman et al. [51].

As a further relevant point, it is expected that two raters are unbiased or that the two (or more) raters classify the same proportion of items (or subjects) into the different categories of the considered qualitative variable. However, it has to be stated that the agreement between two (or more) raters it is not required to be "perfect", but it is conceded that the agreement is less than perfect as long as the difference from the perfect agreement ($k=1$ or $k=k_{MAX}$) can be considered not clinically relevant. In addition, in the scientific context, it is not possible to demonstrate a null difference of H_0: $k=1$, as the expression of a perfect agreement. So, similarly to the non-inferiority clinical trials, the null and alternative hypotheses reverse their role into a hypothesis of a difference (the maximum threshold beyond which there is no agreement) and a hypothesis of no difference, respectively. Then, the maximum difference allowed for considering two drugs "clinically equivalent" becomes the maximum difference allowed for considering two (or more) raters "pragmatically concordant" or "in an almost perfect agreement".

About this point, we have to disagree, at least from a practical point of view, with the second part of the affirmation reported in Choudhary and Nagaraja's book [11] according to which "In certain situations, an assumption of no inter-rater bias appears reasonable. One is where the same rater rates a subject twice, as in a repeatability study, and the other is the case where we can draw a random sample of two raters from the available raters and use their ratings of the same subject. In both these cases, we can assume unbiasedness of two raters". Indeed, we think that if we refer to the population of all available raters, even with similar knowledge and training, in practice some discrepancies can be expected. In any case, unbiased raters lead to the condition of "exchangeability" from which the intra-class coefficient derives.

Formula 1, or its sampling counterpart expressed in terms of observed proportions, can be re-written by replacing p_0 and p_e with their p_{ij} and becomes, after some simplifications, the 12.5 formula on page 259 of the Choudhary and Nagaraja's book [11] reported here for readers' convenience:

$$k = \frac{2\left(p_{11}p_{22} - p_{12}p_{21}\right)}{p_{12} + p_{21} + 2\left(p_{11}p_{22} - p_{12}p_{21}\right)}$$

This formula is useful for considering that $k = 0$ only if the odds ratio of the 2 by 2 contingency table ($OR = p_{11}p_{22}/p_{12}p_{21}$) is equal to 1. In addition, it allows expressing Cohen's kappa in terms of a multinomial distribution from which the Maximum Likelihood (ML) estimates can be calculated and their standard errors can be obtained by using the well-known properties of the ML estimators.

In addition, Choudhary and Nagaraja's book [11], to which the interested reader is referred, shows: (i) a further formulation of Cohen's kappa that allows to explain the paradox of having a high observed agreement proportion with a low and even negative kappa value; (ii) a further parameterization useful for investigating the bounds of the kappa value in the case of a 2x2 contingency table; (iii) the expression of kappa as the Concordance Correlation Coefficient introduced by Lin [52]; and (iv) the expression of Cohen's kappa as an intraclass coefficient introduced in the case of multiple raters. Finally, an interesting formulation of Cohen's kappa is shown in the case of the "agreement with a gold standard" for which the well-known sensitivity and specificity probabilities, in addition to the prevalence parameter of the interesting condition, have to be considered. In this case, when the sensitivity and the specificity are known, the kappa value is a function of the prevalence and, accordingly, it is possible to calculate its maximum value.

For the above reported case of an observer's coding of events compared with a known, accurate standard, kappa would represent the reliability and the diagonal proportions of the agreement matrix would represent the rater's accuracy for individual codes [51]. However, according to Bakeman et al. [51] in the usual agreement study with fallible observers, "the tendency to make similar mistakes allows the interobserver kappa to be higher than the kappa estimating reliability, obtained by comparing a fallible to an infallible rater". Finally, if "both raters' errors were random and independent, then the interobserver kappa will represent a lower bound of the reliability". Many properties of Cohen's kappa, together with its relationship with other indices are reported in several Warrens papers downloadable from his Web site [53] to which the interested reader is addressed.

Cohen's kappa pitfalls

Multiple factors influence the kappa value, such as the rater's bias, the prevalence of the categories, and the number of the categories.

It has to be immediately stated that an important assumption underlying the use of the kappa is that errors associated with raters are independent, as it has been stressed, among others by Thompson and Walter [54], Shoukri [55], and Brennan and Prediger [33].

Cohen's kappa: raters bias effect: A "bias" between the raters A and B is present when the raters differ in assessment how often a condition occurs. When this occurs the marginal distributions of the raters are obviously unequal, leading to asymmetrical tables. So, the "bias" is the extent to which the raters disagree on the proportion of positive (or negative) cases, for a contingency table 2x2. In addition, the Bias Index (BI) is defined as the difference between the proportions of "Yes" of the two raters and it is estimated by the absolute difference between the marginal proportions of the first row and first column in the case of a 2x2 contingency table, or equivalently, by the difference between the proportions of the cells not lying on the principal diagonal of a 2x2 contingency table, usually defined as "b" and "c", respectively. Namely, the bias index is calculated as: $|b – c| / n$, being "n" the number of the paired subjects to be rated. The absolute value of BI has a minimum of 0 when the cell proportions shown above are equal or the two marginal proportions are equal; a maximum value of 1 is reached when or one or the other of these two proportions is equal to 1. The presence of the bias increases the kappa value since it decreases the agreement by chance (p_e).

Byrt et al. [56] defined a "bias-adjusted kappa (BAK)" by averaging the original values in the cells not lying on the principal diagonal [$m = (b + c)/2$]. However, although derived from a different way, BAK is in fact Scott's π [31] that differs from Cohen's kappa in terms of how p_e is calculated. The presence of the bias between raters gives rise to the second kappa paradox, according to Feinstein and Cicchetti [47], consisting in the fact that when the bias is large, kappa is greater than when the bias is small or absent. Otherwise, in contrast to prevalence, the effect of bias is greater when kappa is small than when it is large [56]. So, just as with prevalence (see after), the magnitude of kappa should be interpreted in the light of the bias index. In addition, for fixed observed agreement between the raters, Cohen's kappa penalizes raters with almost equal marginal compared to raters who produce different classification proportions, as it has been formally proved by Warrens[57]. Finally, it has to be remembered the possibility of calculating a theoretical value of Cohen's kappa starting from the row and column marginal and the accuracy/inaccuracy values attributed to the raters, according to Gardner's proposal [58] followed by Bakeman et al. [51] and by Bakeman [59]. Of course, in this case, the filling of the cells of the contingency table is based on the value of accuracy less than 1 for fallible raters.

Cohen's kappa: prevalence effect: Cohen's kappa is influenced by the prevalence of the categories of the considered variable. For example, an observed proportion of agreement of 0.8 gives a kappa of 0.6 when the row and the column marginal are equal to 0.5 (without a prevalence effect) and a value of 0.375 when the row and the column marginal are equal to 0.80 and 0.20, respectively. Byrt et al. [56] firstly defined a "Prevalence Index" (PI) as the difference between the probability of "Yes" and the probability of "No", estimated by the absolute difference between the proportions in the cells lying on the principal diagonal of a contingency table 2 by 2, usually indicated as "a" and "d", respectively. Consequently PI = $| a - d | / n$, where "n" is the number of paired ratings. Disregarding the absolute value, PI takes values from -1 (when $a = 0$ and $d = n$) to +1 (when $a = n$ and $d = 0$) and is equal to 0 when "Yes" and "No" are equally probable, i.e. when the average prevalence of "Yes" is 0.5. In fact, if the prevalence index is great (i.e. the prevalence of a positive rating is either very large or very small), the agreement by chance is also great and kappa is consequentially reduced, as it has been shown by Brennan and Silman [60].

Then Byrt et al. [56] defined a "Prevalence-Adjusted, Bias-Adjusted Kappa (PABAK)" by replacing the observed values in the principal diagonal cells by their average [$m = (a + d)/2$]. Rather surprisingly, PABAK is the same as Bennett et al.'s S coefficient [30], but obtained through a different derivation.

Gjørup [61] proposed an estimate of the prevalence of positive diagnosis by summing the proportion of the cell "yes-yes" with the mean of the proportions of the cells "yes-no" and "no-yes". It has to be reported the Cicchetti and Feinstein's proposal [62] of resolving the Cohen's kappa paradoxes by calculating the observed proportions of positive and negative agreement. Particularly, for a 2x2 contingency table with "a" and "d" the observed frequencies in the cells on the principal diagonal, "b" and "c" those in the adjacent cells and "n" the total number of the observations, the observed proportion of positive agreement (p_{pos}) is: $p_{pos} = 2a/(n + a - d)$, and the observed proportion of negative agreement (p_{neg}) is: $p_{neg} = 2d/(n - a + d)$. However, this solution does not reflect the effect of bias.

Furthermore, Cicchetti and Feinstein's indices [62] are closely related to Byrt et al.'s prevalence index [56], being $p_{pos} = (Po + PI) / (1 + PI)$ and $p_{neg} = (Po - PI)/ (1 - PI)$. Indeed, it has to be endorsed the Cicchetti and Feinstein's statement [62] that "no single omnibus index of agreement can be satisfactory for all purposes".

As a strategy for interpreting 2x2 agreement tables and reporting results, firstly the presence of bias should be assessed using Byrt et al.'s [56] Bias Index (BI) that it has to be "interpreted both in substantive terms (is this amount of bias important in this particular context?) and in terms of its effect on kappa". Furthermore, in presence of a relevant BI, it has to be done an exhaustive investigation to discover its cause, and it may be inappropriate or unnecessary to quote only an index of agreement, but it would be preferable to report the Cohen's kappa components reflecting the observed agreement, the bias index and the prevalence index. It has to be remembered the Sim and Wright's suggestion [63]:"when comparisons are made between agreement studies it can be misleading to report kappa values alone, and it is recommended that researchers should also discuss the effects of bias and prevalence."

Cohen's kappa: number of categories and marginal uniformity: Let's consider 2x2, 3x3, 4x4 and, finally, 5x5 uniform contingency tables. The same observed agreement proportion of $p_o = 0.7$(say) allows obtaining kappa values of 0.4, 0.55, 0.6, and 0.625. This fact corresponds to require for the same kappa values a lower observed agreement proportion. Otherwise, for the same kappa value of 0.4, p_o is 0.7, 0.6, 0.55, and, finally 0.52. As a further example let's consider non-uniform rows (columns) marginal with 0.10 for all the marginal, except for the last equal to 1 minus the sum of the previous ones. In this case a kappa = 0.6 is obtained from p_o of 0.928, 0.864, 0.808, and 0.76. Finally, for sake of completeness, the same observed agreement proportion of 0.8 allows to obtain kappa values of -0.1(1), 0.41176, 0.583(3), and finally, 0.6(6). So, it is very well-evident that increasing the number of categories allows having greater kappa values at the same value of the observed agreement proportion. Finally, the greater the non-uniformity of the marginal, the greater the observed agreement proportion (p_o) must be for obtaining satisfactory kappa values.

Weighted kappa: In the case in which the categories are naturally ordered, it is immediate to diversify the level of the agreement/disagreement, being a disagreement between two adjacent categories of the contingency table less important than a disagreement between categories that are more distant from each other.

So, Cohen [64] introduced the weighted kappa for a "nominal scale agreement" with an exemplification based on disagreement weights, being a weight equal to 0 given to the cells lying on the principal diagonal and values of 1 and of 3 given to progressively more distant categories. Moreover, Cohen [64] introduced also the weighted kappa for an agreement scaling with a weight equal to 1 for the cells on the principal diagonal and smaller weights as the cells symmetrically move away from the principal diagonal.

It has to be said that the main criticism against the weighted kappa is the fact that the weights can be arbitrarily chosen. However, very popular schemes are the so called "linear" and "quadratic" weights. The first scheme gives weights equal to $|i - j| / (I-1)$ consisting in the absolute difference between the number of the rows and the number of the columns divided by the number of the categories (I) minus one. In addition, according to the quadratic scheme, the weights are equal to $(i –j)^2 / (I-1)^2$, being the squared difference divided by the squared number of categories minus one.

Warrens [65] stressed the fact that the quadratic weights are more frequently used, since, in this case, the weighted kappa can be interpreted as an intraclass correlation coefficient; then, Warrens [65] highlighted some interesting properties of the weighted kappa, being Warrens' paper [65] together with Cohen's paper [64] very recommendable for readers interested in this topic.

Variance of kappa and of weighted kappa

The problematic aspects of this topic have been very well synthesized in the first sentence of the Fleiss et al.'s paper [66]:"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." The above text has been also reported by Kraemer et al. [67] by adding: "This wry comment by Fleiss et al. in 1979 continues to characterize the situation with regard to the kappas coefficients up to the year 2001, including not only derivation of correct standard errors, but also the formulation, interpretation and application of kappas."

The exact variance for the unweighted and weighted kappa has been shown by Everitt [68], even if the variance of the weighted kappa has been shown in a generally cryptic formula (formula 17, on page 101) with an "S" notation shortly defined as "S denotes the appropriate summation over the whole table". In addition, the pertinent calculations were not detailed and it has been reported only the variance value for the considered example 0.004417, very similar to the value of 0.004425 obtained from the approximate variance formula.

According to Everitt [68], let's define m_{ij} the absolute frequencies and w_{ij} the weights (indicating the level or seriousness of the disagreement and, of course, equal to zero for the diagonal cells in which there are the frequencies of agreement) of the ij cell (i-th row and j-th column for I rows and J columns, respectively); in addition, r_i and c_j are the marginal of the rows and of the columns, respectively, and, finally, N is the total sample size. So, using Everitt's notation [68], the probability density function of the m_{ij} is given by:

$$P(m_{ij}) = \frac{\prod_{i}^{I} r_i! \prod_{j}^{J} c_j!}{\prod_{i}^{I}\prod_{j}^{J} m_{ij} N!}$$

In addition, if the marginal are considered fixed, the marginal distribution of a single m_{ij} can be shown to be a (central in the case of k = 0, under H_0) hypergeometric distribution given by:

$$P(m_{ij}) = \frac{r_i!(N-r_i)!c_j!(N-c_j)!}{N!m_{ij}!(r_i-m_{ij})!(c_j-m_{ij})!(N-r_i-c_j+m_{ij})!}$$

However, according to Stevens [69], the above distribution has been previously shown by Yates [70] when marginal totals are fixed, provided that the entry into columns is independent of the entry into rows:

Therefore, the mean and variance of m_{ij} are, respectively:

$$\mathrm{Mean}(m_{ij}) = \frac{r_i c_j}{N};\ \mathrm{Var}(m_{ij}) = \frac{r_i(N-r_i)c_j(N-c_j)}{N^2(N-1)}$$

The covariance (Covar) between any two cells on the same row (m_{ij} and $m_{ij'}$, say) is:

$$\mathrm{Covar}(m_{ij}, m_{ij'}) = \frac{-r_i c_j c_{j'}(N-r_i)}{N^2(N-1)}$$

$$\mathrm{Covar}(m_{ij}, m_{i'j}) = \frac{-r_i r_{i'} c_j(N-c_j)}{N^2(N-1)}$$

Otherwise, for any two cells on the diagonal (m_{ij} and $m_{i'j'}$ for i = j, corresponding to m_{ii} and $m_{i'i'}$), the covariance (Covar) is:

$$\mathrm{Covar}(m_{ij}, m_{i'j'}) = \mathrm{Covar}(m_{ii}, m_{i'i'}) = \frac{r_i c_i r_{i'} c_{i'}}{N^2(N-1)}$$

that it is equal to the covariance between two diagonally adjacent opposed cells such as ($m_{ij'}$ and $m_{i'j}$).

Finally, for any two diagonally opposed cells (m_{ij} and $m_{i'j'}$, say), the Covar is:

$$\mathrm{Covar}(m_{ij}, m_{i'j'}) = \frac{r_i r_{i'} c_j c_{j'}}{N^2(N-1)}$$

Then, according to Everitt [68] the mean of k and of k_w are equal to zero, since:

$$E(k) = \frac{1}{1-p_e}\{E(p_0) - p_e\}$$,where E denotes the mathematical expectation and p_e is fixed by the marginal totals. Now:

$$E(p_0) = \frac{1}{N}\sum_{i=1}^{I} E(m_{ii}) = \frac{1}{N}\sum_{i=1}^{I}\frac{n_i M_i}{N} = \frac{1}{N^2}\sum_{i=1}^{I} n_i M_i = p_e$$

Therefore the expected value of k is zero.

The expectation of k_w is given by:

$$E(k_W) = 1 - \left(\sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} P_{eij} \right)^{-1} E\left(\sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} P_{oij} \right);$$

$$E\left(\sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} P_{o_{ij}} \right) = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} E\left[P_{o_{ij}} \right] = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} \frac{r_i M_j}{N^2} = \sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} P_{e_{ij}}$$

And it is also equal to zero.

Mielke et al. [71] wrote a Fortran program (VARKAP) for calculating the exact variance of the weighted kappa from which the calculation steps are more understandable and, in addition,considered contingency tables more than the 3x3 table examplified in Everitt's paper [68].

Of course, the program works also for the unweighted kappa which is a special case of the weighted kappa when the weights (w_{ij}) are equal to 1 for i ≠ j and 0 otherwise.

We translated Mielke et al.'s program [71] into SAS®/IML language [72] and also in the open source R language [73] and we obtained the same variance values at the sixths decimal figure shown in Mielke et al.'s example [71]. Furthermore, with the same programs we obtained 0.0042902 instead of the value of 0.004417 reported by Everitt [68], perhaps owing to the computational inaccuracies of the computer programs written and used in the years around 1970. It is questionable if the actual difference of about 13 thousandths instead of 8 millionths is still sufficiently small to be considered also irrelevant to justify, together with calculation difficulties, the preference for the approximate variance. Of course, nowdays, the difficulties of calculation are outdated, but it has to be pointed out that the exact variance of kappa (weighted or not) has been proposed only for the case of k = 0, as it is also reported in Mielke et al. [71], and, therefore, it is not useful for testing an hypothesis of at least an acceptable agreement, given by kappa values of 0.4, at least.

Indeed, it has also to be said that a statistical null hypothesis test of k_0=0 is rather not interesting and even unacceptable; in fact, the test must have a null hypothesis of a poor agreement in order to be rejected and, consequently, to conclude for a satisfactory enough agreement between the raters, according to the primary objective of an agreement study. Particularly, it has to be reported that k = 0.4 is generally considered the minimum threshold for a satisfactory agreement as it has been stressed by Everitt [68] and by Mielkeet al. [71], even if we strongly suggest a value of 0.5 in agreement also with Svanholm [49].

Mielke et al.'s formula [71] for the exact variance of the weighted kappa is:

$$\sigma^2_{\hat{k}} = \frac{1}{N-1} \left(\sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij} R_i C_j \right)^{-2} \left[\sum_{i=1}^{r} \sum_{j=1}^{r} w_{ij}^2 R_i (N - R_i) C_j (N - C_j) \right.$$

$$\left. - \sum_{i=1}^{r} \sum_{j \neq k} w_{ij} w_{ik} R_i (N - R_i) C_j C_k \; - \sum_{i \neq j} \sum_{k=1}^{r} w_{ik} w_{jk} R_i R_j C_k (N - C_k) + \sum_{i \neq k} \sum_{j \neq l} w_{ij} w_{kl} R_i R_k C_j C_l \right]$$

Where, in addition to the "i" and "j"subscripts for the rows and the columns of the square contingency table, there are two other subscripts "k" for two terms and "l" for the fourth and last term between the square brackets. Furthermore, R_i (or R_j or R_k) and C_j (C_k or C_l) are the marginal frequencies of the rows and of the columns, respectively, and the summation is from 1 to r, being r the number of the rows that, obviously, for a square table, corresponds to the number of the columns. Apart from the term at the denominator outside the square brackets and for the first term inside that are the expression of the usual double summation on the rows and columns of a contingency table, the second and third term (between the square brackets) are the expression of a triple and a quadruple summation, respectively. Particularly, for the second term, the summation is done for all rows ("i" = 1 to r) and for all columns except for the case in which the subscripts "j" and "k" are equal; in practice, it has to program a double summation with one index ("j", in this case) ranging from 1 to "r" − 1 and one another index ("k", in this case) ranging from "j" + 1 to "r".

The same approach is followed for the third term between the square brackets; in this case the subscript "i" ranges from 1 to "r" − 1, while the subscript "j" ranges from "i" +1 to "r" and the subscript "k" ranges from 1 to r. The last term between the square brackets requires a much more complicated approach in which there are four summations with the subscript "i" ranging from 1 to "r" − 1, the subscript "k" ranging from "i" +1 to "r", the index “j” ranging from 1 to “r “– 1 and, finally, the index “l” ranging from “j” + 1 to “r”.

Anyhow, the above outlined calculations are well understood (and also obtained) by reading a program written in SAS®/IML language [72] or by a program written in the open source software R [73] available from the authors. Curiously, it has to be said that the above reported value of 0.004417 for the exact variance of the weighed kappa in Everitt’s paper [68] has been also mentioned in Fleiss et al.’s paper [50] as: “Everitt (1968, p.102) found the exact variance to be 0.004417, which would indicate that the expression given in Equation 9 somewhat underestimates the exact value. The fact that the above value of 0.004425 from the approximate variance is practically equal to the exact variance value of 0.004417 *(a difference of 8 millionths: note from the authors)* supports the use of the approximate variance easier to calculate”.

Furthermore, the values of the kappa approximate variance shown by Cohen [32] are overestimated and not correct, according to Fleiss et al. [50]. Indeed, Cohen [32] derived the variance of the kappa (the weighted kappa will be introduced later) by considering p_e “as a constant” and, consequently, a fixed quantity and p_o as “if it were the population value”; in addition, the above assumption has been considered as adequate “ordinarily p_e will not vary greatly relative to k, particularly with large n (i.e., ≥ 300)”. Accordingly, the variance of p_o is p_o (1-p_o)/n and the kappa variance becomes the variance of p_o multiplied by the square of the constant value (1-p_e) at the denominator. The square root of the variance is the kappa standard error given by:

$$\sigma_k = \sqrt{\frac{p_0\left(1-p_0\right)}{n\left(1-p_e\right)^2}}$$

That is the equation 7 of the Cohen’s paper [32]. However it has to be noted that the use of the Greek letter “σ” should be associated to the Greek letter “π" instead of the Latin letter “p”.

Accordingly to the increase of n, it is possible to assume that the k sampling distribution will approximate the Gaussian distribution with the consequent calculation of the 95% and 99% confidence intervals by using the quantiles 1.96 and 2.58 of the standard Gaussian distribution, respectively; then, the testing of two independent kappa coefficients or a kappa against an expected value can be carried out by means of an approximate Z-test.

The formula for the approximate kappa variance from Fleiss et al. [50] is:

$$\text{s.e.}(k) = \sqrt{\frac{\tau(k)}{N}};\ \text{with}$$

$$\tau(k) = \frac{1}{\left(1-p_e\right)^4}\left\{\sum_{i=1}^{I} p_{ii}\left[\left(1-p_e\right)-\left(p_{i\cdot}+p_{\cdot i}\right)\left(1-p_0\right)\right]^2 + \left(1-p_0\right)^2 \sum_{i=1}^{I}\sum_{j=1,\, j\neq i}^{J} p_{ij}\left(p_{i\cdot}-p_{\cdot j}\right)^2 - \left(p_0 p_e - 2p_e + p_0\right)^2\right\}$$

Where $p_{i.}$ are the rows marginal and $p_{.i}$($p_{.j}$) are the columns marginal, respectively, p_0 is the sum of the observed proportion of agreement (the sum of the proportions of the cells on the principal diagonal of the contingency table), and p_e is the expected proportion of agreement by chance or in the case of a true independence or a true no agreement, corresponding to the sum of the proportions of the cells on the diagonal of the contingency table obtained by multiplying the corresponding rows and columns marginal proportions.

In the case of testing Cohen’s kappa H_0: k = 0, the above variance becomes the formula 14 of the Fleiss et al.’s paper [50]; namely:

$$\text{Var}(k_0) = \frac{1}{N\left(1-p_e\right)^2}\left\{\sum_{i=1}^{I} p_{i.}p_{.i}\left[1-\left(p_{i.}+p_{.i}\right)\right]^2 + \sum_{i=1}^{I}\sum_{j=1,\, j\neq i}^{J} p_{i.}p_{.j}\left(p_{.i}-p_{.j}\right)^2 - p_e{}^2\right\}$$

In the case of the weighted kappa, the variance (formula 8 Fleiss et al.’s paper [50])is:

$$\text{Var}(k_w) = \frac{1}{N\left(1-p_e\right)^4}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\left[w_{ij}\left(1-p_e\right)-\left(\bar{w}_{i.}+\bar{w}_{.j}\right)\left(1-p_0\right)\right]^2 - \left(p_0 p_e - 2p_e + p_0\right)^2\right\}$$

In the case of testing H_0: k_w = 0, the pertinent variance (formula 9 of Fleiss et al.’s paper [50])is:

$$\text{Var}(k_{w0}) = \frac{1}{N\left(1-p_e\right)^2}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i.}p_{.j}\left[w_{ij}-\left(\bar{w}_{i.}+\bar{w}_{.j}\right)\right]^2 - p_e{}^2\right\}$$

Where $\bar{w}_{i.} = \sum_{j=1}^{r} w_{ij}p_{.j}$ and $\bar{w}_{.j} = \sum_{i=1}^{r} w_{ij}p_{i.}$

Actually, Fleiss et al. [50] proposed firstly the asymptotic correct formula of the variance of the weighted kappa where w_{ij} are the weights for each cell of the contingency table since the unweighted kappa and its variance formulas are considered as a particular case of the weighted kappa.

The weights considered by Fleiss et al. [50] are equal to 1 for the cells on the main diagonal of the contingency table and are values arbitrarily chosen by the experimenter, assumed, without loss of generality, to lie between 0 and 1. In this case, the weights can be defined as "agreement weights" since their maximum values are on the "agreement cells" lying on the principal diagonal of the contingency table. It has to be taken into account that, for ordinal variables, a disagreement between adjacent rows or columns (class 1 vs. class 2, for example) is less serious that one between not adjacent rows or columns (class 1 vs. class 3, for example) or even more so between distant rows or columns (class 1 vs. class 4, for example). Of course, in the case of nominal qualitative variables without a natural order of the classes, the contingency table has to be appropriately built with rows (columns) in order of similarity. Apart from the previously reported variance formulas for illustrating their theoretical background, it has to be considered that in the sample size calculation, firstly it has to specify the kappa values under the null (k_0) and alternative (k_A) hypothesis, then the prevalence of the categories (rows and columns marginal) from which the proportion of the agreement by chance is calculated. Then, as a third step, it has to calculated the proportion of the observed agreement under the null (H_0: p_0) and alternative (H_A: p_A) hypothesis as:

$$\pi_0 = k(1-\pi_e)+\pi_e$$

Of course, the subscript "A" replaces the "0" in the case of the alternative hypothesis, and the Latin letter "p" replaces the Greek letter "π" in the case of the observed values. Finally, as a fourth step, the cells of the contingency table must be filled under the null and the alternative hypothesis to calculate the corresponding kappa variances. It is very well evident that only in the simplest case of a 2x2 contingency table there is a unique reference table, given the rows marginal (r_1, r_2), the columns marginal (c_1, c_2) and the kappa value (k) from which the observed agreement proportion is calculated. Indeed, the proportions of the four cells (p_{11}, p_{12}, p_{21}, and p_{22}) can be obtained as: (i) $p_{22} = p_0 - p_{11} \rightarrow p_{22} = p_0 - (r_1 - p_{12}) = p_0 - r_1 + p_{12}$; then since $p_{12} = c_2 - p_{22}$, it is straightforward to obtain $p_{22} = p_0 - r_1 + c_2 - p_{22} \rightarrow p_{22} = (p_0 - r_1 + c_2)/2$. The remaining proportions are consequently: (ii) $p_{11} = p_0 - p_{22}$; (iii) $p_{12} = r_1 - p_{11}$; (iv) and, finally, $p_{21} = c_1 - p_{11}$.

In the case of a symmetrical contingency table 3x3, the condition of dividing π_0 proportionally to the rows (columns) marginal and, then,of dividing the disagreement in the cells outlying the principal diagonal proportionally to the rows (columns) marginal does not allow to obtain a unique table unless some other conditions are added such as the condition $p_{12} = p_{21}$, $p_{23} = p_{32}$, and $p_{13} = p_{31}$, leading to a symmetrical pattern of the cells of the contingency table. So, with row and column marginal equal to 0.5, 0.3, and 0.2 leading to π_e = 0.38, and with k = 0.4, we obtain p_0 = 0.628 that can to be proportionally divided on the principal diagonal cells; namely 0.628*0.5 = 0.314, 0.628*0.3 = 0.1884, and 0.628*0.2 = 0.1256. Then, by dividing the disagreement proportion of 0.186 (from the difference: 0.5 – 0.314) in the first row proportionally to the marginal as 0.186*0.3 / (0.3 + 0.2) and 0.186*0.2 / (0.3 + 0.2), we obtain 0.1116 to be put in the $\text{cell}_{(1,2)}$ and 0.0744 to be put in the cell $_{(1,3)}$, respectively.

The above values have to be put in the $\text{cell}_{(2,1)}$ and $\text{cell}_{(3,1)}$, respectively for the symmetry; since the sum: $p_{11} + p_{12} + p_{13}$ (0.314 + 0.1116 + 0.0744) equals the first row marginal of 0.5, the sum: $p_{12} + p_{22}$ (0.1116 + 0.1884) equals the second row marginal of 0.3, and, finally the sum: $p_{13} + p_{33}$ (0.0744 + 0.1256) equals the third row marginal of 0.2. Finally, the remaining $\text{cell}_{(2,3)}$ and $\text{cell}_{(3,2)}$ are filled with 0.

The above calculations correspond to having written an algebraic system of six equations; namely, three equations for the fixed rows marginal, two equations for the fixed columns marginal (three minus 1, since the condition for the third column is implied in the previous three row conditions) and one equation for the condition that the sum of the proportions in the diagonal cells have to be equal to π_0. So, in order to obtain a solution, the 9 equations have to be reduced to six by equaling two by two the proportions on the symmetrical cells out of the diagonal, for example. However, it is not guaranteed that the algebraic solution leads to a contingency table with the cells showing a symmetrical pattern.

In addition, it may happen that all cells are filled by non-zero values, as it occurs for rows (columns) marginal of 0.4, 0.3 and 0.3, leading to π_e of 0.34, and for k = 0.4, leading to π_0 =

0.604. In this case, $p_{11} = 0.2416$, $p_{22} = 0.1812$, and $p_{33} = 0.1812$, $p_{12} = 0.0792$, $p_{13} = 0.0792$, $p_{21} = 0.0792$, $p_{23} = 0.$ 0.0396, $p_{31} = 0.0792$, and $p_{32} = 0.0396$. It can be seen that $p_{23} = p_{32} = 0.0396$, are different from zero and that their sums are equal to the rows and columns fixed marginal.

However, this approach can fail in some cases. For example, with row (column) marginal equal to 0.8, 0.1, 0.1 leading to π_e of 0.66, and with kappa = 0.4 giving π_0 of 0.796, we obtain $p_{11}= 0.6368$, $p_{22} = p_{33} = 0.0796$. Then, the disagreement in the first row equal to 0.1632 is equally split in the $cell_{1,2}$ and $cell_{1,3}$ with the consequence that 0.0816 + 0.0796 = 0.1612, a value greater than the fixed second and third row (column) marginal equal to 0.10.

In any case, the above contingency table not fulfilling the condition of fixed rows (columns) marginal can be submitted to an iterative process by increasing the $cell_{(1,1)}$ and decreasing all other cells until the above condition is fulfilled. However, the 3x3 contingency table obtained is, again, a not unique table fulfilling the conditions of the rows (columns) marginal and of the observed agreement given by the sum of the cells on the principal diagonal.

For example, both the 3x3 tables (0.732 0.034 0.034; 0.034 0.032 0.034; 0.034 0.034 0.032) and (0.732 0.068 0.000; 0.000 0.032 0.068; 0.068 0.000 0.032) fulfill the condition of having the same rows (0.8, 0.1, 0.1) and columns (0.8, 0.1, 0.1) marginal and the condition of having the same observed agreement proportion of 0.796 or the same kappa value of 0.4.

It has also to be remembered the proposal of Gardner [58] followed also by Bakeman et al. [51] and by Bakeman [59] of filling the cells of the contingency table based on the accuracy less than 1 for fallible raters. Indeed, the above calculations are based on a perfect accuracy of the raters.

So, starting from a cxc accuracy matrix of the rater "A" (A_{ic}) with on the principal diagonal the accuracy values and on the remaining cells the complement to 1 of the accuracy divided by the number of the categories minus 1, and from a similar cxc accuracy matrix for the rater "B" (B_{jc}), and the rows (columns) vector of the category prevalence (π_c), it is possible to obtain a baseline matrix from which the value of kappa can be calculated. This quite laborious procedure is clearly shown in the downloadable Bakeman's Technical Report [59] to which the interested readers are referred.

It has to be noted that even if the approach with no fallible or fallible raters allows obtaining a unique agreement matrix, there is no guarantee that this matrix corresponds to the population matrix of the multinomial distribution from which the kappa value may have been generated.

Furthermore, Gardner's approach [58] and the "symmetrical approach" shown before lead to a cxc agreement matrix without considering the value of the kappa variance that it will be possible to calculate. In any case, there are several cxc agreement contingency tables fulfilling the main requirements of the fix rows (columns) marginal and of the kappa value (or of the observed agreement).

Sample size calculation

Sample size calculation according to Flack et al. [74]: In order to obtain a unique contingency table Flack et al. [74] introduced the criterion of selecting the table with the maximum asymptotic standard error proposed by Fleiss et al. [50] leading to a conservative sample size calculation.

Flack et al.'s approach [74] was also followed by PASS® 13 and PASS® 16 [75] in the procedures "Kappa Test for Agreement Between Two Raters", described in Chapter 811, and "Confidence Intervals for Kappa", described in Chapter 819, respectively.

Flack et al. [74] considered the "large-sample standard error" formula given by Fleiss et al. [50] and concluded that, "given the marginal and the kappa value, the maximum variance value can be obtained by maximizing the double summation of the variance formula by placing all of the off-diagonal probability in the cells corresponding to the largest marginal and putting zero on the remaining off-diagonal cells". In addition, "it has to minimize the single summation term that is subtracted from the first by maximizing the proportions in the cells corresponding to the smallest marginal". Flack et al. [74] gave an example of symmetrical table with marginal (0.1, 0.2, 0.3, and 0.4) and kappa = 0.8 from which it is easily obtained the disposition of the cell proportions in order to have the maximum of the kappa variance.

We recommend a mathematical method based on the linear programming (LP, also called linear optimization) that allows to obtain a best outcome (such as maximum or a minimum) in a mathematical model whose requirements are represented by linear relationships, as it has been stressed by Pratt and Hughes [76], which optimize (usually maximize or minimize) a linear objective function of many variables, subject to linear constraints. The constraints, in this case are given by the fixed rows and columns marginal and by the observed agreement proportion, fixed by the kappa values under H_0 and H_A (see Appendix A. Linear Programming).

To this aim, it can be used the LPSOLVE subroutine in SAS®/IML of SAS® 9.4 [72] or the linear programming (LP) solver in the OPTMODEL procedure of SAS®/OR, reported by Pratt and Hughes [76]. Otherwise, it can be used the "IpSolve" R Package from Berkelaar et al. [77]. The package "IpSolve" is also used by the function N2.cohen.kappa. R in the package "irr" [78]. We have generalized this function in a program written in the open source R language to consider the full expression of the variance of Cohen's unweighted and weighted kappa. In addition, we have reported the correct number of the sufficient constraints, together with the theory of the LP used in this particular context in the Appendix A. In any case, the interested readers can be referred to the almost exhaustive reference of Sallan et al. [79].

However, it has to be noted that the algebraic solution of the system under the linear programming approach can lead to a situation, not acceptable from the point of view of an agreement study, consisting of zero values for some cells on the principal diagonal together with some, perhaps, relevant values in the "disagreement cells. It has to be said that it is not possible, from the theoretical point of view that the agreement is only for some classes of the variables with a complete disagreement for the remaining ones. The same is true if all disagreement is present in only a few cells of the contingency table, particularly if these cells are just those adjacent to the principal diagonal.

It has also to be said that with the linear programming approach, it is possible to determine the contingency probability table with fixed rows (columns) marginal and the Cohen's kappa value for obtaining any predetermined value of the variance.

Sample size calculation according to Altaye et al. [87,89] and Donner et al. [92,94]: Furthermore, under a model parallel to the common correlation model used for the case of continuous variables that implies an equal correlation coefficient between any pair of variables, Donner and Eliasziw [80] proposed a "goodness of fitting (GOF)" approach to develop confidence interval and significance-testing procedures for the kappa statistic. Donner and Eliasziw [80] called their model "the common correlation model for dichotomous data" and developed it for dichotomous data and for two raters under the assumption that there is no "rater bias", leading to uniform underlying success proportions of the two raters.

Particularly, Donner and Eliasziw's [80] "goodness of fitting" approach brings together the disagreement proportions and, in the case of a 2x2 contingency table, calculates the sum of the squared difference between the observed and expected proportions divided by the expected proportion. The resulting statistic, in the case of binary variables and two raters, has a limiting χ^2 distribution with one degree of freedom. It has to be noted that the Donner and Eliasziw's approach [80] of putting together the disagreement proportions implies that it is possible to consider only the case of "agreement or not" disregarding the weighted kappa based on a disagreement gradually evaluated from slight to severe.

Then, Donner and Eliasziw [80] show how to obtain the values of the lower and upper confidence limit of the kappa by resolving a cubic equation; similarly, it is also possible to solve for a single root the cubic equation for a one-sided 100(1 - a) % lower confidence limit. It has to be reported that the "GOF" approach gave conditional empirical coverages close to the nominal value from a Monte Carlo study. So, the "GOF" resulted to be superior to two methods proposed by Bloch and Kraemer [81], one derived by a "large sample standard error" and the other obtained by a "variance stabilizing transformation "aimed to improve the accuracy of the confidence interval estimate.

In addition, Donner and Eliasziw [82] showed how to use the goodness-of-fit test procedure for sample size calculations in the context of reliability studies. Particularly, it has to be considered the central χ^2 distribution with one degree of freedom under the null hypothesis and the non-central χ^2 distribution and its non-centrality parameter with one degree of freedom under the alternative hypothesis.

So, the sample size (n) formula from Donner and Eliasziw [82] is:

$$n=\lambda(1,1-\beta,\alpha)\left\{\frac{\left[\pi(1-\pi)(k_1-k_0)\right]^2}{\pi^2+\pi(1-\pi)k_0}+\frac{2\left[\pi(1-\pi)(k_1-k_0)\right]^2}{\pi(1-\pi)(1-k_0)}+\frac{\left[\pi(1-\pi)(k_1-k_0)\right]^2}{(1-\pi)^2+\pi(1-\pi)k_0}\right\}^{-1}$$

Where l is the non-centrality parameter, 1-β is the power and α is the statistical significance.

Interestingly, the non-centrality parameter λ for 1 degree of freedom is, simply, the squared sum of the 1-$\alpha/2$ and 1-β percentiles of the Z-distribution. Then, for a significance level of 0.0975 and a power of 0.80, corresponding to $z_{1-\alpha/2}$ = 1.959963985 and $z_{1-\beta}$= 0.841621234, l = 7.848885936 (say 7.849); for a significance level of 0.0975 and a power of 0.90, corresponding to $z_{1-\alpha/2}$ = 1.959963985 and $z_{1-\beta}$ = 1.281221566, l = 10.50528378 (say 10.505 instead of the 10.507, reported in the table on page 1518 of the Donner and Eliasziw's paper [82]).

Then, the approach described by Donner and Eliasziw [80,82] has been extended by Donner et al. [83] to testing the homogeneity of k independent kappa statistics of the intraclass form. In addition, Donner [84] provided sample size formulas and tables for designing studies comparing two or more inter-observer agreement or concordance coefficients. Furthermore, always for the case of two raters and a dichotomous variable, Donner [85] showed the sample size requirements for a prespecified expected width or a lower limit of a confidence interval of the intraclass kappa statistic. Bahadur's model [86] was subsequently applied by Altaye et al. [87] to modeling interrater agreement among multiple raters; particularly, the maximum likelihood estimators were obtained by a reparameterization of the Bahadur's model [86] shown by George and Bowman [88]. Thereafter, Altaye et al. [89] extended their previous results by proposing a χ^2 goodness-of-fit test based on the Dirichlet multinomial distribution for considering multiple raters and/or polytomous nominal variables. It has to be noted that the Dirichlet multinomial model expresses the joint distribution of the ratings and it has been used in the case of data obtained from a cluster sampling as it has been shown by Brier [90] and for making inference about the intraclass correlation coefficient in the context of twin studies as it has been shown by Bartfay et al. [91]. Altaye et al. [89] reported that their model allows to have coverage and type I error proportion close to nominal (Table 2 and Table 3) and to obtain a sample size formula for the required number of subjects and raters that provides predetermined power to test statistical kappa hypotheses. As the number of raters increases, the required number of subjects decreases, but this sample size saving rapidly diminishes after the accrual of five raters, as it is shown in Table IV of Altaye et al.'s paper [89]. This sample size calculation approach corresponds to an extension of the case of two raters with a binary outcome variable shown by Altaye et al. [87] to the case of multiple raters and to polytomous variables. A different formulation of the Dirichlet multinomial distribution that allows considering the case when the assumption of mutual independence does not hold is shown in the Appendix B. Dirichlet Multinomial Distribution.

Donner and Rotondi [92] showed sample size requirements for interobserver agreement studies with a binary outcome in terms of the number of subjects (N) and raters (nr) that would allow the expected lower bound of a 95% confidence limit for Cohen's kappa to be equal or greater than a required threshold. Of course, the required threshold has to be chosen for having an adequate agreement level. Donner and Rotondi's approach [92] is based on the equivalence between Cohen's kappa and the intraclass correlation coefficient obtained from a one-way random effects model shown by Fleiss [93], and a parsimonious model for correlated binary proposed by Bahadur [86]. Donner and Rotondi's paper [92] shows two sample size tables: the first (Table 2) reports the number of subjects (N) "required to ensure that the expected lower limit of a 95% one-sided confidence limit for k·is no less than k_L" for four values of k_0 (0.5, 0.6, 0.7, and 0.8), two values of k_L (0.40 and 0.60 for each k_0 value), three values of π (0.1, 0.3, and 0.5), and four different number of raters (n = 2, 3, 4 and 5). Then, Donner and Rotondi's Table 3 [92] reports, for the same values of π and number of raters, the "Expected lower limit of a 95% one-sided confidence limit" for k, with fixed sample sizes of 25, 50, 100, 150, and 200, and with k_0 ranging from 0.5 to 0.8 by step of 0.10.

The sample sizes calculations were performed by means of the package "kappaSize" written in the open source R language by Rotondi [94]. The package "kappaSize" [94], in addition to the sample size based on the lower confidence limit for k, allows to calculate a power base sample size for comparing two kappa (k_0 under the null hypothesis and k_1 under the alternative one) for 2 to 5 categories and for 2 to 5 raters (PowerBinary, Power3Cats, Power4Cats, and Power5Cats).

These sample sizes will be compared with those obtained from Flack et al. 's procedure [74] and from our two models leading to three sample size calculations in the results section in the case of two raters. It has to be stressed that the sample size obtained by PASS® 13/16[75] in the case of two raters corresponds to those calculated according to Flack et al. [74].

However, it has to be strongly pointed out that the Donner and Rotondi's approach [92] does not "… ensure that the expected lower limit of a 95% one-sided confidence limit for k is no less than k_L" since, firstly the above result is a probabilistic and not a deterministic event, and, secondly, the probability of obtaining the above result is only about 0.50, owing to the fact that the "confidence interval power" has not been taken into account.

Indeed, with 2 raters and binary variables with marginal probabilities of 0.6 and 0.4, k = 0.6 under the null hypothesis, the "desired expected lower confidence limit" for kappa (k_L) equal to 0.5 and $\alpha = 0.05$, the required sample size is 299 from the "CIBinary" function of the package "kappaSize" [94]. Then, the 95%CI of kappa of 966 samples out of 1,000 simulated samples includes the kappa simulation parameter 0.6 with an adequate coverage of 0.966, but the lower 95% CI was greater than the "desired expected lower confidence limit" of 0.5 in only 545 (54.5%) samples. Similar results have been obtained from another simulation case with 2 raters, multinomial variables (3 classes) with marginal probabilities of 0.6, 0.3, and 0.1, k = 0.6 under the null hypothesis, the "desired expected lower confidence limit" for kappa (k_L) equal to 0.5 and $\alpha = 0.05$.

The required sample size is 245 from the "CI3Cats" function of the package "kappaSize" [94]. Then, the 95% CI of kappa of 964 samples out of 1,000 simulated samples includes the kappa simulation parameter 0.6 with an adequate coverage of 0.964, but the lower 95%CI was greater than the "desired expected lower confidence limit" of 0.5 in only 555 (55.5%) samples.

So, in our opinion Donner and Rotondi's suggestion [92] is not shareable, unless, after the first sample size calculation, a second iterative procedure is implemented to increase the first sample size until the area under a non-central χ^2 distribution and the required threshold k_L has a sufficiently satisfactory value corresponding to the probability of obtaining the required result (CI power). Hong et al. [95] followed Donner and Rotondi's approach [92] even if not openly stated. Indeed, they reported Donner and Rotondi's Table 1 [92] with the same inaccurate 10.507 value instead of 10.505 as we shown before and produced nomogram for sample sizes calculation for interobserver agreement studies with only two raters for several prevalence patterns of 2, 3, 4, and 5 categories. However, Hong et al. [95] sample sizes calculation is based on the comparison between two agreement proportions instead of two kappa values. The motivation of their approach is "Because a large difference in kappa values would produce a negligible difference in proportion of agreement, it seems reasonable that the level of agreement in a sample size calculation is considered in terms of the proportion of agreement rather than a kappa value." However, owing to the complex relationships between marginal probabilities, observed agreement proportions and kappa values, it is our opinion that researchers should be trained to infer in terms of the maximum achievable kappa value given the marginal probabilities and the relevant difference between two kappa values that it is sensible to postulate in an agreement study.

The comparison between the sample sizes proposed by Hong et al. [95] and Flack et al. [74] or PASS®13/16 [75], is very similar to that between and Altaye et al. [87,89] and Donner et al. [92,94] and Flack et al. [74] or PASS®13/16 [75] considered in the results section. Hence, it will not be considered in detail.

Interestingly, Von Eye and Mun [24] reported a "more general approach to power analysis for k, recently proposed by Indurkhya et al. [96], based on a Dirichlet multinomial distribution allowing to derive a χ^2-distributed statistic for the null hypothesis that $k_0 = 0$ and a consequent sample size equation". So, von Eye and Mun [24] showed a Table 1.4 "Minimum Required Sample Sizes for $\alpha = 0.05$ and p = 0.8 (power)" adapted by summarizing two tables of Indurkhya et al.'s paper [96]. However, Indurkhya et al.'s paper [96] is not actually present and, consequently downloadable, from the internet site of the journal. So, it not possible to understand the details of the proposed methods. However these sample sizes correspond only in the case of three categories to the sample sizes calculated according to Altaye et al. [87,89] and Donner et al. [92,94]. (See Appendix C. Sample Sizes shown in the von Eye and Mun's book [24] for a more detailed discussion on this topic).

Sample size calculation according to Choudhary and Nagaraja [18], and Cantor [97]: It has also to be remembered the very pragmatic presentation of the sample size calculation given by Choudhary and Nagaraja [18] without considering the problems of the different variances values, very likely because it has been shown only the case of two raters and two categories.

They used a different formulation of the Fleiss et al. [50] formula of the kappa variance given by:

$$\mathrm{Var}(k) = \frac{1}{N\left(1-p_e\right)^2}\left\{\sum_{i=1}^{2} p_{ii}\left[1-\left(p_{i\bullet}+p_{\bullet i}\right)\left(1-k\right)\right]^2+\left(1-k\right)^2 \bullet\left[p_{12}\left(p_{1\bullet}+p_{\bullet 2}\right)^2 +p_{21}\left(p_{2\bullet}+p_{\bullet 1}\right)^2\right]-\left[k-p_e\left(1-k\right)\right]^2\right\}$$

This variance can be used, under the assumption of a Normal distribution in the case of large sample, to calculate approximate confidence intervals and to test the agreement hypotheses (H$_0$: k≤k0 vs. H$_1$: k>k$_0$, for example). In addition, the sample size calculation formula (12.28 in the paragraph 12.4.7 Sample Size Calculations) of Choudhary and Nagaraja' book [18] is simply given by:

$$N = \hat{\sigma}^2 \frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{\left(k_0-k_1\right)^2}$$

Where $\hat{\sigma}^2$ is an estimate of the kappa variance and $\zeta_{1-\square}$ and $\zeta_{1-\square}$ are the (1-α) and (1-β) quantiles of the standard Normal distribution. It is interesting to note that Choudhary and Nagaraja's suggestion [18] to take into account the sample size increases the further one moves away from the situation in which the rows (columns) marginal are equal to 0.5, is to use preliminary estimates of the prevalence of the classification classes and, then, to insert in the sample size formula the greater variance obtained by substituting k$_0$ or k$_1$. Indeed, for rows (columns) marginal of 0.6 and 0.4, k$_0$ = 0.4 and k$_1$ = 0.7, α = 0.025 and power = 0.80, the sample sizes calculated by using k$_0$ and k$_1$ are 76 and 46, respectively, instead of 66 and 67 calculated according to Cantor [97] and PASS®13/16 [75], respectively.

In addition, for rows (columns) marginal of 0.5 and 0.5, k$_0$ = 0.3 and k$_1$ = 0.5, a = 0.025 and power = 0.80, the sample sizes calculated by using k$_0$ and k$_1$ are 179 and 147, respectively, instead of 169 and 168 calculated according to Cantor [97] and PASS®13/16 [75], respectively. So, it seems rather problematic to follow Choudhary and Nagaraja's suggestion [18].

It has also to be reported the sample size calculation formula proposed by Cantor [97] in which the variances under the null and alternative hypothesis are separated together with their pertinent quantiles:

$$N_{\mathrm{Cantor}} = \frac{\left(z_{1-\alpha}\sqrt{\hat{\sigma}_0}+z_{1-\beta}\sqrt{\hat{\sigma}_1}\right)^2}{\left(k_0-k_1\right)^2}$$

It has to be stressed that this formulation that considers the variances under the null and the alternative hypothesis, is more usual in the sample size calculation settings. Finally, the sample size shown in Cantor's paper [97] with the same conditions as the latter (rows (columns) marginal of 0.5 and 0.5, k$_0$ = 0.3 and k$_1$ = 0.5, power = 0.80,) except for α = 0.05 is 131 (132 rounding at the higher integer), 141 and 116 according to Choudhary and Nagaraja [18], and, finally, 133 from PASS® 13/16 [75].

Aims of the Paper

First and immediate aim of our research is to give the mathematical theory supporting Flack et al.'s proposal [74] by using the Linear Programming (LP) to directly obtain the contingency table with the maximum value of the kappa variance. Nonetheless, it has to be remembered that using LP, it is also possible to obtain several kappa variance values, such as the minimum or any predetermined value. The possibility of calculating the maximum and the minimum kappa variance values allows obtaining intermediate sample size values which could guarantee the actual feasibility of an agreement study. This point is fully considered in the Appendix A. Linear

Programming with an example of sample size calculation, to which readers are referred.

Our second aim is to propose a generalization of the “common correlation model for dichotomous variables” to multinomial ones, leading to a new theoretic starting point for the sample size calculation procedure. Particularly, under this model, we have imposed two constraints: the first, less restrictive, is that only the cells on the main diagonal have a common correlation coefficient, leading to the “Partial Common Correlation Model (PCCM)” and the second, more restrictive, is that all cells of the contingency table have a common correlation coefficient, leading to the “Full Common Correlation Model (FCCM)”, Of course, these two models give two different sample sizes calculation procedures.

Our third aim is to compare the sample sizes obtained from the above common correlation models with the sample sizes obtained according to Flack et al. [74] or PASS®13/16 [75], according to the “Goodness of fit” model of Altaye et al. [87,89] and Donner et al. [92,94], and for sake of completeness with the values calculated with the minimum kappa variance value. A further aim is to confirm the conservative characteristic of the sample sizes from Flack et al. [74] or PASS® 13/16 [75], by calculating the empirical power by means of a simulation study with samples having a sample size obviously calculated according to Flack et al. [74]. In addition, we have also calculated the coverage of the 95% Confidence Interval of the sample kappa to evaluate whether the conservative sample sizes guaranteed the nominal coverage or not. Finally, a further relevant aim of our paper is to give some suggestions for obtaining sample sizes suitable for the actual feasibility of an agreement study.

METHODS

The sample size calculation procedure requires the knowledge of the kappa variance given by Fleiss et al. [50]. It has to be observed that the variance is known and unique only if all joint probabilities π_{ij} of the contingency table are determined. It is well known that for 2x2 contingency tables the “unity variance” of Cohen’s kappa is uniquely determined; so, it is possible to calculate the correspondent variance under the null (H_0) and alternative (H_A) hypothesis to be inserted into the sample size calculation formula. It has to be noted that the kappa variance is conveniently defined the “unity kappa variance”, since it is obtained with a sample of only one unity in order to remove the influence of the sample size on its value. Otherwise, for square tables cxc with c>2, the cell probabilities π_{ij} are not uniquely determined, and it is not possible to calculate a unique *Var (k)* and, consequently, a well-defined sample size value.

Flack et al. [74] calculated the sample size after having determined the cell probabilities in order that the kappa variance is a maximum under the null hypothesis (H_0) and the alternative one (H_A). However, the procedure of obtaining the cell probabilities configuration is not adequately described. We have shown how to obtain the maximum (and the minimum) of the kappa variance by resorting to the linear programming (Appendix A). More generally, it is possible to determine the cell probabilities for obtaining a particular value of the kappa variance such as its maximum or its minimum or any other particular value. Furthermore, for counterbalancing the fact that sample sizes calculated according to Flack et al. [74] can be too much conservative, we have calculated the sample sizes under two common correlated multinomial models by extending the common correlated binary model.

Common Correlation Model: Contingency Tables 2x2.

The common correlation model in the case of 2x2 contingency tables has been considered, among several authors, also by Bloch and Kraemer [81]. Particularly, given two identical correlated Bernoulli variables X and Y, the joint probabilities π_{ij} of their corresponding contingency table are uniquely determined, given the correlation coefficient ρ between X and Y.

Let’s represent the probabilities of a 2x2 table:

		X		
		1	0	
Y	1	π_{11}	π_{12}	π
	0	π_{21}	π_{22}	$1-\pi$
		π	$1-\pi$	

With

$$\pi_{11} = P(X = 1 \cap Y = 1) = E(X \cdot Y)$$

$$= E(X \cdot Y) - E(X)E(Y) + E(X)E(Y)$$

Remembering that: $Cov(X,Y) = E(X \cdot Y) - E(X)E(Y)$

We obtain:

$$\pi_{11} = Cov(X,Y) + E(X)E(Y)$$

Remembering that:

$$E(X) = E(Y) = \pi \quad and\ Var(X) = Var(Y) = \pi(1-\pi)$$

and that:

$$Cov(X,Y) = \rho\sqrt{Var(X)Var(Y)} = \rho\pi(1-\pi)$$

We obtain:

$$\pi_{11} = P(X = 1 \cap Y = 1) = \rho\pi(1-\pi) + \pi^2$$

And consequently:

$$\pi_{12} = \pi_{21} = \pi - \pi_{11} = \pi(1-\pi)(1-\rho)$$

$$\pi_{22} = (1-\pi) - \pi_{12} =$$

$$= (1-\pi)(1-\pi+\pi\rho)$$

In addition, Cohen's kappa corresponds to the correlation coefficient between the two X and Y variables: $k = \rho$.

The demonstration is obtained by calculating, from the above reported probabilities, π_0 and π_e and, then, the kappa value:

$$\pi_0 = \pi_{11} + \pi_{22} = (1-\pi)(1-\pi+2\pi\rho) + \pi^2$$

$$\pi_e = \pi^2 + (1-\pi)^2$$

$$k = \frac{\pi_0 - \pi_e}{1-\pi_e} = \cdots = \frac{2\rho\pi(1-\pi)}{2\pi(1-\pi)} = \rho$$

Then, if the rows and columns marginal of a 2x2 contingency tables, are fixed, together with a fixed Cohen's kappa value, the joint probabilities π_{ij} are uniquely determined. In other words, there is only one 2x2 contingency table with those particular rows and columns marginal and that particular Cohen's kappa value. Consequently, also the kappa variance is uniquely determined. So, after having fixed the rows and columns marginal, the k_0 value under the null hypothesis (H_0) and the k_A value under alternative hypothesis (H_A), it is possible to calculate the corresponding unique variance values to be inserted in the sample size calculation formula.

Common correlation model: contingency tables cxc.

Let's consider two correlated identical qualitative multinomial variables (X and Y) together with their joint probabilities square table. Taking into account the previously reported common correlation model, the cxc table can be reduced to a 2x2 one. Particularly, the probability value (π_{ij}) of one cell has to be kept unchanged and the remaining cells on the c-1 columns have to be collapsed by summing their values in order to obtain a cx2 table. Finally, the cx2 table can be collapsed to a 2x2 table by summing the respective cell probabilities on the c-1 rows. So, it will be possible to calculate the joint probability among the unchanged cell and the remaining cells obtained by summing the remaining cells (c-1 for the columns and c-1 for the rows, respectively).

For example, starting from the cell$_{(i,j)}$ with probability π_{ij}(say, the first cell$_{(1,1)}$ with probability π_{11}), the cxc table can be collapsed in the following way.

		Y		
		j	c-j	
X	i	π_{ij}	$(\pi_{i.} - \pi_{ij})$	$\pi_{i.}$
	r-i	$(\pi_{.j} - \pi_{ij})$	$(1 - \pi_{i.} - \pi_{.j} + \pi_{ij})$	$1 - \pi_{i.}$
		$\pi_{.j}$	$1 - \pi_{.j}$	

The two indicator variables X* and Y* of this table are:

		Y*		
		1	0	
X*	1	π_{ij}	$(\pi_{i.} - \pi_{ij})$	$\pi_{i.}$
	0	$(\pi_{.j} - \pi_{ij})$	$(1 - \pi_{i.} - \pi_{.j} + \pi_{ij})$	$1 - \pi_{i.}$
		$\pi_{.j}$	$1 - \pi_{.j}$	

Then, the cell probability π_{ij} is:

$$\pi_{ij} = P(X = i \cap Y = j) = P(X^* = 1 \cap Y^* = 1)$$

$$= P(X^* \cdot Y^* = 1) = Cov(X^*, Y^*) + E(X^*)E(Y^*)$$

Let's ρ_{ij} be the correlation coefficient between X^* *and* Y^*.

Remembering that:

$$E(X^*) = \pi_{i.}, E(Y^*) = \pi_{.j}$$

$$and\ Var(X^*) = \pi_{i.}(1-\pi_{i.}),$$

$$Var(Y^*) = \pi_{.j}(1-\pi_{.j})$$

and

$$Cov(X^*, Y^*) = \rho_{ij}\sqrt{Var(X^*)Var(Y^*)} =$$

$$= \rho_{ij}\sqrt{\pi_{i\cdot}(1-\pi_{i\cdot})\pi_{\cdot j}(1-\pi_{\cdot j})},$$

we obtain:

$$\pi_{ij} = \rho_{ij}\sqrt{\pi_{i.}(1-\pi_{i.})\pi_{.j}(1-\pi_{.j})}+\pi_{i.}\pi_{.j}$$

The above described operation of collapsing the original cxc table into a 2x2 table, has to be iterated for all remaining cells on the rows and, then, on the columns of the contingency table.

In general, the probability value of each cell is given by:

$$\pi_{ij} = \begin{cases} \rho_{ij}\sqrt{\pi_{i\cdot}(1-\pi_{i\cdot})\pi_{\cdot j}(1-\pi_{\cdot j})} + \pi_{i\cdot}\pi_{j\cdot} & for\ i \neq j \\ \rho_{ii}\pi_{i\cdot}(1-\pi_{i\cdot}) + \pi_{i\cdot}^2 & for\ i = j \end{cases}$$

So, the joint probabilities π_{ij} can be expressed in terms of the rows and columns marginal probabilities together with their correlation coefficient. Obviously, the knowledge of the marginal probabilities and of the correlation coefficient allows calculating all joint probabilities of the table.

In the case of a 2x2 contingency table, the two assumptions (knowledge of the rows and columns marginal together with the Cohen's kappa value) allow to obtain a unique table. However, for contingency square tables more than 2x2, these two assumptions are not sufficient for obtaining a unique table. Indeed, it has to consider that there are 2c independent linear constraints; particularly, c constraints for the rows, c-1 constraints for the columns and 1 constraint for the principal diagonal:

$$\sum_{j=1}^{c} \pi_{ij} = \pi_{i\cdot}\ for\ i = 1, \dots, c$$

$$\sum_{i=1}^{c} \pi_{ij} = \pi_{\cdot j}\ for\ j = 1, \dots, (c-1)$$

$$\sum_{i=1}^{c} \pi_{ii} = \pi_0$$

So, starting from the (c-1)(c-1) degrees of freedom (π_{ij}, *or equivalently* ρ_{ij}, that are free to vary) of a cxc contingency table with fixed marginal, by adding one more constraint due to the observed agreement proportion on the diagonal, the degrees of freedom (π_{ij} values actually free to vary) are:

$$c^2 - 2c + 1 - 1 = c^2 - 2c = c(c-2)$$

Then, the usual agreement conditions do not allow determining the cell probabilities in the case of square contingency tables with more than two rows (columns).

Common correlation model (ccm) extensions

Since the contingency tables more than 2x2 are not determined, we imposed some assumptions on their correlation structure in order to extend the models already proposed and to calculate their pertinent sample sizes, to be compared with those available by the current literature.

Partial common correlation model (PCCM). Partial determination of the correlation structure by assuming a common correlation of the diagonal cells of the contingency table: It is sensible to assume that the agreement level of two raters in assigning the same subject to the same category is constant among the categories of the multinomial variable. So, the correlation coefficients among the cells on the principal diagonal will be the same, leading to: $\rho_{ii}=\rho$ for $i=1,2,\dots c$.. The further assumption of a common correlation coefficient ρ on the principal diagonal cells, allows writing the theoretical proportions of the observed agreement and of the agreement by chance as:

$$\pi_0 = \sum_{i=1}^{n}\pi_{ii} = \rho\sum_{i=1}^{n}\pi_{i\cdot}(1-\pi_{i\cdot}) + \sum\nolimits_{i=1}^{n}\pi_{i\cdot}^2 = \rho\left(\sum\nolimits_{i=1}^{n}\pi_{i\cdot} - \sum\nolimits_{i=1}^{n}\pi_{i\cdot}^2\right) + \sum\nolimits_{i=1}^{n}\pi_{i\cdot}^2 = \rho(1-\pi_e) + \pi_e$$

From which:

$$\rho = \frac{\pi_0 - \pi_e}{1 - \pi_e} = k$$

The above formula highlights that the correlation coefficient (ρ) of the probabilities of the cells on the principal diagonal corresponds to the Cohen's kappa, in analogy with what happens for the 2x2 contingency tables. The above formulation represents the first generalization to cxc square contingency tables of the "common correlated model for dichotomous variables".

Then, it is possible to calculate the joint probabilities (π_{ij}) on the principal diagonal given by:

$$\pi_{ii} = \rho\pi_{i\cdot}(1-\pi_{i\cdot}) + \pi_{i\cdot}^2\ for\ i = 1, \dots c$$

In addition, these probabilities fulfill the appreciable property of respecting the order structure of the marginal probabilities; that is, if $\pi_{i\cdot} \leq \pi_{\cdot j}$, then $\pi_{ii} \leq \pi_{jj.}$

Thanks to this assumption, the probability values of the cells on the principal diagonal are obtained by considering the structure imposed by the marginal probabilities and the kappa value.

So, the not constrained joint probabilities are:

$c(c-2)-(c-1)=c(c-3)+1$

Of course, this common correlation for multinomial variables has some consequences on the sample size calculation that will be described later in the "result section".

Full Common Correlation Model (FCCM). Full determination of the correlation structure by assuming a common correlation of all cells of the contingency table: By adding a second more restrictive assumption on the correlation coefficients ρ_{ij} *with* $i \neq j$, we can completely determine the correlation structure of the contingency table.

Let's assume that the correlations ρ_{ij} with $i \neq j$ are the product of two components: $\rho_{ij} = \rho\gamma_{ij}$ one due to the raters, the constant ρ, already considered in the first assumption and the other (γ_{ij}) which expresses the correlation between the categories, considered pairwise, given by:

$$\gamma_{ij} = \frac{-\pi_{i\cdot}\pi_{\cdot j}}{\sqrt{\pi_{i\cdot}(1-\pi_{i\cdot})\pi_{\cdot j}(1-\pi_{\cdot j})}}$$

This second assumption is:

$$\rho_{ij} = -\rho\frac{\pi_{i\cdot}\pi_{\cdot j}}{\sqrt{\pi_{i\cdot}(1-\pi_{i\cdot})\pi_{\cdot j}(1-\pi_{\cdot j})}} \quad \text{for each } i \neq j$$

Then, substituting ρ_{ij} in the π_{ij} previous formula of the PCCM, we have:

$$\pi_{ij} = \rho_{ij}\sqrt{\pi_{i\cdot}(1-\pi_{i\cdot})\pi_{\cdot j}(1-\pi_{\cdot j})} + \pi_{i\cdot}\pi_{\cdot j} =$$

$$= (1-\rho)\pi_{i\cdot}\pi_{\cdot j} \quad for\ i \neq j$$

So, all cell probabilities are determined:

$$\pi_{ij} = \begin{cases} \rho\pi_{i\cdot}(1-\pi_{i\cdot}) + \pi_{i\cdot}^2 & for\ i = j \\ (1-\rho)\pi_{i\cdot}\pi_{\cdot j} & for\ i \neq j \end{cases}$$

And by keeping the previously obtained results, we obtain:

$$\rho = \frac{\pi_0 - \pi_e}{1-\pi_e} = k$$

So, with this full model, obtained under a further restrictive condition in comparison to the first, it is possible to calculate the kappa variance, necessary for the sample size calculation, without resorting to LP. Furthermore, it is possible to demonstrate that also the weighted kappa is equal to the common correlation coefficient.

Indeed, remembering that:

$$\pi_O = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\pi_{ij}\ and$$

$$\pi_e = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\pi_{i\cdot}\pi_{\cdot j}$$

In the case of the complete common correlation model, we have:

$$\pi_{ij} = \begin{cases} \rho\pi_{i\cdot}(1-\pi_{i\cdot}) + \pi_{i\cdot}^2 & for\ i = j \\ (1-\rho)\pi_{i\cdot}\pi_{\cdot j} & for\ i \neq j \end{cases}$$

So, we obtain:

$$\pi_O = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\pi_{ij} =$$

$$= \sum_{i}^{c}[\rho\pi_{i\cdot}(1-\pi_{i\cdot}) + \pi_{i\cdot}^2] + \sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c} w_{ij}(1-\rho)\pi_{i\cdot}\pi_{\cdot j} =$$

$$= \rho\left[\sum_{i}^{c}\pi_{i\cdot}(1-\pi_{i\cdot}) - \sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c} w_{ij}\pi_{i\cdot}\pi_{\cdot j}\right]$$

$$+\sum_{i}^{c}\pi_{i\cdot}^2 + \sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c} w_{ij}\pi_{i\cdot}\pi_{\cdot j} =$$

$$= \rho\left[\sum_{i}^{c}\pi_{i\cdot} - \left(\sum_{i}^{c}\pi_{i\cdot}^2 + \sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c} w_{ij}\pi_{i\cdot}\pi_{\cdot j}\right)\right]$$

$$+\sum_{i}^{c}\pi_{i\cdot}^2 + \sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c} w_{ij}\pi_{i\cdot}\pi_{\cdot j}$$

Recalling that:

$$\sum_{i}^{c}\pi_{i\cdot}^{2}+\sum_{i=1}^{c}\sum_{j=1;i\neq j}^{c}w_{ij}\pi_{i\cdot}\pi_{\cdot j}=\sum_{i=1}^{c}\sum_{j=1}^{c}w_{ij}\pi_{i\cdot}\pi_{\cdot j}=\pi_{e}$$

We obtain:

$$\pi_0=\rho(1-\pi_e)+\pi_e$$

And, consequently:

$$\rho=\frac{\pi_0-\pi_e}{1-\pi_e}=k_w$$

Finally, we have written a program in the open source R language that calculates the sample sizes according to Donner and Rotondi [92] for any number of raters without resorting to the functions, specific for each number of raters of the package "kappaSize" from Rotondi [94]. (Appendix D. R Code).

RESULTS

Variance

Variance of 3x3 contingency tables under the Partial Common Correlation Model: Under the usual conditions of a symmetrical weight (w_{ij}) matrix and of uniform marginal probabilities, the variance formula is unchanged if each probability in position (i,j) is replaced by the corresponding probability in the symmetrical position (j,i). So, given a probability matrix M and its transpose M^T, we obtain that: $VarK(M^T)=VarK(M)$, being VarK (M) the "unity kappa variance" of the matrix M. In addition, given a matrix M, it is always possible to obtain a symmetrical matrix $(\pi_{ij}=\pi_{ji})$ that keeps the same marginal, the same π_0 and the same kappa:

$$M_{Sim}=(M+M^T)/2$$

By using the Fleiss et al.'s [50] formula, the unity kappa variance of this matrix is:

$$VarK(M_{Sim})=\frac{1}{2}VarK(M)+\frac{1}{2}VarK(M^T)=VarK(M)$$

So, in conclusion, the symmetrical matrix $(M_{Sim}\ with\ \pi_{ij}=\pi_{ji})$ keeps the same marginal, the same π_0 and the same kappa and, in addition, has the same unity kappa variance as the matrix M.

By applying the above outlined symmetry condition to the 3x3 probability tables that under the agreement condition of the PCCM have only one degree of freedom (only one cell is free to vary), we obtain that all cells are fixed, as it is shown in the following equation in which the term between the squared brackets is the number of the constraints imposed by the symmetry.

$$gl=c(c-3)+1-[\frac{(c-1)(c-2)}{2}]$$
$$=\frac{c(c-3)}{2}$$

Then, for c = 3 the degrees of freedom are zero.

So, it exists only one 3x3 symmetrical matrix fulfilling all agreement conditions. Consequently, in this case, it is possible to obtain, without resorting to LP, the unique values of the variance under the null (H_0) and the alternative hypothesis (H_A), respectively, to be inserted in the sample size calculation formula.

Variance for uniform cxc contingency table: It is possible to demonstrate that, in the case of uniform marginal, the unity variance of the unweighted kappa is unique. So, it is possible to use the same procedure above outlined for the 3x3 tables in order to calculate the pertinent sample sizes.

Indeed, starting from the Fleiss et al. [50] formula of the unweighted kappa:

$$Var(k)$$
$$=\frac{1}{N(1-\pi_e)^4}\left\{\begin{matrix}\sum_{i=1}^{k}\pi_{ii}[(1-\pi_e)-(\pi_{i\cdot}+\pi_{\cdot i})(1-\pi_0)]^2\\+\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k}\pi_{ij}[(\pi_{i.}+\pi_{,j})(1-\pi_0)]^2\\-(\pi_0\pi_e-2\pi_e+\pi_0)^2\end{matrix}\right\}$$

Since the uniform marginal, we have: $\pi_{i\cdot}=\pi_{\cdot j}=\pi=1/k$ for each i,j.

Remembering that:

$$\sum_{i=1}^{k}\pi_{ii}=\pi_0$$

And that

$$\sum_{i=1}^{k}\sum_{j=1,i\neq j}^{k}\pi_{ij}=1-\pi_0$$

And by considering each term within the curly brackets, we can write:

$$A=\sum_{i=1}^{k}\pi_{ii}[(1-\pi_e)-(\pi_{i\cdot}+\pi_{\cdot i})(1-\pi_0)]^2=$$

$$= [(1 - \pi_e) - 2\pi(1 - \pi_0)]^2 \sum_{i=1}^{k} \pi_{ii} \ =$$

$$= [(1 - \pi_e) - 2\pi(1 - \pi_0)]^2 \cdot \pi_0$$

$$B = \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \pi_{ij}[(\pi_{i.} + \pi_{.j})(1 - \pi_0)]^2 =$$

$$= [2\pi(1 - \pi_0)]^2 \sum_{i=1}^{k} \sum_{j=1, i \neq j}^{k} \pi_{ij} =$$

$$= [2\pi(1 - \pi_0)]^2 (1 - \pi_0) = 4\pi^2(1 - \pi_0)^3$$

$$C = (\pi_0\pi_e - 2\pi_e + \pi_0)^2$$

So, it is possible to notice that the above defined terms as A, B and C depend only on the parameters π_0, π_e *and* π, that, in the agreement context are fixed "a priori". So, also the unity kappa variance is fixed, since it does not depend on any of the joint probabilities (π_{ij}).

Sample size calculation

Preliminary considerations: The sample sizes have been calculated according to Flack et al. [74] (SS-Flack) with the maximum value of the kappa variance and also with the minimum value of the kappa variance (SS-Flack-min), to Altaye et al. [87,89] and Donner and Rotondi [92,94] (SS-Donner), to our method (A&C) based on the multinomial partial common correlation model both for a maximum and a minimum value of the kappa variance (SS-A&C-max, and SS-A&C-min), and, finally, to our method (A&C-full) based on the multinomial full common correlation model leading to have only one value of the kappa variance (SS-A&C-full). We considered four different scenarios of the rows (columns) marginal: from the uniform case to a noticeable non-uniform pattern. In addition, we considered eighteen null and alternative hypotheses given by three values of k_0 = 0.4, 0.5, and 0.6 and respective k_A values given by k_0 + 0.05, k_0 + 0.10, to 0.90 with steps of 0.10, and, finally, with k_A = 0.90 + 0.05. Finally, we considered a significance level of α= 0.05 one-tailed and two tailed, and power = 0.80. So, there are seventy two scenarios of sample size calculations for each statistical significance level.

Results about the sample size calculation will be shown by comparing SS-Flack and SS-Donner, SS-Flack and SS-A&C-max and SS-A&C-min, SS-A&C-max and SS-A&C-min and, then, SS-A&C-max, SS-A&C-min and SS-A&C-full. The comparison between SS-Flack and SS-Flack-min has been reported just for sake of completeness, being the respective values so different. The Flack et al.'s approach [74] and the PCCM model do not allow to uniquely determine the cell probabilities and, consequently, to have only one value of the kappa variance. However, by resorting to LP, it is possible to calculate the maximum and the minimum value of the kappa variance, and consequently a maximum and a minimum value sample size. However, Flack et al.'s approach [74] has the wider variance interval and, obviously, also the wider sample size interval. Otherwise, PCCM model gives much narrower intervals that, in addition, are within the Flack et al.'s intervals [74].

FCCM allows to uniquely determining the cell probabilities, and consequently the kappa variance and the sample size. In addition, the sample size is equal or within the sample size values calculated according to the PCCM with the minimum and maximum value of the kappa variance and only within those calculated according to Flack et al. [74] with the minimum and maximum value of the kappa variance.

In the case of uniform marginal probabilities, Flack et al.'s [74], PCCM, and FCCM approaches uniquely determine the kappa variance. So, the sample size is unique and the same for the three models. Regarding the PCCM for 3x3 tables, we demonstrated that the kappa variance is unique and, consequently, there is only one sample size value that is, in addition, equal to the sample size obtained with FCCM. In the case of uniform marginal probabilities, SS-Donner are always greater than the sample sizes calculated according to the other methods, perhaps owing to the fact that they are calculated for assessing a condition of agreement or not instead of a differentiated agreement assessed by the other procedures. The case of marginal non-uniform probabilities gives diversified results that will be detailed later.

Sample sizes. Contingency tables 2x2: Table SS1 shows the sample sizes for the case of two raters and 2x2 contingency tables at a significance level of α=0.05 one-tailed and two-tailed, respectively and power =0.80. Particularly, Table SS1 shows the k_0 and k_A values, the rows (columns) marginal and the four previously reported sample sizes (SS-Flack, SS-Donner, SS-A&C-max, SS-A&C-min, and SS-Flack-min).

After having found the required null (k_0) and alternative (k_A) hypothesis on the first two columns on the left, it has to select the appropriate pattern of the rows (columns) marginal on the next two columns on the right; then, it is possible to read for α= 0.05 one-tailed and two tailed, the above reported five sample sizes (SS-Flack, SS-Donner, SS-A&C-max, SS-A&C-min, and SS-Flack-min). So, it is possible to choose among the SS-Flack, SS-A&C-max, and SS-A&C-min the sample size more suitable for the actual feasibility of the agreement study. Indeed, SS-Donner has to be considered only in the "agreement or not" case. Even if SS-Flack-min values are extremely more appealing, their choice is not recommended since they are at their minimum, leading to obtain lower and even much lower empirical power values. The SS-A&C-full calculated according to the FCCM are not been reported in this table and in the other sample size tables since they are equal to SS-A&C-max or less than a few units; furthermore, a program in the free source language R for their calculation has been reported in Appendix D.

Comparison between SS-Flack and SS-Donner: By combining the two significance levels for a total of 144 cases, the comparison between SS-Flack and SS-Donner gives a difference mean of 2.71 (31.45, for the absolute difference), median of -12 (14, for the absolute difference) and a mean absolute percent difference between 14.87 and 20.11% depending on whether the minimum or the maximum sample size value is placed in the denominator. SS-Flack are equal in one case (0.7%), greater in 22 cases (15.3%) and lower in the remaining 121 ones (84%). The above outlined differences increase from the one-tailed to the two-tailed significance level.

Particularly, the comparison between SS-Flack and SS-Donner gives a difference mean of 0.83 (28.17, for the absolute difference), median of -11.5 (13, for the absolute difference) with a mean percent absolute difference between 15.67-21.58% in the case of α = 0.05 one-tailed and a difference mean of 4.59 (34.74, for the absolute difference), median of -13 (14.5, for the absolute difference) with a mean percent absolute difference between 14.07-18.65% in the case of α= 0.05 two-tailed.

Then, considering the increasing non-uniformity, it has to be stressed that in the case of uniform marginal and of the 0.6, 0.4 symmetrical marginal, SS-Flack are always lower than SS-Donner. Otherwise, SS-Donner are lower in the case of a greater non-uniformity (0.8, 0.2 and 0.9, 0.1), and of the greatest sample sizes for the hypotheses with k_0= 0.4 and k_A = 0.45 or k_A = 0.50, k_0= 0.5 and k_A = 0.55 or k_A = 0.60, and k_0= 0.6 and k_A = 0.65.

In the case of uniformity, the comparison between SS-Flack and SS-Donner gives a difference mean of -19.19 (19.19, for the absolute difference), median of -12.5 (12.5, for the absolute difference) with a percent difference mean between 16.48-22.43%. Of course, being the mean of the difference and the mean of the absolute difference the same, all SS-Flack are lower than SS-Donner.

Then, considering the degree of non-uniformity of the marginal, there is a difference mean of -17.72 (17.72, for the absolute difference), median of -13.5 (13.5, for the absolute difference) with a percent difference mean between 16.42-22.61% in the case of a little non-uniformity with marginal equal to 0.6 and 0.4. In this case 8 SS-Flack values out of 36 (22.2%) are greater than SS-Donner ones and one value is the same.

The differences are lower in the case of the relevant non-uniformity of 0.8 and 0.2, being the difference mean of 1.39 (20.44, for the absolute difference), median of -10.0 (13.0, for the absolute difference) with a mean percent absolute difference between 13.89-18.84%.

Finally, in the case of an extreme non-uniformity (0.9, 0.1), the difference means are greatest with a difference mean of 46.39 (68.44 for the absolute difference), median of -11.0 (20.0, for the absolute difference) with a mean percent absolute difference between 12.69 and 16.56%. In this case, fourteen SS-Flack values out of 36 (38.9%) are greater than SS-Donner ones. The above results hold also for SS-Flack-min, since for the 2x2 contingency table, the variance is unique.

Comparison between SS-Flack (SS-Flack-min) and SS-A&C-max (SS-A&C-min): SS-Flack and SS-A&C-max are always the same, being unique the "unity" kappa variance in the case of 2x2 contingency tables. Hence, SS-Flack-min and SS-A&C-

min are also the same, and so it is between SS-Flack and SS-Flack-min and between SS-A&C-max and SS-A&C-min.

Comparison between SS-A&C-max and SS-Donner: Of course, the comparison between SS-Donner and SS-A&C-max (SS-A&C-min) gives the same results as that between SS-Flack and SS-Donner.

Comparison between SS-A&C-max, SS-A&C-min and SS-A&C-full: Since the probability table 2x2 is unique, SS-A&C-max, SS-A&C-min and SS-A&C-full are all the same.

Sample sizes. Contingency tables 3x3: Table SS2 shows the sample size results for 3x3 contingency tables at a significance of 0.05 one-tailed and two-tailed, respectively.

Comparison between SS-Flack and SS-Donner: By combining the two significance levels for a total of 144, the comparison between SS-Flack and SS-Donner gives a difference mean of 25.37 (36.00, for the absolute difference), median of -4 (8, for the absolute difference) with a percent difference mean between -8.79 and -12.23% (13.38 and 17.33 for the percent absolute difference) depending on whether the minimum or the maximum sample size value is placed in the denominator. SS-Flack are greater in 45 cases (31.25%), equal in three cases (2.08%) and lower in the remaining 96 ones (66.67%).

The differences increase from one-tailed case with mean of 21.38 (median -4, mean of the percent difference from -9.85 to -14.04%), and mean of the absolute difference of 31.63 (median 8, mean of the percent difference of 14.25 to 18.92%) to the two-tailed case with mean of 29.38 (median -8.5, mean of the percent difference from -7.72 to -10.42%), and mean of the absolute difference of 40.38 (median 9, mean of the percent difference from 12.51 to 15.74%).

In the case of uniform marginal, SS-Flack are always lower than SS-Donner. Otherwise, in the case of non-uniformity increasing and of the greatest sample sizes for the cases with k_0= 0.4 and k_A = 0.45 or k_A = 0.50, k_0= 0.5 and k_A = 0.55 or k_A = 0.60, and k_0= 0.6 and k_A = 0.65, SS-Donner are lower. Particularly, in the case of uniformity, the difference mean is -10.94 (median -8, mean of the percent difference ranging from -15.7 to -21.48) and the mean of the absolute and percent differences are: 10.94, 15.7% and 21.48% (median of 8).

Then, for the first step of non-uniformity (0.50, 0.25, 0.25), the difference mean is 13.39 (median -3, mean of the percent difference ranging from -9.82 to -13.54) and the mean of the absolute differences is of 20.11 (median of 6; mean of the percent absolute difference ranging from 12.81% to 16.71%). The differences increase at the second non-uniformity step (0.60, 0.30, 0.10), being the difference mean 53.22 (median 1, -2.71% and -4.15%) and the absolute mean 57.94 (median 7, 12.67% and 15.52%). The differences decrease at the greatest non-uniformity (0.80, 0.10, 0.10) since the difference mean is 45.83, (median -2.5, -6.92 and -9.75%) together with the absolute mean of 55.0 (median 11, 12.33% and 15.63%). Of course, the above differences are much more relevant if we consider SS-Flack-min.

Comparison between SS-Flack and SS-A&C-max: SS-Flack and SS-A&C-max are equal in the case of uniformity of rows (columns) marginal. Otherwise, SS-Flack are greater in 104 cases (72.22%) except for the four cases of k_0= 0.4 and k_A = 0.90 or k_A = 0.95, k_0= 0.5 and k_A = 0.95, and k_0= 0.6 and k_A = 0.95 at α = 0.05 one-tailed, in which the very small sample sizes are the same. Globally, the difference mean is 27.92 (median 2, percent difference meanfrom5.43% to 6.05%) with the mean of the absolute differences of 27.93 (median 2, percent difference meanfrom5.44% to 6.05%). As usual, the differences increase from the one-tailed case with mean of 24.4 (median 2, 5.20% and 5.80%) to the two-tailed case with mean of 31.4 (median 3, 5.67% and 6.31%). For these sample sizes the raw and absolute differences are the same since SS-Flack are greater than or equal to SS-A&C-max. In the cases of the non-uniformity, but with two equal marginal (0.50, 0.25, 0.25, and 0.80, 0.10, 0.10), difference means are of 22.36 and 29.83 (median 2.5 and 4, respectively) with percent difference mean of 5.20-5.54% and of 4.34-4.56% depending on the denominator, respectively.

Greater differences have been found in the case of a greater non-uniformity with all three marginal different (0.60, 0.30, 0.10) with mean of 59.5 (median 8) and means of the percent difference from 12.19 to 14.11%.

Of course, the sample sizes calculated by using the minimum kappa variance value (SS-Flack-min) are much lower than those calculated under the approach based on the maximum kappa variance value(SS-Flack). Indeed, these sample sizes have been shown in the tables just for giving the idea of the results obtained at the opposite extreme side of the extremely conservative approach proposed by Flack et al. [74]. Actually, it is sensible to think that the “optimum sample size” is between these two extreme values, perhaps as their mean.

Particularly, the comparisons between SS-Flack and SS-Flack-min give a difference of 0 in the case of marginal uniformity, as a further confirmation of the uniqueness of the probability matrix in this case with, consequently, a unique variance value. Otherwise, SS-Flack are always greater than SS-Flack-min with a percent difference of 35.59% (min = 0, Q1 = 9.38%, median = 36.47%, Q3 = 48.33%, and max = 93.36%).

The differences increase from the one-tailed case with mean 178.9 (median of 13.5) and percent difference means of 35.23% and 128.9% to the two-tailed case with mean 229.1 (median 18.5) and percent difference means of 35.95% and 138.97%. As usual, the differences increase as the non-uniformity increases: in fact, mean values are of 112.5 (median 15, percent differences are from 29.34% to 42.73%, depending on the maximum or minimum value placed in the denominator) in the case of marginal equal to 0.50, 0.25, 0.25, of 200.6 (median 29, percent differences are from 40.9% to 69.9%) in the case of marginal equal to 0.60, 0.30, 0.10, and of 503.2 (median 72, from 72.1% to 423.2%) in the case of marginal equal to 0.80, 0.10, 0.10. The comparison between SS-Flack and SS-A&C-min is the same as the comparison between SS-Flack and SS-A&C-max, since in the case of 3x3 tables there is a unique variance value under the common correlation multinomial model for calculating SS-A&C and, consequently SS-A&C-max are equal to SS-A&C-min.

Comparison between SS-Donner and SS-A&C-max: The comparison between SS-Donner and SS-A&C-max (or SS-A&C-min, being the two SS-A&C all the same) gives a mean of the absolute difference of 14.45 (median 8, mean percent of 14.4 and 19.4%) and a difference mean of 2.54 (median 7, 13.8-18.8%), being 128 SS-Donner values (88.9%) greater than SS-A&C-max, fifteen lower and one equal in the case of 0.6, 0.3, 0.1 with $k_0 = 0.5$ and $k_A = 0.55$, at $\alpha = 0.05$ one-tailed. The (absolute) differences increase a little from the one-tailed case with a mean 3.07 (median 6.5, 14.67% and 20.4%) together with means of the absolute and percent differences of 13.15, 15.21% and 20.96% (median 8) to the two-tailed case with mean of 2.03 (median 7, 12.98% and 17.32%) together with mean absolute and percent differences of 15.8, 13.57% and 17.94%.

In the case of marginal uniformity, there is a difference mean of 10.9 (median 8), percent difference mean from 15.7% to 21.5%; being all SS-Donner greater than SS-A&C-max, raw and absolute differences are the same.

The differences are smaller for the case of 0.50, 0.25, 0.25 marginal with difference mean of 8.9 (median of 4), percent difference mean from 14.62% to 19.6%, depending on the denominator; being all SS-Donner greater than SS-A&C-max the raw and absolute differences are the same.

The differences are, again, smaller for the case of 0.60, 0.30, 0.10 marginal with difference mean of 6.3 (median 5, 14.2% and 19.3%) together with the mean of the absolute differences of 9.2 (median 5) and mean percent absolute difference from 14.3 to 19.5%. Finally, the greatest differences are for the case of 0.80, 0.10, 0.10 marginal in which there are some SS-Donner lower than SS-A&C-max; indeed, the difference mean is of -16.0 (median 6, 10.8% and 14.9%) together with a mean of the absolute differences of 28.7 (median 11) and means of the absolute percent differences of 12.9-17.2%.

Comparison between SS-A&C-max, SS-A&C-min and SS-A&C-full: Since the probability table 3x3 is unique, SS-A&C-max, SS-A&C-min and SS-A&C-full are all the same.

Sample sizes. Contingency tables 4x4: Table SS3 shows the sample size results for the case of two raters and 4x4 contingency tables at a significance level of α=0.05 one-tailed and two-tailed, respectively.

Comparison between SS-Flack and SS-Donner: By combining the two significance levels for a total of 144 cases, the comparison between SS-Flack and SS-Donner gives a difference mean of 38.79 (45.53, for the absolute difference), median of -2 (6, for the absolute differences) with a mean percent of the absolute differences between 13.38% and16.69%.SS-Flack are greater in 58 cases (40.28%), equal in three (2.08%) and lower in the remaining83 (57.6) ones. The above outlined differences are greater for the two-tailed significance level. Particularly, the comparison between SS-Flack and SS-Donner gives a difference mean of 33.42 (39.89, for the absolute differences), median of -2.0 (6, for the absolute difference) with percent difference means of -4.94% and -7.1% (13.9% and 17.6% for the means of the percent absolute differences) in the case of α = 0.05 one-tailed. There is a difference mean of 44.18 (51.18, for the absolute difference), median of -2 (7 for the absolute difference) with a percent difference mean of -6.24% and -4.58% (12.79% and 15.79% for the absolute percent differences) in the case of α= 0.05 two-tailed.

In the 36 cases of uniform marginal SS-Flack are always lower than SS-Donner; in addition, SS-Flack are lower in another 47 cases (total of 83, 57.6%). Then, in the case of non-uniformity and of the greatest sample sizes for k_0= 0.4 and k_A = 0.45 or k_A = 0.50, k_0= 0.5 and k_A = 0.55 or k_A = 0.60, and k_0= 0.6 and k_A = 0.65 SS-Donner are lower in 58 (40.3%) cases. This pattern occurs at greater differences between k_0 and k_A the more rows (columns) non-uniformity increases ranging from 0.40, 0.30, 0.20, 0.10 to 0.70, 0.10, 0.10, 0.10. Finally there are three cases of equality when k_0= 0.4 or 0.5 or 0.6 and k_A = 0.80.

Particularly, in the case of uniformity, the mean is -8.03 (median -6, mean of the percent difference ranging from -11.5 to -19.23; being all SS-Flack lower than SS-Donner, the raw and absolute differences are the same.

Then, for the first step of non-uniformity (0.40, 0.30, 0.20, 0.10) the mean is 49.97 (median 2, mean of the percent difference ranging from 13.48 to 15.59) and the mean of the absolute and percent differences are: 52.58, (median of 5 with the mean of the percent absolute difference ranging from 12.9% to 15.75%. The differences increase at the second non-uniformity step (0.60, 0.20, 0.10, 0.10), being the mean 67.56 (median 3, 1.59% and 1.56%) and the absolute mean 70.5 (median 7, 13.56% and 16.52%).

The differences decrease at the greatest non-uniformity (0.70, 0.10, 0.10, 0.10) since the mean is 45.69, (median 5.0, -4.13 and -5.97%) together with the absolute difference mean of 51.0 (median 7.5, 12.47% and 15.28%).

Comparison between SS-Flack and SS-A&C-max: SS-Flack and SS-A&C-max are equal in the 36 (25.0%) cases of uniformity of rows (columns) marginal. Otherwise, SS-Flack are always greater (108 cases; 75.0%) with a difference mean of 40.56 (median = 3) and means of the percent difference of 9.29 and 10.85%. Being SS-Flack always greater (or equal) than SS-A&C-max, the raw and absolute differences are the same.

The differences increase from the one-tailed to the two-tailed significance level.

Particularly, from a mean (row and absolute) of 35.53 (median 3) and means percent difference of 9.21 and 10.74% for the one-tailed case, to a mean (row and absolute) of 45.58 (median 4) and means percent difference of 9.37 and 10.96% for the two-tailed case. In the case of uniformity, SS-Flack and SS-A&C-max (SS-A&C-min) are equal.

In addition, the differences increase from the lower non-uniformity (0.40, 0.30, 0.20, 0.10) with a difference mean of 55.4 (median of 7.5 and percent difference meanof14.84-17.63%) to the intermediate non-uniformity of 0.60, 0.20, 0.10,10 with a difference mean of 67.3 (median of 9.0 and percent difference mean ranging from 14.71 to 17.43%); then, there is a decrease with a difference mean of 39.5 (median of 5.5 and percent difference meanfrom 7.63 to 10.08%) in the case of marginal equal to 0.70, 0.10, 0.10 and 0.10. Being SS-Flack greater than or equal to SS-A&C-max (SS-A&C-min), the comparison between SS-Flack and SS-A&C-min are similar to the comparison between SS-Flack and SS-A&C-max, since the difference between SS-A&C-max and SS-A&C-min is very limited, being the mean of the differences equal to 1.12 (min, Q1, median, and Q3 equal to 0 and max = 19.0). Of course, SS-Flack-min, calculated by using the minimum kappa variance value, are much lower than those calculated under the approach based on the maximum kappa variance value (SS-Flack), apart from the case of the uniformity in which the sample sizes are the same. Indeed, the difference mean is 177.8 (min = 0, Q1= 1.50, median = 15.0, Q3 = 103.0, and max = 2223.0).

Comparison between SS-Donner and SS-A&C-max: The comparison between SS-Donner and SS-A&C-max gives a difference mean of 1.76 (median 5, 12.92%-17.38%) and a mean absolute difference of 10.12 (median of 6, 13.5 – 17.98%), being 122(84.7%)SS-Donner greater than SS-A&C-max, 19 (13.2) lower, and three equal in the case of 0.7, 0.10, 0.1, 0.10 with k_0 = 0.4 and k_A = 0.60 in the one-tailed and two-tailed case and in the case of 0.4, 0.3, 0.2, 0.1 with k_0 = 0.4 and k_A = 0.50 at α = 0.05 two-tailed.

The differences decrease from the one-tailed to the two-tailed significance level. Particularly, from a mean (row and absolute) of 2.11 (median 5) and a percent difference mean from 13.35

to 18.60% for the one-tailed case, to a mean (row and absolute) of 1.40 (median 6) and a percent difference mean from 12.18 to 16.14% for the two-tailed case. In the case of uniformity, SS-Donner and SS-A&C-max (SS-A&C-min) differ of a mean of 8.3 (median of 6, percent from 14.47% to 19.23%). In addition, at the lower non-uniformity (0.40, 0.30, 0.20, 0.10) there is a difference mean of 5.47 (median of 5 and percent difference mean ranging from 13.64 to 18.24%); then, for the intermediate non-uniformity of 0.60, 0.20, 0.10,10 there is a decrease with a mean of -0.27 (median 4.5, percent difference means of 14.41 and 16.91%).There is a further decrease with a difference mean of -6.19, (median of 5 and percent difference mean ranging from11.16 to 15.12%)in the case of marginal equal to 0.70, 0.10, 0.10 and 0.10. Being SS-A&C-max very similar to SS-A&C-min, as already reported, the differences between SS-Donner and SS-A&C-min are practically the same.

Particularly, considering all 144 results together, the difference mean is 2.88 (median of 6; 13.24% - 17.37%) together with a mean of the absolute differences of 10.0 (median of 10.74; 13.73% to 18.32%), being SS-Donner greater than SS-A&C-min in 126 cases (87.5%), lower in only 16 cases (11.11%) and equal in two cases (1.39%) with marginal of 0.7, 0.10, 0.1, 0.10 with k_0 = 0.4 and k_A = 0.60 in the one-tailed and two-tailed case.

The differences are practically the same for the one-tailed case and the two-tailed case: indeed, there is a mean of 3.08 (median 5, 13.93-19.02%) with mean absolute and percent difference of 9.11 (median 6), means percent of 14.2% and 19.18% for the one-tailed case and a mean of 2.67 (median 6, 12.54%-23.67%) with mean absolute and percent difference of 10.89 (median 6.5), means percent of 13.05% and 17.1% for the two-tailed case. The differences are greater for the case of marginal uniformity with difference mean of 8.03 (median 6), percent difference mean from 14.5 to 19.2%, depending on the denominator; being SS-Donner always greater than SS-A&C-max raw and absolute differences are the same. Then, the differences decrease for the first level of non-uniformity, with difference mean of 7.64 (median 6) and percent difference mean from 12.13 to 13.81%.

A further decrease occurs at the second non-uniformity level, with difference mean of 2.03 (median 5) and percent difference meanfrom 13.13 to 18.02%. Finally, for the case of 0.70, 0.10, 0.10, 0.10 marginal the differences are equal to those for SS-A&C-max, being the two sample sizes the same.

Comparison between SS-A&C-max and SS-A&C-min: Particularly, combining the 144 results together, the difference mean is 1.11 (median of 0, 0.36% and 0.38%), being SS-A&C-max greater than or equal than AA-A&C-min, differences and absolute differences are the same. SS-A&C-max are equal to SS-A&C-min in 109 (75.7%) and greater in 35 (24.3%) cases; particularly, they are equal in the uniformity case (0.25, 0.25, 0.25, 0.25) and in the case of marginal equal to 0.7, 0.1, 0.1, 0.1. The differences increase a little from the one-tailed case with mean of 0.97 (median 0, 0.33% and 0.34%) to a mean of 1.26 (median 0, 0.29% and 0.29%) in the two-tailed case. Finally, the differences are very similar in the two non-uniformity cases: a mean of 2.17 (median 0, 0.58% and 0.59%) for the pattern of 0.4.0.3.0.2.0.1 and mean of 2.31 (median 0, 0.86% and 0.91%) for the pattern 0.6, 0.2, 0.1, 0.1. However, the differences are greater at the lowest differences between k_0 and k_A.

Comparison between SS-A&C-max (SS-A&C-min) and SS-A&C-full: SS-A&C-full is lower than SS-A&C-max with a mean difference between SS-A&C-max and SS-A&C-full of 0.85 (min, Q1, median, and Q3 equal to zero and max = 15.0) with a mean of the percent differences of 0.31% in 32 (22.2%)cases and equal in the remaining 112 (77.8%) cases. Furthermore, SS-A&C-full is greater than SS-A&C-min in 20 (13.9%) cases and equal in the remaining 124 (86.1%) cases. Particularly, the mean difference between SS-A&C-min and SS-A&C-fullof -0.26 (min = -5, Q1, median, Q3, and max equal to 0) with a mean of the percent differences of -0.045%. As an important consequence of the previously reported “common correlation model for the diagonal cells of the contingency tables for multinomial variables”, it has to be stressed that the interval between the maximum and minimum variance value is very narrow, being equal to zero for the 2x2 and 3x3 contingency tables. Then, in the case of the 4x4 contingency tables, the difference mean is of 1.12 (min, Q1, median, Q3 equal to 0 and max = 19.0), leading to a maximum difference between the minimum and maximum sample size values ranging from 0 and 19. The maximum of 19 occurred in the case of k_0 = 0.4, k_A= 0.45, marginal probabilities of 0.60, 0.20, 0.20 a = 0.05 two-tailed and power = 0.80. Generally, difference of more than ten occurred when the difference between k_0 and k_A is only of 0.05. As a conclusion, the maximum and minimum sample sizes calculated under this model are practically the same with differences of only a few units and are lower that the sample sizes calculated according to Flack et al. [74].

Furthermore, the sample size calculated under the "full correlation model" (SS-A&C-full) are always within the SS-A&C-max and SS-A&C-min, and allows to have a unique contingency table pattern without recurring to the LP procedure. Finally, considering the 4x4 contingency table, the differences between these sample sizes, pair wise considered, are in the 77.8% equal to SS-A&C-max and in the 86.1% equal to SS-A&C-min.

Influence of the non-uniformity of the marginal and of considering the weighted kappa instead of the unweighted kappa on the required sample size under the same null and alternative hypothesis: As a further and relevant point, there is the influence of the non-uniformity on the required sample size.

The following table considers for a 3x3 contingency table, two testing hypotheses at α = 0.05 one tailed with k_0 = 0.6 and k_A = 0.8 and k_0 = 0.4 and k_A = 0.45 for seven pattern of the rows (columns) marginal ranging from the uniformity to a very extreme non-uniformity. Finally, there are three sample sizes: SS-Flak for the unweighted kappa (n-W), for the weighted kappa with linear weights (W-Linear), and for the weighted kappa with quadratic weights (W-Quadr.).

π1	π2	π3	k_0 = 0.6 and k_A = 0.8			k_0 = 0.4 and k_A = 0.45		
			SS-Flack			SS-Flack		
			n-W.	W-Linear	W-Quadr.	n-W.	W-Linear	W-Quadr.
0.333	0.333	0.333	75	116	162	1,679	2,653	3,896
0.500	0.300	0.200	89	122	172	2,159	2,689	3,996
0.600	0.300	0.100	105	139	217	2,618	2,913	4,473
0.700	0.200	0.100	123	158	223	2,960	3,442	4,713
0.800	0.100	0.100	162	200	249	3,677	4,407	5,567
0.900	0.050	0.050	302	359	442	6,653	8,137	10,058
0.990	0.005	0.005	2,848	3,287	4,011	60,858	75,027	90,961

It has to be noted that the sample size (SS-Flack) from the uniform pattern increases of the 3,692% for the first hypothesis and of 3,513% for the second one.

Furthermore, it has to be noted that the weighted kappa requires greater sample sizes than the unweighted one, and that the quadratic weighting scheme is even more demanding. Particularly, the mean percent increase of the W-Linear is of 30.05% (min = 15.41%, Q1 = 18.87%, median = 28.46, Q3 = 37.08%, max = 54.67%) for the first hypothesis and of 25.08% (min =11.26, Q1 = 16.28%, median = 22.31%, Q3 = 24.55%, max = 58.01%) for the second hypothesis. In addition, the percent increase mean of the W-Quadr. is of 76.87% (min = 40.83%, Q1 = 46.36%, median 81.30%, Q3 = 106.67%, max = 116.00%) for the first hypothesis and of 71.32% (min = 49.46%, Q1 = 51.18% median = 59.22%, Q3 = 85.09%, max = 132.04) for the second hypothesis.

Finally, compared to W-Linear, the sample size of W-Quadr. increases of 35.36% (min = 22.03%, Q1 = 23.12, median = 39.65%, Q3 = 41.14%, max = 56.12%) for the first hypothesis and of 36.73% (min = 21.24%, Q1 = 23.61%, median = 36.93, Q3 = 48.61%, max = 53.55%) for the second hypothesis.

Simulation study

Firstly, it has to be stressed that the simulation studies on the kappa properties have to be planned with a well-defined variance-covariance matrix (π_{ij} have to be known) in addition to the rows (columns) marginal and the observed agreement proportion, given the Cohen's kappa value. It has to be noted that c(c-2) cells are "free", being not determined by the constraints given by the fixed rows (columns) marginal and by the kappa value. Of course, the simulation results will depend on how the π_{ij} "free"are determined. We simulated under the Flack et al.'s [74] condition of a maximum kappa variance value under which the π_{ij} "free" are fixed. So, the calculated sample size will be the maximum among all sample sizes calculated for contingency tables with the same marginal and observed agreement proportion. Consequently, this is the most conservative condition, leading to the rejection of the null hypothesis more than the expected. Of course, this situation is not in agreement with the principle that, in the biomedical research, the sample sizes have to be adequate for try to prove the primary objective of the research with a satisfactory probability level, but also as parsimonious as possible in according to ethical requirements. Finally, it has to be stressed that the aims of the simulation study, based on the multinomial distribution, are to assess the actual power of testing a kappa vs. an expected value and the actual coverage of the 95% confidence interval of kappa with so conservative sample sizes.

Our simulation study has been carried out for contingency tables 3x3 and 4x4 on 1,000 samples with their sample size calculated accordingly to Flack et al. [74] for a significance value of $\alpha = 0.05$ two-tailed and power of 0.80 or 0.90, under four scenarios of rows (columns) marginal (0.33, 0.33, 0.33; 0.50, 0.25, 0.25; 0.6, 0.3, 0.1;0.80, 0.10, 0.10), and six null and alternative hypotheses (k_0= 0.4 with k_A = 0.6, 0.7, and 0.8, and k_0 = 0.6 with k_A = 0.8, 0.9, and 0.95). In total, there are 24 different combinations for each power value. For the 4x4 table the four scenarios of the rows (columns) marginal were of 0.25, 0.25, 0.25, 0.25; 0.40, 0.30, 0.20, 0.10; 0.60, 0.20, 0.10, 0.10, and 0.70, 0.10, 0.10, 0.10.

Table SIM.1A for the 3x3 contingency table shows the descriptive statistics (mean, SD, 95% CI limits, Median, Minimum, First Quartile- Q1- Third Quartile - Q3 – Maximum) of the raw, percent, absolute, and absolute percent bias of the estimates of k_0 and k_A and of the coverage. Table SIM.1B shows the bias (raw, raw percent, absolute and absolute percent) of the power for the simulations with sample sizes calculated for a power of 0.80 and of 0.90, respectively. Values have been truncated to the sixth decimal figure. Of course, the significance results are for a P-value $\leq$ 0.25 at the right tail of the distribution, being the sample size calculated for a significance level of 0.05 two-tailed and taking into account that a P-value lower $\leq$0.25 at the left tail of the distribution corresponds to the Type III error or the “Sign error”.

In the case of a 3x3 contingency table, the absolute bias mean for k_0 and k_A combining together the two simulations with sample sizes calculated for a power of 0.80 and of 0.90, respectively, are 0.006874 and 0.003596, respectively (median 0.006766, and 0.002821, respectively). In addition, the percent absolute bias mean compared to the theoretical value are 1.393691% and 0.464093%, respectively (median of 1.415729%, and of 0.350172%, respectively). The empirical values of k_0 and k_A are always lower than their theoretical values.

The coverage mean bias is -0.018017 (median -0.017050) and the mean of the coverage percent bias is -1.896491 (median = -1.794736), with empirical values always lower than the theoretical value of 0.95. Furthermore, 50% of the values are lower than 0.933 with a minimum coverage value of 0.903.

In addition, the empirical power is always more than the theoretical value of 0.80 and in 22 (91.7%) cases is more than the theoretical value of 0.90; indeed, it is less than the theoretical value of 0.90 in only two cases (8.3%) with k_0 = 0.4 and k_A = 0.60 or 0.80 and marginal of 0.80, 0.10, 0.10.

Table SIM.2A and Table SIM.2B show the corresponding results for the simulation of the 4x4 table. The absolute mean bias for k_0 and k_A combining together the two simulations with sample sizes calculated for a power of 0.80 and of 0.90, respectively, are 0.0060711 and 0.012590, respectively (median 0.004646, and 0.002651, respectively). In addition, the mean percent absolute bias compared to the theoretical value are 1.252031% and 1.568907%, respectively (median 0.876641%, and 0.378128%, respectively).

The empirical k_0 values are lower than their theoretical values in 46 (95.8%) cases. In two cases (one in the simulation under a power of 0.80 and one under a power of 0.9) are greater. Particularly, k_0 = 0.402820 instead of 0.4 with a power of 0.80 and k_0 = 0.401112 instead of 0.4 with a power of 0.90 always in the case of marginal equal to 0.40, 0.30, 0.20, 0.10.

In addition, the empirical k_A values are lower than their theoretical values in 39 (81.3%) cases. In 9 cases (four in the simulation under a power of 0.80 and five under a power of 0.9) are greater. Particularly, for a power of 0.80, k_A = 0.90056and k_A= 0.94781 instead of 0.9with marginal equal to 0.25, 0.25, 0.25, 0.25 and 0.70, 0.10, 0.10, 0.10; for the other two cases, k_A = 0.89618 and = 0.89887 instead of 0.80 both with marginal equal to 0.70, 0.10, 0.10, 0.10. Furthermore, for a power of 0.90, k_A = 0.80005 and k_A= 0.90014 instead of 0.8 and 0.9, respectively with marginal equal to 0.40, 0.30, 0.20, 0.10; then, k_A = 0.89792 and k_A= 0.89816 instead of 0.8 and k_A = 0.9400 instead of 0.9 with marginal equal to 0.70, 0.10, 0.10, 0.10.

The mean of the coverage bias is -0.017562 (median -0.014000) and the mean of the coverage percent bias is equal to -1.925438 (median -1.631578). There is one coverage value equal to 0.95are five coverage values (10.4%) more than the theoretical value of 0.95 (3 for a power of 0.80 and 2 for a power of 0.90), always in the case of k_0 = 0.4 and k_A = 0.6. Particularly, in the case of marginal 0.25, 0.25, 0.25, 0.25, a value of 0.951 with a power of 0.80 and another value of 0.952 with a power of 0.90. In addition, a value of 0.955 in the case of marginal 0.40, 0.30, 0.20, 0.10 with a power of 0.80; then, with marginal 0.60, 0.20, 0.10, 0.10, values of 0.953 and 0.960 with power of 0.80 and of 0.90, respectively. In addition, the empirical power is always more than the theoretical value of 0.80 or of 0.90 with mean biases of 0.087333 and 0.041500 (median 0.067500 and 0.036500), respectively.

DISCUSSION

Firstly, it has to be said that the sample size calculation for an agreement study on qualitative variables to be analyzed with Cohen's kappa it is not a very easy task. Indeed, if the k_0 and k_A under the null and alternative hypotheses can be obtained by a careful search of the pertinent literature (k_0) and from the difference that can be considered as clinically relevant in the particular context of the agreement study (k_A), the subsequent step of fixing the prevalence of the categories on which the raters judgment has to be divided, it is likely to require a more laborious search of the pertinent literature. Then, given for granted that also the rows (columns) marginal have been sensibly found, the sample size calculation is still undefined since the contingency tables more than 2x2 cannot be uniquely defined unless further constraints are imposed. We have described the Falk et al. [74] condition of obtaining the contingency table with the maximum value of the kappa variance leading to a conservative sample size that it could be too much demanding compared to the actual feasibility of the agreement study.

For this case, we have shown a mathematical approach (linear programming) to be devised to obtain the maximum or the minimum value of the kappa variance or any other prefixed value (Appendix A: theoretical aspects and the R code) avoiding laborious attempts. Then, we have commented the Donner et al.'s suggestion [80] of calculating the sample size for an agreement study based on the kappa statistics in order that the lower limit of the kappa 95% confidence interval is greater than a required lower acceptable agreement margin. We have shown that the above result is not only not absolutely certain to be obtained but, also, to be obtained at a satisfactory adequate probability value, as usually it occurs for the sample sizes calculated on the precision of the confidence interval without considering the confidence interval power, as the probability of obtaining the required precision. So, this approach can be considered not shareable, unless the power of the confidence interval is taken into account by increasing the sample size calculated on the precision of the confidence interval or, equivalently, on obtaining a lower confidence limit greater than a prefixed agreement threshold until a satisfactory probability of obtaining the required precision result is achieved.

Then, we have shown that the Altaye et al. [87,89] and Donner et al.'s proposal [92,94] of calculating the sample size under the "goodness of fitting model" is generally too much demanding in terms of sample sizes and we have suggested that it could be taken into account in some limited cases with contingency tables 3x3 and 4x4 in which they are lower, but only if the "agreement or not" condition is acceptable and pertinent to the aims of the study. In addition, the statistical theory on which the "goodness of fitting model" is based and the Dirichlet multinomial distribution have been detailed in Appendix B. Then, in Appendix D has been reported a program written in the open source R language that allows to calculate the sample sizes for any number of raters, under the "goodness of fitting model" and the "agreement or not" condition, without resorting to specific functions each for a defined number of raters of the Rotondi's kappaSize package [94].

In addition, we have shown two other sample size calculation methods, based on the extension of the common correlation model for dichotomous variables. The first is the partial common correlation model with the relevant property of having equal or similar maximum and minimum values of the kappa variance, calculated by means of the LP procedure, leading to suppose that it has been obtained the less biased estimates of the true population variance; consequently, it could be argued that the "sample size as correct as possible" was calculated for the envisaged scenario of rows (columns) marginal and null and alternative hypothesis. Furthermore, we have refined our sample size calculation approach by formulating a second method, based on the full common correlation model, leading to obtain a unique probability table and only one value of the kappa variance without resorting to the LP procedure. A program written in the open source language R for calculating the sample sizes according to the full correlation model has been reported in Appendix E.

It has to be noted that in the case of 2x2 contingency tables, all our proposed methods (SS-A&C-max, SS-A&C-min, and SS-A&C-full) gave equal sample sizes which are, in addition, equal to those calculated according to Flack et al. [74]. SS-Donner are generally greater in the case of uniform marginal and sometimes lower, particularly in the case of the greatest calculated sample sizes; so, it is possible to argue that they can be used in these cases, under the "agreement or not" condition, in order to save resources without an important reduction of the power. Otherwise, in the case of 3x3 contingency tables, all our proposed methods (SS-A&C-max, SS-A&C-min, and SS-A&C-full) give the same sample sizes that are, in addition, lower than or, at maximum, equal to those obtained from the Flack et al.'s approach [74]. So, our methods may be recommended instead. In addition, SS-A&C-max are also almost always lower than SS-Donner, but, again in the about 11% of the cases in which these latter are smaller, they could be recommended under the "agreement or not "condition.

More diversified is the situation of the 4x4 contingency tables. However, as our sample sizes (SS-A&C-max, SS-A&C-min) may be recommended since they are smaller or, at the most, the same as SS-Flack. More precisely, SS-A&C-full, always between or equal to SS-A&C-min and SS-A&C-max, could be a more sensible choice. Of course in the limited number of cases in which SS-Donner is lower, again, it could be considered as an alternative for an "agreement or not" study.

More relevant is the fact that Donner et al.'s [92,94] proposal can be applied to agreement studies with more than two raters, taking also into account that the sample size decreases as the number of raters increases.

This is a very important point, since the commercial software such as PASS® 13/16 and nQuery®[100] make sample size calculation for only two raters. So it has to resort to the kappaSize package in R [94] for doing the pertinent sample size calculations or to our program, written in the open source R language, which uses only one function instead of the different formulas, each for a defined number of raters, of the kappaSize package (Appendix D). However, a relevant drawback of the "goodness of fitting model" proposed by Donner et al. [92,94] is the fact that it is based on the "agreement or not" without allowing for a differentiate level of disagreement through weighted kappa statistics. So, apart from the case of a study involving more than two raters it can be recommended only in a few cases in which its sample size is lower and the object of the study is the "agreement or not" situation. Of course, the sample sizes can be calculated with the minimum value of the kappa variance for the null (k_0) and alternative (k_A) hypothesis, leading to the lowest sample size. Perhaps, in this case there are no problems on the actual feasibility of the study but, nonetheless, there are some concerns on the actual possibility of demonstrating a difference clinically relevant; consequently, this approach cannot be recommended.

So, it has to be concluded that the possibility of having some sample sizes calculation procedures can be considered as a very useful tool in order to choose the sample size value more suitable for the actual feasibility of the study in the particular scenario of rows (columns) marginal and null and alternative hypothesis.

As a further remark, it has to be considered that the rows (columns) marginal pattern, the number of the categories of the variable, and also the number of raters influence the sample sizes. Since the required sample size increases as the marginal get more even, being the number of the categories fixed by the kind of the considered variable and the number of raters fixed by the kind of the agreement study, researchers are strongly advised to be very careful on the categories prevalence. Perhaps, if the knowledge of the distribution of the prevalence is unknown, a sensitivity sample size calculation starting from the uniform marginal pattern with the minimum sample size to an intermediate pattern has to be warmly recommended to choose the most reasonable sample size.

As a final consideration, it has to be stressed that the very similar maximum and minimum values of the kappa variance under the "partial common correlation model" allows avoiding the big differences between the sample sizes calculated with the maximum or minimum kappa variance according to Flack et al. [74]. Furthermore, the unique kappa variance value under the FCCM leads to sample sizes always within those calculated from the maximum and minimum values of the kappa variance under the PCCM and allows to argue that this latter model has to be strongly recommended since it avoids the need of calculating a maximum kappa variance value by resorting to LP procedure and, in addition, it refers to its population contingency table. So, it seems possible to conclude that with the FCCM we have obtained the optimum sample size without the rather forced a priori decision of using the maximum kappa variance according to Flack et al.'s approach [74].

Operative conclusions

In the case of 2 raters with uniform marginal, PCCM, FCCM and Flack et al.'s method [74] give the same sample sizes value that are lower than SS-Donner; so, any one of the three methods can be chosen.

In the case of 2 raters with non-uniform marginal, given as absolutely sensible the common correlation condition for the cells on the principal diagonal, the PCCM can be preferred to the Flack et al.'s [74] method since the sample sizes are smaller than or at maximum equal. Furthermore, if the common correlation condition for all cells of the contingency table can be considered sensible, the FCCM can be preferred to PCCM and Flack et al.'s method [74] since the sample sizes are smaller than or at maximum the same and the kappa variance is easier calculated. Donner et al. [92,94] method can be considered in the few cases in which the sample sizes are lower, provided that the "agreement or not" is the primary objective of the study and in the case of more than two raters, being the only one currently available. Of course it is also possible to calculate the maximum and minimum sample size according to PCCM and Flack et al.'s method [74] in order to have the complete sample sizes scenario and to choose the value which guarantees as much as possible the effective feasibility of the agreement study.

REFERENCES

1. Beal SL. (1989). Sample Size Determination for Confidence Intervals on the Population Mean and the Difference Between Two Population Means. Biometrics. 45: 969-977.
2. Cesana BM, Reina G, Marubini E. (2001). Sample Size for Testing a Proportion in Clinical Trials: A "Two-Step" Procedure Combining Power and Confidence Interval Expected Width. The American Statistician. 55: 288-292.
3. Cesana BM. (2004). Sample size for testing and estimating the difference between two paired and unpaired proportions: a "two-step" procedure combining power and the probability of obtaining a precise estimate. Statistics in Medicine. 23: 2359-2373.
4. Jiroutek MR, Muller KE, Kupper LL, Stewart PW. (2003). A New Method for Choosing Sample Size for Confidence Interval-Based Inferences. Biometrics. 50: 580-590.
5. Cesana BM, Antonelli P. (2018). Sample size calculations need to be adequate and parsimonious. *Journal of* Clinical Epidemiology. 108: 140-141.
6. Cesana BM, Antonelli P. (2010). A new approach to sample size calculations for the power of testing and estimating population means of Gaussian distributed variables. Biomed Stat Clin Epidemiol. 4: 67-78.
7. Cesana BM, Cavaliere F. (2016). Basics to perform and present statistical analyses in scientific biomedical reports Part 3. Minerva Anestesiologica. 82:1032-1035.
8. Marín-Martínez F, Sánchez-Meca J. (2009) Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. Educational and Psychological Measurement.
9. G*Power version 3.1.9.4 supported by the Department of Psychology of the University of Düsseldorf, Germany.
10. Wilkinson L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*. 54: 594-604.
11. Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, 1969, and Second Edition. USA: Lawrence Erlbaum Associates.
12. Lenth RV. (2012) Some Practical Guidelines for Effective Sample Size Determination. The American Statistician. 55: 187-193.
13. Lipsey MW, Puzio K, Yun C, Hebert M.A, Steinka Fry k. et al. (2012). Translating the Statistical Representation of the Effects of Education Interventions In to More Readily Interpretable Forms PDF. United States: U.S. Dept of Education, National Center for Special Education Research, Institute of Education Sciences, NCSER 2013-3000.
14. Sawilowsky S. (2009). New effect size rules of thumb. Journal of Modern Applied Statistical Methods. 2: 467-474.
15. Ellis PD. (2010). The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results.
16. https://en.wikipedia.org/wiki/Effect_size.

17. Jiroutek MR, Turner JR. (2018). Editorial Why it is nonsensical to use retrospective power analyses to conduct a postmortem on your study. 20: 408-410.
18. Choudhary PK, Nagaraja HN. (2017). Measuring Agreement: Models, Methods, and Applications.
19. Barlow W. (1996). Measurement of interrater agreement with adjustment for covariates. 52: 695-702.
20. Tanner MA, Young MA. (1985). Modeling agreement among raters. Journal of the American Statistical Association. 80: 175-180.
21. Nelson KP, Edwards D. (2008). On population based measures of agreement for binary classifications. Canadian Journal of Statistics. 36: 411-426.
22. Agresti A. (1992). Modeling patterns of agreement and disagreement. Statistical Methods in Medical Research. 1: 201-218.
23. Agresti A. (2013). *Categorical Data Analysis* Third Edition 2013 by John Wiley & Sons. Inc., Hoboken, New Jersey, USA.
24. Von Eye A, Mun EY. (2005). Analyzing Rater Agreement: Manifest Variable Methods. Lawrence Erlbaum Associates.
25. Lienert GA, Krauth J. (1975). Configural Frequency Analysis as a statistical tool for defining types. Educational and Psychological Measurement. 35: 231-238.
26. von Eye A. (2001). Configural Frequency Analysis version 2000 program for 32 bit operating systems. Methods of Psychological Research-Online. 6: 129-139.
27. von Eye A. (2002). Configural Frequency Analysis- methods, models, and applications. Mahwah, NJ: Lawrence Erlbaum Associates.
28. Guttman L. (1946). An approach for quantifying paired comparisons and rank order. Annals of Mathematical Statistics. 17: 144-163.
29. Goodman GD, Kruskal WH. (1954). Measures of association for cross classifications. Journal of the American Statistical Association. 49: 732-764.
30. Bennett EM, Alpert R, Goldstein AC. (1954). Communications through limited response questioning. Public Opinion Quarterly. 18: 303-308.
31. Scott WA. (1995). Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly. 19: 321-325.
32. Cohen JA. (1960). A coefficient of agreement for nominal scales. Educ Psychol Meas. 20: 37-46.
33. Brennan RL, Prediger DJ. (1981). Coefficient kappa: Some uses, misuses, and alternatives. Educ Psychol Meas. 41: 687-699.
34. Zwick R. (103). Another look at interrater agreement. Psychological Bulletin. 103: 374-378.
35. Krippendorff K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. Human Communication Research. 30: 411-433.
36. DeMast J. (2007). Agreement and kappa-type indices. The American Statistician. 61: 148-153.
37. Warrens MJ. (2010). Inequalities between kappa and kappa-like statistics for k×k tables. Psychometrika. 75: 176-185.
38. Warrens MJ. (2013). A Comparison of Cohen's Kappa and Agreement Coefficients by Corrado Gini. IRRAS 16. 3: 345-351.
39. Gini C. Indice di omofilia e di rassomiglianza e loro relazioni col coefficiente di correlazione e con gli indici di attrazione. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 1914-1915; Serie 8 Tomo LXXIV, Parte Seconda, 74: 583-610.
40. Gini C. Nuovi contributi alla teoria delle relazioni statistiche. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 1914-1915, Serie 8, Tomo LXXIV, Parte Seconda: 1903-1942.
41. Janson S, Vegelius J. (1979). On generalizations of the G index and the Phi coefficient to nominal scales. Multivariate Behavioral Research. 14: 255-269.
42. Popping R. Overeenstemmingsmaten voor Nominale Data. Rijksuniversiteit Groningen. Groningen, 1983.
43. Ato M, López JJ, Benavente A. (2011). A simulation study of rater agreement measures with 2x2 contingency tables. Psicológica. 32: 385-402.
44. Gwet K. (2008). Computing inter-rater reliability and its variance in presence of high agreement. British Journal of Mathematical & Statistical Psychology. 61: 29-48.
45. Aickin M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen’s kappa. Biometrics. 46: 293-302.
46. Martín A, Femia P. (2004). Delta: a new measure of agreement between two raters. British Journal of Mathematical and Statistical Psychology. 57: 1-19.
47. Feinstein AR, Cicchetti DV. (1990). High agreement but low kappa: I. Resolving the paradoxes. J Clin Epidemiol. 43: 543-549.
48. Landis JR, Koch GG. (1977). The measurement of interrater agreement for categorical data. Biometrics. 33: 159-174.

49. Svanholm H, Starklint H, Gunderson HJG, Fabricus J, Barlebo H, et al. (1989). Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS. 97: 689-698.

50. Fleiss JL, Cohen J, Everitt BS. (1969). Large sample standard errors of kappa and weighted kappa Psychological Bulletin. 72: 323-327.

51. Bakeman R, Quera V, McArthur D, Robinson BF. (1997). Detecting Sequential Patterns and Determining Their Reliability With Fallible Observers. Psychological Methods. 2: 357-370.

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

53. Warrens MJ.: http://www.matthijswarrens.nl/.

54. Thompson WD, Walter SD. (1988). A reappraisal of the kappa coefficient. J Clin Epidemiol. 1988;41:949-958.

55. Shoukri MM. (2004). Measures of Interobserver Agreement. 2004 Boca Raton, Fla: Chapman & Hall/CRC.

56. Byrt T, Bishop J, Carlin JB. (1993). Bias, Prevalence and Kappa. J Clin Epidemiol. 46: 423-429.

57. Warrens MJ. (2010). A Formal Proof of a Paradox Associated with Cohen's Kappa. Journal of Classification. 27: 322-332.

58. Gardner W. (1995). On the reliability of sequential data: Measurement, meaning, and correction. In JM. Gottman Ed., The analysis of change. pp. 339-359.

59. Bakeman R. (2018). KappaAcc: Deciding Whether Kappa is Big Enough by Estimating Observer Accuracy. Technical Report 28 January 20. Georgia State University Atlanta, GA 30303.

60. Brennan P, Silman A. (1992). Statistical methods for assessing observer variability in clinical measures. BMJ. 304: 1491-1494.

61. Gjørup T. (1988). The Kappa Coefficient and the Prevalence of a Diagnosis. Methods Inf. Med. 2704: 184-186.

62. Cicchetti DV, Feinstein AR. (1990). High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 43: 551-558.

63. Sim J, Wright CC. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 85: 257-268.

64. Cohen J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 70: 213-220.

65. Warrens MJ. (2013). Cohen's weighted kappa with additive weights Adv Data Anal Classif. 7: 41-55.

66. Fleiss JL, Nee JCM, Landis JR. (1979). Large sample variance of kappa in the case of different sets of raters. Psychol Bull. 86: 974-977.

67. Kraemer HC, Periyakoil VS, Nod A. (2004). Chapter 1.3 Agreement Statistics; Tutorial In Biostatistics; Kappa coefficients in medical research. Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by RB. D'Agostino 2004 John Wiley & Sons, Ltd.]

68. Everitt BS. (1968). Moments of the Statistics Kappa and Weighted Kappa. The British Journal of Mathematical and Statistical Psychology. 21: 97-103.

69. Stevens WL. (2011). The distribution of entries in a contingency table with fixed marginal totals. Annals of Human Genetics. 83: 238-244.

70. Yates F. (1934). Contingency Tables Involving Small Numbers and the χ2 Test. Supplement to the Journal of the Royal Statistical Society. 1: 217-235.

71. Mielke PW Jr, Berry KJ, Johnston JE. (2005). A Fortran Program for Computing the Exact Variance of Weighted Kappa. Perceptual and Motor Skills. 101: 468-472.

72. SAS/IML® 12.3 User's Guide. Copyright © 2013, SAS Institute Inc., Cary, NC, USA

73. R Core Team 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

74. Flack VFA, Afifi A, Lachenbruch PA, Schouten HJ. (1988). A Sample Size Determinations for the two Rater Kappa Statistic. Psychometrika. 53: 321-325.

75. PASS 13 Power Analysis and Sample Size Software 2014. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass. PASS 16 Power Analysis and Sample Size Software 2018. NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.

76. Pratt R, Hughes E. (2011). Linear Optimization in SAS/OR ® Software: Migrating to the OPTMODEL Procedure Paper 200-2011 SAS Global Forum.

77. Package 'lpSolve', (2020). Version 5.6.15 Title Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs.

78. Packages "irr". Authors: Gamer M, Lemon J, Fellows I, Singh P. and maintainer: Gamer M., version: 0.84.1.

79. Sallan JM, Lordan O, Fernandez V. (2015). Modeling and solving linear programming with R.

80. Donner A, Eliasziw M. (1992). A Goodness-Of-Fit Approach to Inference Procedures for the Kappa Statistic: Confidence and Sample Size Estimation Interval Construction, Significance-Testing. Statistics In Medicine. 11: 1511-1519.

81. Bloch DA, Kraemer HC. (1989). 2x2 kappa coefficients: measures of agreement or association. Biometrics. 45: 269-287.

82. Donner A, Eliasziw M. (1994). Statistical Implications of the Choice Between a Dichotomous or Continuous Trait in Studies of Interobserver Agreement. Biometrics. 50: 550-555.

83. Donner A, Eliasziw M, Klar N. (1996). Testing the Homogeneity of Kappa Statistics. Biometrics. 52: 176-183.

84. Donner A. (1998). Sample Size Requirements for the Comparison of two or more Coefficients of Inter-Observer Agreement. Statist. Med. 17: 1157-1168.

85. Donner A. (1999). Sample size requirements for interval estimation of the intraclass kappa statistic. Communications in Statistics - Simulation and Computation. 28: 415-429.

86. Bahadur R. (1961). A representation of the joint distribution of responses to n dichotomous items, in H. Solomonon, ed., Studies in Item Analysis and Prediction, Stanford, California: Stanford Mathematical Studies in the Social sciences.

87. Altaye M, Donner A, Klar N. (2001). Inference procedures for assessing interobserver agreement among multiple raters. Biometrics. 57: 584-588.

88. George EO, Bowman D. (1995). A full likelihood procedure for analyzing exchangeable binary data, Biometrics. 51: 512-523.

89. Altaye M, Donner A, Eliasziw M. (2001). A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Statist. Med. 20: 2479-2488.

90. Brier SS. (1980). Analysis of contingency tables under cluster sampling. Biometrika. 67:591-596.

91. Bartfay E, Donner A, Klar N. (1999). Testing the equality of twin correlations with multinomial outcomes. Annals of Human Genetics. 63: 341-349.

92. Donner A, Rotondi MA. (2010). Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. The International Journal of Biostatistics. 6: 31.

93. Fleiss J. (1981). Statistical Methods for Rates and Proportions. Wiley, New York.

94. Package ‘kappaSize’, Title: Sample Size Estimation Functions for Studies of Interobserver Agreement Version 1.1 released on February 20, 2015 and Version 1.2 released on November 26, 2018, Author: Michael A. Rotondi, Maintainer: Michael A. Rotondi.

95. Hong H, Choi Y, Hahn S, Park SK, Park BJ. (2014). Nomogram for sample size calculation on a straightforward basis for the kappa statistic Annals of Epidemiology. 2014; 24: 673-80.

96. Indurkhya A, Zavas LH, Buka SL. (2004). Sample size estimates for inter-rater agreement studies. Methods of Psychological Research-On line, 2004 in press.

97. Cantor AB. (1996). Sample-Size Calculations for Cohen's Kappa. Psychological Methods. 1: 150-153.

98. Mosimann J. (1962). On the compound multinomial distribution, the multivariate distribution, and correlations among proportions. Biometrika. 49: 65-82.

99. Rotondi MA, Donner A. (2012). A Confidence Interval Approach to Sample Size Estimation for Interobserver Agreement Studies with Multiple Raters and Outcomes. Journal of Clinical Epidemiology. 65: 778-784.

100. Elashoff JD. (2007). nQuery Advisor ® 2007 Version 7.0 User’s Guide. Statistical Solutions Ltd., Republic of Ireland.

APPENDIX

Appendix A: Linear Programming.

Flack et al. [74] proposed to determine the cell probabilities π_{ij} in order that the estimate of the kappa standard error is maximum.

Particularly, it has been suggested of “Placing all of the off-diagonal probability, 1 – p, on the p_{ij} corresponding to the largest rows

or columns marginal”. However, this suggestion was not supported by a detailed mathematical procedure leading to obtain the required and unique contingency table. It has to be noted that the problem of obtaining a contingency table with a determined Cohen’s kappa variance can be solved by recurring to Linear Programming (PL). This calculation can be carried out by SAS® IML [72] as detailed in its User’s Guide https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=imlug&docsetTarget=imlug_genstatexpls_sect011.htm&locale=en).

Generally, one must resort to matrix algebra and it has to specify a vector for the lower bounds and/or upper bounds of the variables (X) and the vector of coefficients (**C**) such that **C`X** is the linear objective function (o.f.). It has to be noted that the above expression in matrix algebra means that the transpose (` or ^{T}) vector of the coefficients pre-multiplies the vector of the variables. By remembering that the linear programming allows obtaining a required value, usually a maximum (max) or a minimum (min), the PL general matrix expression is given by:

$$\max(\min)\, z = \boldsymbol{C}^T \boldsymbol{X}$$

$$\boldsymbol{AX} \lesseqqgtr \boldsymbol{B} \qquad \boldsymbol{X} \geq 0$$

Where z is the objective function (o.f.) to be returned to a maximum (minimum), **X** is the vector of the involved variables, and $\mathrm{AX} \lesseqqgtr \mathrm{B}$ is the set of the non-trivial constraints of **X**.

It has to be stressed that both the o.f. and the constraints must be linear equations.

According to Wikipedia, “In mathematical optimization, a popular algorithm for linear programming is the simplex algorithm developed by Dantzig (George Bernard Dantzig, November 8, 1914 – May 13, 2005, an American mathematical scientist, https://en.wikipedia.org/wiki/Simplex_algorithm) in 1947. The name of the algorithm is derived from the concept of a simplex (a simplex, plural: simplexes or simplices is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions) and was suggested by Motzkin (Theodore Samuel Motzkin, 26 March 1908 – 15 December 1970 was an Israeli-Americanmathematician, https://en.wikipedia.org/wiki/Theodore_Motzkin)”.

The formulation of our problem in terms of the PL has its first step in identifying the objective function to be maximized consisting in the formula of the kappa (kw) variance (reproduced here for easiness of reading, with “c” for the number of categories equal to the number of the rows or of the columns):

$$Var(k_w) = \frac{1}{N(1-p_e)^4}\left\{\sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij}\left[w_{ij}(1-p_e) - (\bar{w}_{i.} + \bar{w}_{.j})(1-p_0)\right]^2 - (p_0 p_e - 2p_e + p_0)^2\right\}$$

where$p_{.j} = \sum_{i=1}^{c} p_{ij}$; $\bar{w}_{i.} = \sum_{j=1}^{c} w_{ij} \cdot p_{.j}$ $\quad and \quad p_{i.} = \sum_{j=1}^{c} p_{ij}$; $\bar{w}_{.j} = \sum_{i=1}^{c} w_{ij} \cdot p_{i.}$

The above formula can be written also as:

$$Var(k_w) = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij} \cdot \frac{\left[w_{ij}(1-p_e) - (\bar{w}_{i.} + \bar{w}_{.j})(1-p_0)\right]^2 - (p_0 p_e - 2p_e + p_0)^2}{N(1-p_e)^4} = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij} \cdot c_{ij}$$

Where:

$$c_{ij} = \frac{\left[w_{ij}(1-p_e) - \left(\bar{w}_{i.} + \bar{w}_{.j}\right)\left(1-p_0\right)\right]^2 - (p_0 p_e - 2p_e + p_0)^2}{N(1-p_e)^4}$$

Of course, in the case of the unweighted kappa the weight matrix for "agreement weights" is a unity matrix and the above formula becomes:

$$c_{ij} = \frac{\left[(1-p_e) - \left(p_{i.} + p_{.j}\right)\left(1-p_0\right)\right]^2 - (p_0 p_e - 2p_e + p_0)^2}{N(1-p_e)^4}$$

So, the kappa variance is a linear function of the probabilities of the table.

In order to obtain the previously reported formula$z = \boldsymbol{C}^T\boldsymbol{X}$, where **C** and **X** are vectors, the matrix of the probabilities (p_{ij}) and the matrix of the coefficients (c_{ij}) have to be transformed in their corresponding vectors, by stacking the columns of each matrix one below the other; for example, the probability matrix (**P**) becomes the row vector **X** and the coefficients matrix (**F**, say) becomes the vector **C**:

$$\begin{bmatrix} p_{11} & \cdots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix} \rightarrow [p_{11} \cdots \quad p_{k1} \cdots \quad p_{1k} \cdots \quad p_{kk}]^T = \boldsymbol{X}^T$$

$$\begin{bmatrix} c_{11} & \cdots & c_{1k} \\ \vdots & \ddots & \vdots \\ c_{k1} & \cdots & c_{kk} \end{bmatrix} \rightarrow [c_{11} \cdots \quad c_{k1} \cdots \quad c_{1k} \cdots \quad c_{kk}]^T = \boldsymbol{C}^T$$

The operator "T" (superscript) for "transpose" has been written since a row vector it is shown to save space instead of a column vector. The second step consists in the formulation of the constraints, taking into account that the row and column constraints together with the constraint pertinent to the fixed kappa value are equalities and, consequently, linear expressions. Also these constraints have to be transformed, similarly to the transformation made on the variables and coefficients matrices in order to be expressed in the matrix formula:

$\boldsymbol{AX = B}$

Let’s take the simple case of a 3x3 contingency table together with its 3x3 probabilities matrix (**P**) and coefficients matrix (**F**).

So:

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \boldsymbol{F} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

The row and column constraints are, respectively:

$$p_{11} + p_{12} + p_{13} = p_{1.} p_{11} + p_{21} + p_{31} = p_{.1}$$

$$p_{21} + p_{22} + p_{23} = p_{2.} p_{12} + p_{22} + p_{32} = p_{.2}$$

$$p_{31} + p_{32} + p_{33} = p_{3.} p_{13} + p_{23} + p_{33} = p_{.3}$$

It has to be noted that one of these constraints is redundant (for example the third), but since this fact does not affect this procedure, it has been retained for the sake of clarity.

Then, the further constraint due to kappa is:

$$p_{11} + p_{22} + p_{33} = p_0$$

So, the column vector **C** and **X** are:

$$\boldsymbol{C}^T = [c_{11}, c_{21}, c_{31}, \cdots, c_{13}, c_{23}, c_{33}] \boldsymbol{X}^T = [p_{11}, p_{21}, p_{31}, \cdots, p_{13}, p_{23}, p_{33}]$$

The first constraint (the sum of the probability values of the first row equals the respective marginal) given by: $p_{11} + p_{12} + p_{13} = p_{1.}$ becomes:

$$[1\ 0\ 0, 1\ 0\ 0, 1\ 0\ 0] \cdot \left[p_{11} p_{21} p_{31}, \cdots, p_{13}\ p_{23} p_{33}\right]^T = p_{1.}$$

And so on for the remaining k row constraints and for the k (or k-1) column constraints.

The constraint due to kappa (the sum of the probability values on the principal diagonal must be equal to the observed proportion of agreement to give a well-defined kappa value) given by: $p_{11} + p_{22} + p_{33} = p_0$ becomes:

$$[1\ 0\ 0, 0\ 1\ 0, 0\ 0\ 1] \cdot \left[p_{11} p_{21} p_{31}, \cdots, p_{13}\ p_{23} p_{33}\right]^T = p_0$$

The matrix **A,** obtained by including all constraints, is:

$$\boldsymbol{A}=\begin{bmatrix}1&0&0&1&0&0&1&0&0\\0&1&0&0&1&0&0&1&0\\0&0&1&0&0&1&0&0&1\\1&1&1&0&0&0&0&0&0\\0&0&0&1&1&1&0&0&0\\0&0&0&0&0&0&1&1&1\\1&0&0&0&1&0&0&0&1\end{bmatrix}$$

It has to be noted that this matrix has 7 constraints (3 for the rows, 3 for the columns and 1 for the kappa or the observed proportion of agreement) and 9 columns since there are 9 probability values and 9 coefficients.

The vector of the known value is given by the known row and column marginal and by the known proportion of the observed agreement, given the kappa value:

$$\boldsymbol{B}=[p_{1.}\,p_{2.}\,p_{3.}\,p_{.1}\,p_{.2}\,p_{.3}\,p_{0}]$$

Then the expression of the constraints in matrix algebra is:

$\boldsymbol{AX}=\boldsymbol{B}$

So, with the previously outlined transformations, we have obtained a linear programming problem to be resolved with the pertinent method.

Summarizing, from a square (cxc) contingency table, we obtain:

1)- **X**: a vector with c^2 elements of the unknown terms (from the **P** matrix of the unknown cell probabilities)

2)- **C**: a vector with c^2 elements of the coefficients of the o.f. (from the **F** matrix)

3)- **A**: a matrix with $(2c+1)$ x (c^2) elements: $(2c+1)$ rows and c^2 columns

4)- **B**: a vector with $(2c+1)$ elements of the known terms.

Procedure in R

In the R settings, the procedure lp(…) of the package lsSolve allows to obtain the solution of our maximization problem (https://CRAN.R-project.org/package=lpSolve).
However, in order to utilize this procedure, a further vector (Dir) has to be made with its elements given by the kind of the inequality together with its direction (< or >) or by the equality (=) of each constraint.

It has to be noted that in the case we are interested in, this vector has to be made of all equalities:

$Dir=["=","=","=","=","=","=","="]$

Essential parameters of the function lp(…) are:

1)-“direction”: “Character string giving direction of optimization”: "min" (default) or "max."

2)-“objective.in”: “Numeric vector of the objective function coefficients”

3)-“const.mat”: “Matrix of numeric constraint coefficients: particularly, one row per constraint, one column per variable”.

4)-“const.dir”: “Vector of character strings giving the direction of the constraint: each value should be one of” "<," "<=," "=," "==," ">," or ">=". “(In each pair the two values are identical.)”

5)-“const.rhs”: “Vector of numeric values for the right-hand sides of the constraints”.

Particularly, in the case of calculating the probabilities table for obtaining the maximum kappa variance, it has to write:

lp(direction = "max", objective. in = C^T, const. mat = A, const. dir = Dir, const. rhs = B)

Otherwise, for the minimum variance, it has to write: direction = "min"

A numerical example:

```
 # the package lsSolve has to be uploaded
#Having fixed the row equal to the column marginal (Pmarg = c(,,,), the kappa value and the number of observation (n_obs):
Pmarg = c(.5,.3,.2); kappa =0.1,n_obs = 200
```

And the weights matrix W

```
> W (weights matrix)
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 1 0
[3,] 0 0 1
```

It is possible to obtain: p_o (the proportion of observed agreement) and p_e (the proportion of the chance agreement)

```
> po; pe
[1] 0.442
[1] 0.380
```

and the matrix **A**: (matrix of the coefficients of the matrix equation **AX = B)**

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1 0 0 1 0 0 1 0 0
[2,] 0 1 0 0 1 0 0 1 0
[3,] 0 0 1 0 0 1 0 0 1
[4,] 1 1 1 0 0 0 0 0 0
[5,] 0 0 0 1 1 1 0 0 0
[6,] 0 0 0 0 0 0 1 1 1
[7,] 1 0 0 0 1 0 0 0 1
```

The vector B of the known terms of the system **AX=B** is:

```
B=[ 0.500 0.300 0.200 0.500 0.300 0.442]
#Intermediate calculations for obtaining the coefficients (cij) of the objective formula (o.f.)
#Calculus of WM (matrix of the quantities w̄i.+w̄.j for the two pedix i and j
> WM_c<-t(W)%*%pmarg
> WM_r<-W%*%pmarg
> WM<-outer(c(WM_r),c(WM_c),"+")
> WM
[,1] [,2] [,3]
[1,] 1.0 0.8 0.7
[2,] 0.8 0.6 0.5
[3,] 0.7 0.5 0.4
> W*(1-pe)
[,1] [,2] [,3]
[1,] 0.62 0.00 0.00
[2,] 0.00 0.62 0.00
[3,] 0.00 0.00 0.62
> WM*(1-po)
[,1] [,2] [,3]
[1,] 0.5580 0.4464 0.3906
[2,] 0.4464 0.3348 0.2790
[3,] 0.3906 0.2790 0.2232
> (W*(1-pe)-WM*(1-po))^2 (first part of the numerator of the formula of the coefficients cij)
[,1] [,2] [,3]
[1,] 0.0038440 0.19927296 0.1525684
[2,] 0.1992730 0.08133904 0.0778410
[3,] 0.1525684 0.07784100 0.1574502
> (po*pe-2*pe+po)
[1] -0.15004
> ((po*pe-2*pe+po))^2 (second part of the numerator of the formula for the coefficients cij)
[1] 0.022512
> (n_obs*(1-pe)^4) (denominator of the formula of the coefficients cij)
[1] 29.55267
> C2_mtx (matrix of the coefficients cij)
[,1] [,2] [,3]
```

```
[1,] -0.0006316857 0.005981217 0.004400832
[2,] 0.0059812175 0.001990583 0.001872216
[3,] 0.0044008325 0.001872216 0.004566025
> C2 is the vector C of the coefficients of the o.f. calculated with the previously reported formula (matrix of the coefficients cij transformed in a vector)
[1] -0.0006316857 0.0059812175 0.0044008325 0.0059812175 0.0019905827
[6] 0.0018722164 0.0044008325 0.0018722164 0.0045660250
```

The vector C, obtained from the previous formula, is:

C_T= [-0.0006316857 0.0059812175 0.0044008325 0.0059812175 0.0019905827 0.0018722164 0.0044008325 0.0018722164 0.0045660250]

```
### calculus of the probability matrix
> library(lpSolve)
```

The vector Dir is:

```
> Dir <- ["=" "=" "=" "=" "=" "=" "="]
```

The command to be performed is:

```
> sol_max<- lp(direction = "max", objective.in = C, const.mat = A, const.dir = Dir, const.rhs = B)
# “the procedure lp returns an object containing, among others, the objective function value at the optimum (maximum or minimum requested) and the vector of the optimal coefficients called solution”.
> sol_max$f.obj.value
[1] 0.004152924
> sol_max$solution
[[1] 0.221 0.279 0.000 0.279 0.021 0.000 0.000 0.000 0.200
> m_max<- matrix(sol_max$solution,n,n) # this command transforms the probability vector in a table
> m_max
[,1] [,2] [,3]
[1,] 0.221 0.279 0.0
[2,] 0.279 0.021 0.0
[3,] 0.000 0.000 0.2
kappa_Fleiss(x = m_max,n_obs = 200)
size po pe kappa var_K
200 0.442 0.38 0.1 0.004152924
```

Where “size” is the number of observations, “po” is the observed agreement proportion for obtaining, given the row and columns marginal, the prefixed kappa value, “pe” is the proportion of the agreement given by chance, kappa is the kappa value and “var_k” is the kappa variance value that is a maximum in this case.

For the minimum variance value, the command becomes:

```
> sol_min<- lp(direction = "min", objective.in = C, const.mat = A, const.dir = Dir, const.rhs = B)
m_min
[,1] [,2] [,3]
[1,] 0.421 0.079 0.0
[2,] 0.079 0.021 0.2
[3,] 0.000 0.200 0.0
$kappa_Fleiss
size    po      pe     kappa      var_K
200     0.442   0.38   0.1        0.001469781
```

Where "size" is the number of observations, "p_o" is the observed agreement proportion for obtaining, given the rows and columns marginal, the prefixed kappa value, "p_e" is the proportion of the agreement given by chance, kappa is the kappa value and "var_k" is the kappa variance value that is a minimum, in this case.

In addition, it is possible also to obtain a probability table with a fixed value of the "unity kappa variance".

Example of sample size calculation according to Flack et al. [74]

```
> kappa_ssize_FL(pmarg = c(.5,.3,.2), kappa0 =.6, kappa =.8, max_min_var = "max", alpha = .05, power = .8, two_one_tail = 2, out = 2)
         size       alpha power  kappa0  kappaA p1 p2 p3   tau_H0      tau_HA      varK_H0     varK_HA
[1,]     88.34428 0.05   0.8      0.6       0.8       0.5 0.3 0.2 0.7263955 0.5419585 0.5276504  0.293719
```

#tau_H0 and tau_HA are the standard errors of the corresponding unity variances (formula 2, page 322 of Flack et al.[74]), under H_0 and H_A hypothesis, respectively.

#Check of the values of the variance under H_0 and under H_A:

```
> Mtx_nxn_PL(pmarg = c(.5,.3,.2), kappa = .6, max_min_var = "max", n_obs = 1, out = 2)
$mtx_vmax
[,1] [,2] [,3]
[1,] 0.376 0.124 0.0
[2,] 0.124 0.176 0.0
[3,] 0.000 0.000 0.2
$kappa_Fleiss
        size     po      pe      kappa     var_K         Var_K_0_exact        Var_K_0_appx
[1,]    1        0.752   0.38    0.6       0.5276504     0.005371089          0.5317378
```

#It has to be noted that "Var k_0_exact" is the value of the exact kappa variance under the null hypothesis (H_0: k = 0.6), according to Everitt [68].

```
> Mtx_nxn_PL(pmarg = c(.5,.3,.2), kappa = .8, max_min_var = "max", n_obs = 1, out = 2)
$mtx_vmax
[,1] [,2] [,3]
```

```
[1,] 0.438 0.062 0.0
[2,] 0.062 0.238 0.0
[3,] 0.000 0.000 0.2
$kappa_Fleiss
     size  po     pe   kappa   var_K      Var_K_0_exact  Var_K_0_appx
[1,]   1  0.876  0.38   0.8   0.293719     0.005371089     0.5317378
```

Appendix B: Dirichlet Multinomial Distribution. Sample Size According to the "Goodness-Of-Fit" approach from Donner and Eliasziw [80,82].

By using the same notation of Altaye et al. [87], let X_{ij}denote the absolute frequency of the ratings on subject i (i = 1, …, n) into category j (j = 1, …, c, being c mutually exclusive categories) rated independently by a set of "n" randomly selected raters. Furthermore, let's assume that the rating probability of each category are P_1, P_2, …, P_c, where $P_1+\cdots+P_c=1$ and, consistently, $X_{i1}+\cdots+X_{ic}=$ n. So, it has to be noted that the above model assumes the rater interchangeability or the marginal homogeneity across raters.

Then, under the usual assumption that the Xij's are mutually independent, it follows that the joint distribution of the X_{ij}'s conditional on Pj (j = 1, 2, …, c) is given by the multinomial distribution:

$$P\left(X_{i1}, X_{i2}, \ldots, X_{ic} \mid n, P_1, P_2, \ldots, P_c\right) = \prod_{j=1}^{c} \frac{P_j^{X_j}}{X_{ij}!} =$$

However when the assumption of mutual independence does not hold, the Pj may alternatively be assumed to follow a Dirichlet distribution with parameters $a_1, a_2, \ldots, a_c$ as it has been shown, among others, by Brier [90] and Mosimann [98], with density function given by:

$$f\left(P_1, P_2, \ldots, P_c \mid a_1, a_2, \ldots, a_c\right) = \frac{\Gamma\left(\sum_{j=1}^{c} a_j\right) \prod_{j=1}^{c} P_j^{a_j - 1}}{\prod_{j=1}^{c} \Gamma\left(a_j\right)}$$

Where $a_j > 0$.

Then the joint distribution of the Xij's is a Dirichlet multinomial and is given by:

$$P\left(X_{i1}, X_{i2}, \ldots, X_{ic}\right) = \frac{n!\,\Gamma\left(\sum_{j=1}^{c} a_j\right) \prod_{j=1}^{c} \Gamma\left(X_{ij} + a_j\right)}{\prod_{j=1}^{c} X_{ij}!\,\Gamma\left(n + \sum_{j=1}^{c} a_j\right) \prod_{j=1}^{c} \Gamma\left(a_j\right)}$$

Equivalently, if we let $\theta=\sum_{j=1}^{c} a_j$, and $\pi_j = a_j / \theta$, where π_j, is the expected value of P_j, we obtain:

$$P\left(X_{i1},X_{i2},...,X_{ic}\right)=\frac{n!\Gamma(\theta)\prod_{j=1}^{c}\Gamma\left(X_{ij}+\theta\pi_j\right)}{\prod_{j=1}^{c}X_{ij}!\Gamma(n+\theta)\prod_{j=1}^{c}\Gamma\left(\theta\pi_j\right)}$$

A different formulation of the Dirichlet Multinomial Distribution.

We consider the following different notation for the factorial algebra:

$$\frac{n!}{\prod_{j=1}^{c}X_{ij}!}=\binom{n}{X_{i1},X_{i2},...,X_{ic}}$$

And the following for the Gamma function.

$$\frac{\Gamma(\theta)}{\Gamma(n+\theta)}=\frac{\Gamma(\theta)}{(n+\theta-1)(n+\theta-2),\ldots,\theta\Gamma(\theta)}=\frac{1}{\theta_{(n)}}$$

$$\text{With: }\theta_{(n)}=\theta(\theta+1),\ldots,(\theta+n-1)=\prod_{h=0}^{n-1}(\theta+h)$$

$$\frac{\Gamma\left(X_{ij}+\theta\pi_j\right)}{\Gamma\left(\theta\pi_j\right)}=\frac{\left(X_{ij}+\theta\pi_j-1\right)\left(X_{ij}+\theta\pi_j-2\right)\cdot\ldots\cdot\theta\pi_j\Gamma\left(\theta\pi_j\right)}{\Gamma\left(\theta\pi_j\right)}=\theta\pi_j\left(\theta\pi_j+1\right)\cdot\ldots\cdot\left(\theta\pi_j+X_{ij}-1\right)=\left(\theta\pi_j\right)_{\left(x_{ij}\right)}$$

$$\text{with }\left(\theta\pi_j\right)_{\left(x_{ij}\right)}=\theta\pi_j\left(\theta\pi_j+1\right)\cdot\ldots\cdot\left(\theta\pi_j+X_{ij}-1\right)=\prod_{h=0}^{X_{ij}-1}(\theta\pi+h)$$

So:

$$P\left(X_{i1},X_{i2},..X_{ic}\right)=\frac{n!}{\prod_{j=1}^{c}X_{ij}!}\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\frac{\prod_{j=1}^{c}\Gamma\left(X_{ij}+\theta\pi_j\right)}{\prod_{j=1}^{c}\Gamma\left(\theta\pi_j\right)}=\frac{n!}{\prod_{j=1}^{c}X_{ij}!}\frac{\Gamma(\theta)}{\Gamma(n+\theta)}\prod_{j=1}^{c}\frac{\Gamma\left(X_{ij}+\theta\pi_j\right)}{\Gamma\left(\theta\pi_j\right)}=$$

$$=\binom{n}{X_{i1},X_{i2},...,X_{ic}}\frac{1}{\theta_{(n)}}\prod_{j=1}^{c}\left(\theta\pi_j\right)_{\left(X_{ij}\right)}=\binom{n}{X_{i1},X_{i2},...,X_{ic}}\frac{\prod_{j=1}^{c}\prod_{h=0}^{X_{ij}-1}\left(\theta\pi_j+h\right)}{\prod_{h=0}^{n-1}(\theta+h)}$$

Letting $\theta=\frac{1-k}{k}$ [leading to $k=(1+\theta)^{-1}$], the basic model becomes:

$$P(X_{i1},X_{i2},..X_{ic})=\binom{n}{X_{i1},X_{i2},\dots,X_{ic}}\frac{\prod_{j=1}^{c}\prod_{h=0}^{X_{ij}-1}\left(\frac{1-k}{k}\pi_j+h\right)}{\prod_{h=0}^{n-1}\left(\frac{1-k}{k}+h\right)}=\binom{n}{X_{i1},X_{i2},\dots,X_{ic}}\frac{\prod_{j=1}^{c}\prod_{h=0}^{X_{ij}-1}\left(\frac{(1-k)\pi_j+hk}{k}\right)}{\prod_{h=0}^{n-1}\left(\frac{1-k+hk}{k}\right)}=$$

$$=\binom{n}{X_{i1},X_{i2},\dots,X_{ic}}\frac{\prod_{j=1}^{c}\prod_{h=0}^{X_{ij}-1}((1-k)\pi_j+hk)}{\prod_{h=0}^{n-1}(1-k+hk)}$$

Calculus of the Probability of the Agreement against the "not agreement".

Donner and Eliasziw [80,82] proposed to combine all the cells with a disagreement component into a single cell. This will allow us to test hypotheses concerning the specified value of a single parameter. So, the cxc cells of the probability table are split into c cell with the probability of the perfect agreement and a further cell with the probability of all ratings with a disagreement component.

For the given set of "n" raters, let's $P(_{j1,j2,\dots,jn})$ be the probability that rater 1 chooses category "j1" , rater 2 chooses category "j2", and so on. Then, the probability that all raters chooses the same category j is:

$$P\left(\underbrace{j,j,\dots,j}_{n}\right)=P\left(\underbrace{X_{i1}=0,j\dots,X_{ij}=n,\dots,X_{ic}=0}_{n}\right)=\binom{n}{\underbrace{0,\dots,n,..,0}_{c}}\frac{\prod_{h=0}^{n-1}(\theta\pi_j+h)}{\prod_{h=0}^{n-1}(\theta+h)}=$$

$$=\prod_{h=0}^{n-1}\frac{(\theta\pi_j+h)}{(\theta+h)}=\frac{\theta\pi_j(\theta\pi_j+1)\dots(\theta\pi_j+(n-1))}{\theta(\theta+1)\dots(\theta+n-1)}=$$

$$=\frac{\pi_j(\theta\pi_j+1)\dots(\theta\pi_j+(n-1))}{(\theta+1)\dots(\theta+n-1)}=\frac{\pi_j\prod_{h=1}^{n-1}(\theta\pi_j+h)}{\prod_{h=1}^{n-1}(\theta+h)}$$

That, expressed in terms of k, becomes:

$$P\left(\underbrace{j,j,\ldots,j}_{n}\right)=\frac{\pi_j\prod_{h=1}^{n-1}\left(\theta\pi_j+h\right)}{\prod_{h=1}^{n-1}\left(\theta+h\right)}=\frac{\pi_j\prod_{h=1}^{n-1}\left(\left(1-k\right)\pi_j+hk\right)}{\prod_{h=1}^{n-1}\left(\left(1-k\right)+hk\right)}=\frac{\pi_j\left(\left(1-k\right)\pi_j+k\right),\cdots,\left(\left(1-k\right)\pi_j+\left(N-1\right)k\right)}{\left(\left(1-k\right)+k\right)\left(\left(1-k\right)+2k\right),\ldots,\left(\left(1-k\right)+\left(N-1\right)k\right)}=$$

$$=\frac{\pi_j\left(\left(1-k\right)\pi_j+k\right),\cdots,\left(\left(1-k\right)\pi_j+\left(N-1\right)k\right)}{\left(1+k\right)\left(1+2k\right),\ldots,\left(1+\left(N-2\right)k\right)}$$

For example, with 3 raters and 3 categories, the probability that the three raters rank the i-th subject in the j-th (j=1,2,3) category is:

$$P\left(\underbrace{j,j,j}_{3}\right)=P\left(\underbrace{X_{ij}=3,X_{is,s\neq j}=0}_{3}\right)=\frac{\pi_j\prod_{h=1}^{3-1}\left(\left(1-k\right)\pi_j+hk\right)}{\left(k+1\right),\ldots,\left(1+\left(3-2\right)k\right)}=\frac{\pi_j\left(\pi_j-k\pi_j+k\right)\left(\pi_j-k\pi_j+2k\right)}{\left(k+1\right)}$$

The above formula corresponds to the first three rows of the formula (4) of Altaye et al. [89].

There is also the following recursive version of this formula:

$$P\left(\underbrace{j,j,\ldots,j}_{n}\right)=\frac{\pi_j\prod_{h=0}^{n-1}\left(\theta\pi_j+h\right)}{\prod_{h=0}^{n-1}\left(\theta+h\right)}=\frac{\pi_j\prod_{h=0}^{n-2}\left(\theta\pi_j+h\right)}{\prod_{h=0}^{N-2}\left(\theta+h\right)}\frac{\left(\theta\pi_j+\left(n-1\right)\right)}{\left(\theta+\left(n-1\right)\right)}=P\left(\underbrace{j,j,\ldots,j}_{n}\right)\frac{\left(\theta\pi_j+n-1\right)}{\left(\theta+n-1\right)}=$$

$$P\left(\underbrace{j,j,\ldots,j}_{n}\right)\frac{\left(\left(1-k\right)\pi_j+\left(n-1\right)k\right)}{\left(\left(1-k\right)+\left(n-1\right)k\right)}=P\left(\underbrace{j,j,\ldots,j}_{n}\right)\frac{\left(\pi_j+\left(n-1-\pi_j\right)k\right)}{\left(1+\left(n-2\right)k\right)}$$

Putting $P^{(n)}\left(j\right)=P\left(\underbrace{j,j,\ldots,j}_{n}\right)$, it is possible to write:

$$P^{(n)}\left(j\right)=P^{(n-1)}\left(j\right)\frac{\left(\pi_j+\left(n-1-\pi_j\right)k\right)}{\left(1+\left(n-2\right)k\right)}\text{ and }P^{(1)}\left(j\right)=\pi_j$$

Sample Size Calculation.

The above formulas can be used for the sample size calculation under the “Goodness of fit” approach.

Several Authors have suggested to calculate these probabilities with the equivalent formula from Altaye et al. [87]:

$$P\left(\underbrace{j_{(1)},j_{(2)},\ldots,j_{(n)}}_{n}\right)=\pi_j\left[1+k\sum_{s=1}^{n-1}Z_s+\sum_{s=1}^{n-1}Z_sZ_1+,\ldots,k^{N-1}Z_1Z_2,\ldots,Z_{N-1}\right]\left[\prod_{r=1}^{N-2}\left(1+hk\right)\right]^{-1}$$

$$\text{where }Z_s=\frac{s-\pi_j}{\pi_j}\text{ for }s=1,2,\ldots,n-1$$

Rotondi and Donner [92,94,99] have implemented the package “kappaSize” of the open source R software with four functions (PowerBinary, Power3Cats, Power4Cats,Power5Cats) in order to calculate the sample sizes for a limited number of categories (respectively: 2, 3 4 and 5) and of raters (2 to 6).

However, by using the previously reported formula (in both the recursive and non-recursive versions), it is possible to write a unique function allowing to calculate the sample size for any number of categories and raters.

The software in the open source R language of the above mentioned formula is shown in the Appendix D.

Appendix C: Sample Sizes shown in the von Eye and Mun’s book [24]

von Eye and Mun’s book [24] showed the Table 1.4: “Minimum Required Sample Sizes for $\alpha = 0.05$ and $p = 0.8$ (power)” summarizing two tables of Indurkhya et al.’s paper [96]. Particularly, the first table reports the required sample sizes for the null hypothesis that $k_0 = 0.4$ vs. the alternative one that $k_1 = 0.6$, for $\alpha = 0.05$ and $p = 0.8$ (for power); then the second table reports the required sample sizes for the null hypothesis that $k_0 = 0.6$ vs. the alternative one that $k_1 = 0.8$, for $\alpha = 0.05$ and $p = 0.8$ (for power).

It has to be commented that the sample sizes shown for the case of “Null hypothesis: $k_0 = 0.4$, alternative hypothesis: $k_1 = 0.6$, three categories and number of raters ranging from 2 to 6” are obtained or exactly or with a difference of only one unity from the procedure of Altaye and Donner [87,89] and Donner et al. [92,94] with the package “kappaSize” of R [94], as it is shown in the following table. Between brackets there are the sample sizes calculated from the Rotondi’s R software [94] Power3 Cats “Power-Based Approach for the Number of Subjects Required for a Study of Interobserver Agreement with Three Outcome Categories”, Power 3 Cats(kappa0=, kappa1=, props=C(), raters=, alpha=, power=) where the significance level alpha (α) is two-sided.

Marginal probabilities			**Number of raters**				
π_1	π_2	π_3	2	3	4	5	6
0.10	0.10	0.80	205(206)	113(114)	83(84)	68(69)	59(60)
0.10	0.40	0.50	127(127)	69(69)	50(50)	40(41)	35(35)
0.33	0.33	0.34	107(106)	58(58)	42(43)	35(35)	30(30)

All sample sizes are equal or differ of one unit at most.

The same pattern occurs for the third part of Table 1.4, always with three categories and with null hypothesis: $k_0 = 0.6$ and alternative hypothesis: $k_1 = 0.8$.

Marginal probabilities			**Number of raters**				
π_1	π_2	π_3	2	3	4	5	6
0.10	0.10	0.80	172(173)	102(102)	77(78)	66(65)	58(58)
0.10	0.40	0.50	102(103)	60(61)	46(46)	40(39)	35(34)
0.33	0.33	0.34	87(87)	52(51)	39(39)	33(33)	30(29)

However, sample sizes are very different in the case of four categories.

The following table shows the results for the case of the “Null hypothesis: $k_0 = 0.4$, alternative hypothesis: $k_1 = 0.6$”, corresponding to the second part of Table 1.4 of von Eye and Mun’s book [24].

The sample sizes have been calculated by the function Power4Cats “Power-Based Approach for the Number of Subjects Required for a Study of Interobserver Agreement with Four Outcome Categories” [94].

Marginal probabilities				Number of raters				
π_1	π_2	π_3	π_4	2	3	4	5	6
0.10	0.10	0.10	0.70	102(140)	42(78)	38(58)	32(48)	29(41)
0.10	0.30	0.30	0.30	88(92)	30(51)	30(38)	29(31)	27(27)
0.25	0.25	0.25	0.25	60(87)	28(49)	27(36)	25(30)	25(26)

A similar pattern of a relevant difference occurs also for the case of “null hypothesis: $k_0 = 0.6$, alternative hypothesis: $k_1 = 0.8$”, corresponding to the fourth part of the Table 1.4 of von Eye’s and Mun’s book [24].

The sample sizes between brackets have been calculated by the function Power4Cats “Power-Based Approach for the Number of Subjects Required for a Study of Interobserver Agreement with Four Outcome Categories” [94].

Marginal probabilities				Number of raters				
π_1	π_2	π_3	π_4	2	3	4	5	6
0.10	0.10	0.10	0.70	157(119)	74(71)	68(54)	52(46)	49(40)
0.10	0.30	0.30	0.30	88(78)	30(46)	30(35)	29(30)	27(26)
0.25	0.25	0.25	0.25	60(74)	28(44)	27(34)	25(28)	25(25)

So, it is very strange that the sample sizes are the same or different for only one unit in the 3x3 categories case and very different in the 4x4 categories case with a general increase for the “null hypothesis: $k_0 = 0.4$, alternative hypothesis: $k_1 = 0.6$”, and an increase/decrease pattern for the “null hypothesis: $k_0 = 0.6$, alternative hypothesis: $k_1 = 0.8$”. Furthermore, the sample sizes shown in the Table 1.4 of the von Eye and Mun’s book [24] for the four categories case do not show the nice decreasing pattern at the increasing number of raters as it is shown by the sample sizes calculated by Rotondi’s “Power4Cats” function [94]. So, it seems to be sensible to not rely on the sample sizes shown in von Eye and Mun’s book [24] for the case of 4x4 contingency table. Furthermore, it is difficult to understand why after having written that the sample sizes have been derived for the null hypothesis of $k_0 = 0$, there are only two sample sizes calculated for testing $k_0 = 0.4$ vs. $k_1 = 0.6$ and $k_0 = 0.6$ vs. $k_1 = 0.8$.

Von Eye and Mun’s comment [24] about the decreasing of the sample sizes (Table 1.4) with the increase of the number of raters and the increase of the number of rating categories, is quite logical and expected. Indeed, this is a situation similar to that of a repeated measurements design in which the sample size decreases with the increase of the measurement occasions and also with the increase of the correlation among the repeated measurements. In addition, it is also consistent the increase of the sample sizes with the increase of the marginal non-uniformity.

A further comparison of the sample sizes shown in the Table 1.4 of von Eye and Mun’s book [24] can be done with the sample sizes calculated for only two raters according to Flack et al. [74] or, equivalently by PASS (13/16) [75] as the following table shows for with three categories with null hypothesis: $k_0 = 0.4$; alternative hypothesis, $k_1 = 0.6$, $\alpha = 0.05$ (two-tailed), and power = 0.80.

Between brackets there are the sample sizes calculated according to Flack et al. [74] or Pass (13/16) [75] for $\alpha = 0.05$ (two-tailed and one tailed, respectively).

Marginal probabilities			**Number of raters (2)**
π_1	π_2	π_3	
0.10	0.10	0.80	205 (219 - 171)
0.10	0.40	0.50	127 (145 - 113)
0.33	0.33	0.33	107 (101 - 79)

It is very well evident that the sample sizes calculated with the maximum value of the kappa variance according to Flack et al. [74] or PASS (13/16) [75] are greater in the case of non-uniformity of the marginal and lower in the case of marginal uniformity.

A different pattern occurs for the third part of the Table 1.4, always with three categories, with null hypothesis: $k_0 = 0.6$, alternative hypothesis: $k_1 = 0.8$, $\alpha = 0.05$ (two-tailed), power = 0.80 in the case of two raters.

Between brackets there are the sample sizes calculated according to Flack et al. [74] or Pass (13/16) [75] for $\alpha = 0.05$ (two-tailed and one tailed, respectively).

Marginal probabilities			**Number of raters (2)**
π_1	π_2	π_3	
0.10	0.10	0.80	172 (162 - 126)
0.10	0.40	0.50	102 (98 - 76)
0.33	0.33	0.34	87 (75- 58)

Indeed, rather surprisingly, the sample sizes calculated with the maximum value of the kappa variance according to Flack et al. [74] or PASS (13/16) [75] are always lower.

The following table shows the results for the case of the null hypothesis: $k_0 = 0.4$, alternative hypothesis: $k_1 = 0.6$, $\alpha = 0.05$ (two-tailed), and power = 0.80, corresponding to the second part of the Table 1.4 of von Eye's and Mun's book [24].

Marginal probabilities				**Number of raters (2)**
π_1	π_2	π_3	π_4	
0.10	0.10	0.10	0.70	102(155 - 121)
0.10	0.30	0.30	0.30	88(97 - 76)
0.25	0.25	0.25	0.25	60(83 - 65)

It is very well evident that the sample size values are always greater.

The following table show the sample sizes for the null hypothesis: $k_0 = 0.6$, alternative hypothesis: $k_1 = 0.8$, $\alpha = 0.05$ (two-tailed), power = 0.80, corresponding to the fourth part of the Table 1.4 of von Eye and Mun’s book [24] in the case of only two raters.

Marginal probabilities				Number of raters(2)
π_1	π_2	π_3	π_4	
0.10	0.10	0.10	0.70	157 (114 - 88)
0.10	0.30	0.30	0.30	88 (71 - 55)
0.25	0.25	0.25	0.25	60 (64 - 50)

In this case there is an erratic pattern with smaller values in the case of an important marginal non-uniformity and a little greater value (64 instead of 60)in the case of uniformity.

The above discrepancies make it difficult to rely on the sample sizes shown in von Eye and Mun’s Table 1.4 [24] in planning an agreement study on qualitative variables to be analyzed by means of Cohen’s kappa also for the case of two raters.

Furthermore, Indurkhya et al.’s paper [96] is not actually present and, consequently downloadable from the internet site (https://www.dgps.de/fachgruppen/methoden/mpr-online/) of the journal Methods of Psychological Research (MPR-online) among the three papers of the last issue (N.1, Vol.9, 2004) of MPR-online. Indeed, MPR-online at the end of 2004 was merged with the Spanish journal "*Metodologia de las Ciencias del Comportamiento*" leading to a new journal called "Methodology European Journal of Research Methods for the Behavioral and Social Sciences” that is the official organ of the European Association of Methodology (EAM), published by Hogrefe & Huber until 2019 and then by ZPID on their PsychOpen platform.

Appendix D. R code for the sample size calculation with a number of raters equal or more than two with only a function.

The following is the code of a function in R written for unifying together the four functions (PowerBinary, Power3Cats, Power4Cats, and Power5Cats) of the Package ‘kappaSize’, November 26, 2018, Version 1.2, Date 2018-11-25, Title: Sample Size Estimation Functions for Studies of Interobserver Agreement Author: Michael A. Rotondi, Maintainer: Michael A. Rotondi [94].

The “kappaSize” package has to be loaded

Parameters for this example:

pmarg<-c(.6,.3,.1); kappa<-.6; nr<-2 names(pmarg)<-paste("p",1:n_cat,sep="")

In the vector c(), it has to insert the marginal probabilities (equal of the rows and of the column); kappa: it has to insert the value; “nr” is the number of the raters, 2 in this example.

##

#Recursive Function for calculating the probability of the agreement or not according to Donner et al.

##

```
# Calculus of only one probability

 P_agree<-function(prob,kappa,nr) {
if (nr==1) P<-prob else P<-P_agree(prob,kappa,nr=nr-1)*(prob*(1-kappa)+(nr-1)*kappa)/(1+(nr-2)*kappa)
return(P)
}
# Calculus of all the probabilities.
Probs_agree<-function(pmarg,kappa,nr) {
n_cat<-length(pmarg)
# calculus of the probabilities
probs<-sapply(X = pmarg,FUN = P_agree,kappa=kappa,nr=nr)
P0<-1-sum(probs)
# output
result<-c(P0,probs)
names(result)<-0:n_cat
return(result)
}
#######################################################################################
# Function for calculating the non-central parameter of a | 2 distribution with degree of freedom (df) equal to 1
(df = 1)
#######################################################################################
lambda<-function(alpha,power,two_one_tail,df=1) {
if (two_one_tail==1) alpha<-alpha*2
lambda<-uniroot(f = function(x) {pchisq(q = qchisq(p = 1-alpha,df = df,ncp = 0),
df = df,ncp = x,lower.tail = F)-power},interval = c(0,100))$root
return(lambda=lambda)
}
###########################################################################################
## Function for calculating the Sample Size for the comparison between two kappa values
###########################################################################################
kappa_ssize_Donner<-function(pmarg,nr,k0,kA,alpha,power,two_one_tail,out=1) {
# calculus of the probabilities under the null hypothesis (H0) and the alternative hypothesis (HA)
pi_H0<-Probs_agree(pmarg = pmarg, kappa = k0, nr = nr)
pi_HA<-Probs_agree(pmarg = pmarg, kappa = kA, nr = nr)
# calculus of the non-centrality parameter   (lamba)
lmbda<-lambda(alpha = alpha, power = power, two_one_tail = 2,df=1)
# calculus of the sample size
Den<-sum((pi_HA-pi_H0)^2/pi_H0)
N<-ceiling(lmbda/Den)
# output
if (out==1) return(ssize=N) else {
ris2<-cbind(pmarg=t(pmarg), n_raters=nr, alpha=alpha, power=power, tail=two_one_tail, ssize=N)
nomi_pr<-paste("p",1:length(pmarg), sep="")
colnames(ris2)[1:length(pmarg)]<-nomi_pr
if (out==2) return(ris2) else return(cbind(ris2, lambda=lmbda, Den=Den))
}
}
```

Example 1

"kappa_ssize_Donner" is the function written by one of the authors (PA) in R and "Power3Cats" is the name of the function of the Package 'kappaSize'[94].

```
> kappa_ssize_Donner(pmarg = c(.6,.3,.1),nr = 3,k0 = .4,kA = .6,alpha = .05,power = .8,two_one_tail = 2,out = 2)
p1 p2 p3 n_raters k0 kA alpha power tail ssize
[1,] 0.6 0.3 0.1 3 0.4 0.6 0.05 0.8 2 74
```

The above output, from our function, reports the marginal probabilities, the number of raters, the two value of kappa under the null and the alternative hypothesis, respectively, alpha ( ) power (1 –), number of tails and, finally, the sample size.

```
> Power3Cats(kappa0 = .4,kappa1 = .6,props = c(.6,.3,.1),raters = 3,alpha = .05,power = .8)
A minimum of 74 subjects are required for this study of interobserver agreement.
```

The above output is from the function of the Package 'kappaSize'

Example 2

```
> kappa_ssize_Donner(pmarg = c(.3,.3,.2,.1,.1), nr = 6,k0 = .4, kA = .6,alpha = .05, power = .8, two_one_tail = 2, out = 2)
p1 p2 p3 p4 p5 n_raters k0 kA alpha power tail ssize
[1,] 0.3 0.3 0.2 0.1 0.1 6 0.4 0.6 0.05 0.8 2 25
```

The above output, from our function, reports the marginal probabilities, the number of raters, the values of Cohen's kappa under the null (k0) and the alternative hypothesis (kA), respectively, the significance level (alpha), the power (power), the number of tails (tail) and, finally, the sample size (ssize).

```
> Power5Cats(kappa0 = .4,kappa1 = .6,props = c(.3,.3,.2,.1,.1),raters = 6,alpha = .05,power = .8)
A minimum of 25 subjects are required for this study of interobserver agreement.
Warning: At least one expected cell count is less than five.
Warning: At least one expected cell count is less than five.
Warning: At least one expected cell count is less than five
```

The above output is from the function "Power5Cats" of the Package 'kappaSize'

Appendix E. R function for the sample size calculation in the case of the Full Common Correlation Model (FCCM)

No packages have to be upload for running the following R code.

```
kappa_ssize_PA<-function(pmarg, kappa0,kappaA,alpha=.05,power=.8,two_one_tail=1,out=1) {
n_cat<-length(pmarg)
names(pmarg)<-paste("p",1:n_cat,sep = "")
if (abs(sum(pmarg) - 1) >= 0.001)
stop("Sorry, the three proportions must sum to one.")
for (i in 1:n_cat) {
if ((pmarg[i] >= 1) || (pmarg[i] <= 0))
stop("Sorry, the proportion, pmarg must lie within (0,1).")
}
if ((kappa0 >= 1) || (kappa0 <= 0) || (kappaA <= 0) || (kappaA >= 1))
stop("Sorry, the null and alternative values of kappa must lie within (0,1).")
if ((alpha >= 1) || (alpha <= 0) || (power <= 0) || (power >= 1))
stop("Sorry, the alpha and power must lie within (0,1).")
k0<-kappa0;kA<-kappaA
# probabilities table under H0
p_diag_H0<-k0*pmarg*(1-pmarg)+pmarg^2
PP<-outer(pmarg,pmarg,FUN = "*")
p_off_diag_H0<-(1-k0)*(PP-diag(diag(PP)))
Pij_H0<-p_off_diag_H0; diag(Pij_H0)<-p_diag_H0
# probabilities table under HA
p_diag_HA<-kA*pmarg*(1-pmarg)+pmarg^2
PP<-outer(pmarg,pmarg,FUN = "*")
p_off_diag_HA<-(1-kA)*(PP-diag(diag(PP)))
Pij_HA<-p_off_diag_HA; diag(Pij_HA)<-p_diag_HA
# calculus of the kappa variance by means of the function kappa_Fleiss of R.
```

```
kappa0_rst<-kappa_Fleiss(x = Pij_H0,n_obs = 1)
kappaA_rst<-kappa_Fleiss(x = Pij_HA,n_obs = 1)
varK_H0<-kappa0_rst[,"var_K"]
varK_HA<-kappaA_rst[,"var_K"]
if (two_one_tail==1) {
z_alpha<-qnorm(p = 1-alpha)
z_beta<-qnorm(p =power)
} else {
z_alpha<-qnorm(p = 1-alpha/2)
z_beta<-qnorm(p = power)
}
# calculus of the sample size
N<-(z_alpha*sqrt(varK_H0)+z_beta*sqrt(varK_HA))^2/(kappaA- kappa0)^2
# output
out1<-cbind(size=N); rownames(out1)<-NULL
out2<-cbind(size=N,alpha=alpha,power=power,kappa0=kappa0,kappaA=kappaA,
rating_probs=t(pmarg),tau_H0=sqrt(varK_H0),tau_HA=sqrt(varK_HA),varK_H0=varK_H0,varK_HA=varK_HA)
rownames(out2)<-NULL
if (out==1) return(out1) else return(out2)}
# kappa_Fleiss R function for the calculus of the Cohen’s kappa and its variance.
kappa_Fleiss<-function(x,n_obs=NULL,W=NULL) {
# x table of the relative frequencies of the absolute frequencies (for this latter case, n_obs has to be put to NULL)
if (is.null(n_obs)) n<-sum(x) else n<-n_obs
ncat<-dim(x)[1]
M_prob<-x/sum(x)
## two weight matrices frequently used
# W1 whose elements are: w1(i,j)=1-abs(i-j)/(n_cat-1)
# W2 whose elements are: w2(i,j)=1-(abs(i-j)/(n_cat-1))^2
indx<-which(x>=0,arr.ind = T)
W1<-1-matrix(abs(indx[,"row"]-indx[,"col"])/(ncat-1),ncat,ncat)
W2<-1-matrix((abs(indx[,"row"]-indx[,"col"])/(ncat-1))^2,ncat,ncat)
if (is.null(W)) W<-diag(1,ncat,ncat)
if (!is.matrix(W) ){
```

```
W<-switch(W,
W<-W1,
W<-W2)
}
### calculus of Cohen’s kappa
pmarg_row<-apply(X = M_prob,MARGIN = 1,FUN = sum)
pmarg_column<-apply(X = M_prob,MARGIN = 2,FUN = sum)
M_prob_Ind<- outer(pmarg_row,pmarg_column,FUN = "*")
p0<-sum(W*M_prob)
pc<-sum(W*M_prob_Ind)
kappa<-(p0-pc)/(1-pc)
# calculus of the kappa variance according to the formula 8 (page 324) of Fleiss et al.’s paper [50].
wmarg_row<-apply(X = W*pmarg_column,MARGIN = 2,FUN = sum)
wmarg_column<-apply(X = W*pmarg_row,MARGIN = 2,FUN = sum)
M_marg_sum<-outer(wmarg_row,wmarg_column,FUN = "+")
A<-1/(n*(1-pc)^4)
B1<-W*(1-pc) # matrice {wij*(1-pc)}
B2<-M_marg_sum*(1-p0) #matrix {(wi.+w.j)*(1-p0)} with wi.=sum_j(wij*p.j)
B<-sum(M_prob*(B1-B2)^2)
#double summation within the curly brackets of the formula 8, page 324
C<- (p0*pc-2*pc+p0)^2 # second addendun within the square brackets { } of the previously quoted formula 8, page 324, [50]
var_K<-A*(B-C)
if (var_K<0 | is.nan(var_K)) var_K<-0
out1<-cbind(size=n,po=p0, pe=pc,kappa=kappa,var_K=var_K)
return(out)
}
```

Table SS1: Sample Sizes for k_0, k_A, marginal of a 2x2 contingency table, $\alpha = 0.05$, and power = 0.80.

k_0	k_A	π_1	π_2	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
				1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.4	0.45	0.5	0.5	2,042	2,597	2,078	2,638	2,042	2,597	2,042	2,597	2,042	2,597
0.4	0.50	0.5	0.5	501	638	520	660	501	638	501	638	501	638
0.4	0.60	0.5	0.5	119	153	130	165	119	153	119	153	119	153
0.4	0.70	0.5	0.5	50	64	58	74	50	64	50	64	50	64
0.4	0.80	0.5	0.5	26	34	33	42	26	34	26	34	26	34
0.4	0.90	0.5	0.5	15	19	21	27	15	19	15	19	15	19
0.4	0.95	0.5	0.5	11	15	18	22	11	15	11	15	11	15
0.5	0.55	0.5	0.5	1,811	2,305	1,855	2,355	1,811	2,305	1,811	2,305	1,811	2,305
0.5	0.60	0.5	0.5	441	563	464	589	441	563	441	563	441	563
0.5	0.70	0.5	0.5	103	133	116	148	103	133	103	133	103	133
0.5	0.80	0.5	0.5	42	54	52	66	42	54	42	54	42	54
0.5	0.90	0.5	0.5	21	27	29	37	21	27	21	27	21	27
0.5	0.95	0.5	0.5	15	19	23	30	15	19	15	19	15	19
0.6	0.65	0.5	0.5	1,530	1,950	1,583	2,010	1,530	1,950	1,530	1,950	1,530	1,950
0.6	0.70	0.5	0.5	368	471	396	503	368	471	368	471	368	471
0.6	0.80	0.5	0.5	83	108	99	126	83	108	83	108	83	108
0.6	0.90	0.5	0.5	32	42	44	56	32	42	32	42	32	42
0.6	0.95	0.5	0.5	21	28	33	42	21	28	21	28	21	28
0.4	0.45	0.6	0.4	2,121	2,698	2,148	2,727	2,121	2,698	2,121	2,698	2,121	2,698
0.4	0.50	0.6	0.4	520	663	537	682	520	663	520	663	520	663
0.4	0.60	0.6	0.4	124	159	135	171	124	159	124	159	124	159
0.4	0.70	0.6	0.4	52	67	60	76	52	67	52	67	52	67
0.4	0.80	0.6	0.4	27	35	34	43	27	35	27	35	27	35
0.4	0.90	0.6	0.4	15	20	22	28	15	20	15	20	15	20
0.4	0.95	0.6	0.4	11	15	18	23	11	15	11	15	11	15
0.5	0.55	0.6	0.4	1,887	2,402	1,924	2,442	1,887	2,402	1,887	2,402	1,887	2,402
0.5	0.60	0.6	0.4	459	586	481	611	459	586	459	586	459	586
0.5	0.70	0.6	0.4	107	138	121	153	107	138	107	138	107	138
0.5	0.80	0.6	0.4	44	57	54	68	44	57	44	57	44	57

k_0	k_A	π_1	π_2	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
				1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.5	0.90	0.6	0.4	21	28	31	39	21	28	21	28	21	28
0.5	0.95	0.6	0.4	15	20	24	31	15	20	15	20	15	20
0.6	0.65	0.6	0.4	1,597	2,035	1,645	2,088	1,597	2,035	1,597	2,035	1,597	2,035
0.6	0.70	0.6	0.4	384	492	412	522	384	492	384	492	384	492
0.6	0.80	0.6	0.4	87	113	103	131	87	113	87	113	87	113
0.6	0.90	0.6	0.4	33	44	46	58	33	44	33	44	33	44
0.6	0.95	0.6	0.4	22	29	34	43	22	29	22	29	22	29
0.4	0.45	0.8	0.2	3,110	3,953	3,032	3,849	3,110	3,953	3,110	3,953	3,110	3,953
0.4	0.50	0.8	0.2	766	975	758	963	766	975	766	975	766	975
0.4	0.60	0.8	0.2	183	235	190	241	183	235	183	235	183	235
0.4	0.70	0.8	0.2	77	99	85	107	77	99	77	99	77	99
0.4	0.80	0.8	0.2	39	51	48	61	39	51	39	51	39	51
0.4	0.90	0.8	0.2	22	29	31	39	22	29	22	29	22	29
0.4	0.95	0.8	0.2	16	22	26	32	16	22	16	22	16	22
0.5	0.55	0.8	0.2	2,839	3,612	2,783	3,532	2,839	3,612	2,839	3,612	2,839	3,612
0.5	0.60	0.8	0.2	692	883	696	883	692	883	692	883	692	883
0.5	0.70	0.8	0.2	162	208	174	221	162	208	162	208	162	208
0.5	0.80	0.8	0.2	66	85	78	99	66	85	66	85	66	85
0.5	0.90	0.8	0.2	32	42	44	56	32	42	32	42	32	42
0.5	0.95	0.8	0.2	22	30	35	44	22	30	22	30	22	30
0.6	0.65	0.8	0.2	2,437	3,105	2,418	3,069	2,437	3,105	2,437	3,105	2,437	3,105
0.6	0.70	0.8	0.2	586	750	605	768	586	750	586	750	586	750
0.6	0.80	0.8	0.2	133	172	152	192	133	172	133	172	133	172
0.6	0.90	0.8	0.2	51	67	68	86	51	67	51	67	51	67
0.6	0.95	0.8	0.2	33	44	50	63	33	44	33	44	33	44
0.4	0.45	0.9	0.1	5,418	6,883	5,092	6,465	5,418	6,883	5,418	6,883	5,418	6,883
0.4	0.50	0.9	0.1	1338	1,702	1,273	1,617	1,338	1,702	1,338	1,702	1,338	1,702
0.4	0.60	0.9	0.1	321	411	319	405	321	411	321	411	321	411
0.4	0.70	0.9	0.1	134	173	142	180	134	173	134	173	134	173
0.4	0.80	0.9	0.1	69	89	80	102	69	89	69	89	69	89
0.4	0.90	0.9	0.1	38	50	51	65	38	50	38	50	38	50

k_0	k_A	π_1	π_2	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
				1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.4	0.95	0.9	0.1	28	38	43	54	28	38	28	38	28	38
0.5	0.55	0.9	0.1	5,060	6,437	4,786	6,076	5,060	6,437	5,060	6,437	5,060	6,437
0.5	0.60	0.9	0.1	1,235	1,576	1,197	1,519	1,235	1,576	1,235	1,576	1,235	1,576
0.5	0.70	0.9	0.1	289	372	300	380	289	372	289	372	289	372
0.5	0.80	0.9	0.1	117	152	133	169	117	152	117	152	117	152
0.5	0.90	0.9	0.1	57	75	75	95	57	75	57	75	57	75
0.5	0.95	0.9	0.1	40	53	60	76	40	53	40	53	40	53
0.6	0.65	0.9	0.1	4,397	5,602	4,221	5,359	4,397	5,602	4,397	5,602	4,397	5,602
0.6	0.70	0.9	0.1	1,058	1,354	1,056	1,340	1,058	1,354	1,058	1,354	1,058	1,354
0.6	0.80	0.9	0.1	239	309	264	335	239	309	239	309	239	309
0.6	0.90	0.9	0.1	91	120	118	149	91	120	91	120	91	120
0.6	0.95	0.9	0.1	59	79	87	110	59	79	59	79	59	79

SS-A&C-full are the same as SS-A&C-max and SS-A&C-min.

Table SS2: Sample Sizes for k_0, k_A, marginal of a 3x3 contingency table, α = 0.05, and power = 0.80.

k_0	k_A	π_1	π_2	π_3	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
					1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.4	0.45	0.33	0.33	0.33	1,321	1,679	1,336	1,696	1,321	1,679	1,321	1,679	1,321	1,679
0.4	0.50	0.33	0.33	0.33	326	415	334	424	326	415	326	415	326	415
0.4	0.60	0.33	0.33	0.33	79	100	84	106	79	100	79	100	79	100
0.4	0.70	0.33	0.33	0.33	33	43	38	48	33	43	33	43	33	43
0.4	0.80	0.33	0.33	0.33	17	22	21	27	17	22	17	22	17	22
0.4	0.90	0.33	0.33	0.33	10	13	14	17	10	13	10	13	10	13
0.4	0.95	0.33	0.33	0.33	7	10	12	15	7	10	7	10	7	10
0.5	0.55	0.33	0.33	0.33	1,214	1,544	1,237	1,570	1,214	1,544	1,214	1,544	1,214	1,544
0.5	0.60	0.33	0.33	0.33	297	378	310	393	297	378	297	378	297	378
0.5	0.70	0.33	0.33	0.33	70	90	78	99	70	90	70	90	70	90
0.5	0.80	0.33	0.33	0.33	29	37	35	44	29	37	29	37	29	37
0.5	0.90	0.33	0.33	0.33	14	19	20	25	14	19	14	19	14	19
0.5	0.95	0.33	0.33	0.33	10	13	16	20	10	13	10	13	10	13
0.6	0.65	0.33	0.33	0.33	1,057	1,346	1,089	1,382	1,057	1,346	1,057	1,346	1,057	1,346
0.6	0.70	0.33	0.33	0.33	255	326	273	346	255	326	255	326	255	326
0.6	0.80	0.33	0.33	0.33	58	75	69	87	58	75	58	75	58	75
0.6	0.90	0.33	0.33	0.33	22	29	31	39	22	29	22	29	22	29
0.6	0.95	0.33	0.33	0.33	15	20	23	29	15	20	15	20	15	20
0.4	0.45	0.50	0.25	0.25	1,551	1,972	1,429	1,814	1,426	1,812	1,426	1,812	926	1,174
0.4	0.50	0.50	0.25	0.25	381	486	358	454	352	448	352	448	233	295
0.4	0.60	0.50	0.25	0.25	91	117	90	114	85	108	85	108	58	73
0.4	0.70	0.50	0.25	0.25	38	49	40	51	36	46	36	46	25	32
0.4	0.80	0.50	0.25	0.25	20	26	23	29	19	24	19	24	13	17
0.4	0.90	0.50	0.25	0.25	11	15	15	19	11	14	11	14	7	10
0.4	0.95	0.50	0.25	0.25	8	11	12	15	8	10	8	10	6	7
0.5	0.55	0.50	0.25	0.25	1,398	1,779	1,325	1,682	1,311	1,667	1,311	1,667	963	1,223
0.5	0.60	0.50	0.25	0.25	341	435	332	421	320	409	320	409	239	304
0.5	0.70	0.50	0.25	0.25	80	103	83	106	76	97	76	97	58	74
0.5	0.80	0.50	0.25	0.25	33	42	37	47	31	40	31	40	24	31

k_0	k_A	π_1	π_2	π_3	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
					1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.5	0.90	0.50	0.25	0.25	16	21	21	27	15	20	15	20	12	15
0.5	0.95	0.50	0.25	0.25	11	15	17	21	11	14	11	14	8	11
0.6	0.65	0.50	0.25	0.25	1,196	1,523	1,167	1,481	1,140	1,452	1,140	1,452	919	1,168
0.6	0.70	0.50	0.25	0.25	288	368	292	371	275	352	275	352	224	286
0.6	0.80	0.50	0.25	0.25	65	85	73	93	63	81	63	81	52	67
0.6	0.90	0.50	0.25	0.25	25	33	33	42	24	32	24	32	20	26
0.6	0.95	0.50	0.25	0.25	16	22	24	31	16	21	16	21	13	17
0.4	0.45	0.60	0.30	0.10	2,058	2,618	1,699	2,157	1,719	2,185	1,719	2,185	1,098	1,396
0.4	0.50	0.60	0.30	0.10	504	642	425	540	424	539	424	539	271	345
0.4	0.60	0.60	0.30	0.10	120	153	107	135	102	130	102	130	65	83
0.4	0.70	0.60	0.30	0.10	50	64	48	60	43	55	43	55	28	36
0.4	0.80	0.60	0.30	0.10	26	34	27	34	22	29	22	29	15	20
0.4	0.90	0.60	0.30	0.10	15	19	17	22	12	16	12	16	9	11
0.4	0.95	0.60	0.30	0.10	11	15	15	18	9	12	9	12	7	9
0.5	0.55	0.60	0.30	0.10	1,801	2,293	1,572	1,996	1,572	2,000	1,572	2,000	1,002	1,275
0.5	0.60	0.60	0.30	0.10	437	559	393	499	384	490	384	490	245	312
0.5	0.70	0.60	0.30	0.10	102	131	99	125	90	116	90	116	60	77
0.5	0.80	0.60	0.30	0.10	41	54	44	56	37	48	37	48	25	33
0.5	0.90	0.60	0.30	0.10	20	27	25	32	18	24	18	24	13	16
0.5	0.95	0.60	0.30	0.10	14	19	20	25	13	17	13	17	9	12
0.6	0.65	0.60	0.30	0.10	1,501	1,913	1,378	1,750	1,359	1,731	1,359	1,731	875	1,114
0.6	0.70	0.60	0.30	0.10	360	461	345	438	328	419	328	419	217	276
0.6	0.80	0.60	0.30	0.10	81	105	87	110	75	96	75	96	51	66
0.6	0.90	0.60	0.30	0.10	31	41	39	49	29	38	29	38	20	26
0.6	0.95	0.60	0.30	0.10	20	27	29	36	19	25	19	25	13	17
0.4	0.45	0.80	0.10	0.10	2,893	3,677	2,589	3,287	2,731	3,470	2,731	3,470	200	244
0.4	0.50	0.80	0.10	0.10	714	908	648	822	675	859	675	859	66	79
0.4	0.60	0.80	0.10	0.10	171	219	162	206	163	208	163	208	22	25
0.4	0.70	0.80	0.10	0.10	72	92	72	92	68	87	68	87	10	12
0.4	0.80	0.80	0.10	0.10	37	48	41	52	35	46	35	46	6	7
0.4	0.90	0.80	0.10	0.10	21	27	26	33	20	26	20	26	3	4

k_0	k_A	π_1	π_2	π_3	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
					1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.4	0.95	0.80	0.10	0.10	15	20	22	28	14	19	14	19	2	2
0.5	0.55	0.80	0.10	0.10	2,671	3,398	2,449	3,109	2,554	3,249	2,554	3,249	690	867
0.5	0.60	0.80	0.10	0.10	652	831	613	778	624	796	624	796	184	229
0.5	0.70	0.80	0.10	0.10	153	196	154	195	147	188	147	188	49	60
0.5	0.80	0.80	0.10	0.10	62	80	69	87	59	77	59	77	21	26
0.5	0.90	0.80	0.10	0.10	30	40	39	49	29	38	29	38	10	13
0.5	0.95	0.80	0.10	0.10	21	28	31	39	20	27	20	27	7	9
0.6	0.65	0.60	0.30	0.10	2,307	2,939	2,174	2,760	2,231	2,842	2,231	2,842	1,010	1,278
0.6	0.70	0.60	0.30	0.10	555	710	544	690	538	688	538	688	255	323
0.6	0.80	0.60	0.30	0.10	126	162	136	173	122	158	122	158	62	78
0.6	0.90	0.60	0.30	0.10	48	63	61	77	46	61	46	61	24	31
0.6	0.95	0.60	0.30	0.10	31	42	45	57	30	40	30	40	15	20

SS-A&C-full are the same as SS-A&C-max and SS-A&C-min.

Table SS3: Sample Sizes for k0, kA, marginal of a 4x4 contingency table, α = 0.05, and power = 0.80.

k_0	k_A	π_1	π_2	π_3	π_4	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
						1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.4	0.45	0.25	0.25	0.25	0.25	1,081	1,373	1,089	1,382	1,081	1,373	1,081	1,373	1,081	1,373
0.4	0.50	0.25	0.25	0.25	0.25	268	340	273	346	268	340	268	340	268	340
0.4	0.60	0.25	0.25	0.25	0.25	65	83	69	87	65	83	65	83	65	83
0.4	0.70	0.25	0.25	0.25	0.25	28	35	31	39	28	35	28	35	28	35
0.4	0.80	0.25	0.25	0.25	0.25	14	19	18	22	14	19	14	19	14	19
0.4	0.90	0.25	0.25	0.25	0.25	8	11	11	14	8	11	8	11	8	11
0.4	0.95	0.25	0.25	0.25	0.25	6	8	9	12	6	8	6	8	6	8
0.5	0.55	0.25	0.25	0.25	0.25	1,015	1,290	1,031	1,309	1,015	1,290	1,015	1,290	1,015	1,290
0.5	0.60	0.25	0.25	0.25	0.25	249	317	258	328	249	317	249	317	249	317
0.5	0.70	0.25	0.25	0.25	0.25	59	76	65	82	59	76	59	76	59	76
0.5	0.80	0.25	0.25	0.25	0.25	24	31	29	37	24	31	24	31	24	31
0.5	0.90	0.25	0.25	0.25	0.25	12	16	17	21	12	16	12	16	12	16
0.5	0.95	0.25	0.25	0.25	0.25	9	11	13	17	9	11	9	11	9	11
0.6	0.65	0.25	0.25	0.25	0.25	899	1,145	924	1,173	899	1,145	899	1,145	899	1,145
0.6	0.70	0.25	0.25	0.25	0.25	218	278	231	294	218	278	218	278	218	278
0.6	0.80	0.25	0.25	0.25	0.25	50	64	58	74	50	64	50	64	50	64
0.6	0.90	0.25	0.25	0.25	0.25	19	25	26	33	19	25	19	25	19	25
0.6	0.95	0.25	0.25	0.25	0.25	13	17	19	24	13	17	13	17	13	17
0.4	0.45	0.4	0.3	0.2	0.1	1,520	1,934	1,197	1,519	1,208(-10)	1,535(-13)	1,194(4)	1,517(5)	811	1,029
0.4	0.50	0.4	0.3	0.2	0.1	372	475	300	380	299(-3)	380(3)	295(1)	376(1)	203	258
0.4	0.60	0.4	0.3	0.2	0.1	89	114	75	95	72	92(1)	71(1)	91	50	64
0.4	0.70	0.4	0.3	0.2	0.1	37	48	34	43	31(-1)	39	30	39	22	28
0.4	0.80	0.4	0.3	0.2	0.1	19	25	19	24	16	21(1)	16	20	12	15
0.4	0.90	0.4	0.3	0.2	0.1	11	14	12	16	9	12	9	12	7	9
0.4	0.95	0.4	0.3	0.2	0.1	8	11	10	13	7	9	7	9	5	6
0.5	0.55	0.4	0.3	0.2	0.1	1,337	1,702	1,128	1,432	1,123(-6)	1,428(-7)	1,115(2)	1,417(4)	812	1,031
0.5	0.60	0.4	0.3	0.2	0.1	325	415	282	358	275(-1)	350(-1)	273(1)	348(1)	201	255
0.5	0.70	0.4	0.3	0.2	0.1	76	98	71	90	65	83	65	83	48	62
0.5	0.80	0.4	0.3	0.2	0.1	31	40	32	40	27	34	27	34	20	26
0.5	0.90	0.4	0.3	0.2	0.1	15	20	18	23	13	17	13	17	10	13

k_0	k_A	π_1	π_2	π_3	π_4	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
						1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.5	0.95	0.4	0.3	0.2	0.1	11	14	14	18	9	12	9	12	7	9
0.6	0.65	0.4	0.3	0.2	0.1	1,121	1,428	1,006	1,278	986(-3)	1,256(-4)	982(1)	1,251(1)	758	965
0.6	0.70	0.4	0.3	0.2	0.1	269	345	252	320	238	305(-1)	238	304	185	236
0.6	0.80	0.4	0.3	0.2	0.1	61	79	63	80	55(-1)	70	54	70	43	56
0.6	0.90	0.4	0.3	0.2	0.1	23	31	28	36	21	28	21	28	17	22
0.6	0.95	0.4	0.3	0.2	0.1	15	20	21	27	14	18	14	18	11	15
0.4	0.45	0.6	0.2	0.1	0.1	1,889	2,403	1,463	1,857	1,511(-12)	1,920(-15)	1,496(3)	1,901(4)	513	646
0.4	0.50	0.6	0.2	0.1	0.1	463	590	366	465	373(-2)	475(-4)	370(1)	471	134	168
0.4	0.60	0.6	0.2	0.1	0.1	110	141	92	117	90	115(-1)	89(1)	114	35	44
0.4	0.70	0.6	0.2	0.1	0.1	46	59	41	52	38	49(-1)	38	48	16	20
0.4	0.80	0.6	0.2	0.1	0.1	24	31	23	30	20	26(-1)	20	25	9	11
0.4	0.90	0.6	0.2	0.1	0.1	13	18	15	19	11	15(-1)	11	14	5	6
0.4	0.95	0.6	0.2	0.1	0.1	10	13	13	16	8	11	8	11	4	5
0.5	0.55	0.6	0.2	0.1	0.1	1,665	2,120	1,384	1,757	1,405(-6)	1,788(-9)	1,397(2)	1,777(2)	658	833
0.5	0.60	0.6	0.2	0.1	0.1	405	517	346	440	344(-2)	438(-2)	342	436	167	211
0.5	0.70	0.6	0.2	0.1	0.1	94	121	87	110	81	104	81	104	42	53
0.5	0.80	0.6	0.2	0.1	0.1	38	50	39	49	33	43	33	43	18	23
0.5	0.90	0.6	0.2	0.1	0.1	19	25	22	28	16	21	16	21	9	11
0.5	0.95	0.6	0.2	0.1	0.1	13	18	18	22	12(-1)	15	11	15	6	8
0.6	0.65	0.6	0.2	0.1	0.1	1,395	1,777	1,233	1,566	1,231(-3)	1,568(-5)	1,227(1)	1,562(1)	706	896
0.6	0.70	0.6	0.2	0.1	0.1	335	429	309	392	297	380(-1)	296(1)	379	176	224
0.6	0.80	0.6	0.2	0.1	0.1	76	98	78	98	68	88(-1)	68	87	42	54
0.6	0.90	0.6	0.2	0.1	0.1	29	38	35	44	26	34	26	34	17	21
0.6	0.95	0.6	0.2	0.1	0.1	19	25	26	32	17	23	17	23	11	14
0.4	0.45	0.7	0.1	0.1	0.1	2,056	2,613	1,760	2,235	1,839	2,336	1,839	2,336	314	390
0.4	0.50	0.7	0.1	0.1	0.1	506	645	440	559	455	579	455	579	89	109
0.4	0.60	0.7	0.1	0.1	0.1	121	155	110	140	110	140	110	140	26	31
0.4	0.70	0.7	0.1	0.1	0.1	51	65	49	63	46	59	46	59	12	14
0.4	0.80	0.7	0.1	0.1	0.1	26	34	28	35	24	31	24	31	6	8
0.4	0.90	0.7	0.1	0.1	0.1	15	19	18	23	13	18	13	18	4	4
0.4	0.95	0.7	0.1	0.1	0.1	11	15	15	19	10	13	10	13	3	3

k_0	k_A	π_1	π_2	π_3	π_4	SS-Flack		SS-Donner		SS-A&C-max		SS-A&C-min		SS-Flack-min	
						1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed	1-tailed	2-tailed
0.5	0.55	0.7	0.1	0.1	0.1	1,881	2,393	1,676	2,128	1,726	2,196	1,726	2,196	644	812
0.5	0.60	0.7	0.1	0.1	0.1	459	585	419	532	423	539	423	539	168	210
0.5	0.70	0.7	0.1	0.1	0.1	108	138	105	133	100	128	100	128	43	54
0.5	0.80	0.7	0.1	0.1	0.1	44	57	47	60	41	53	41	53	18	23
0.5	0.90	0.7	0.1	0.1	0.1	21	28	27	34	20	26	20	26	9	11
0.5	0.95	0.7	0.1	0.1	0.1	15	20	21	27	14	19	14	19	6	8
0.6	0.65	0.7	0.1	0.1	0.1	1,617	2,061	1,498	1,902	1,517	1,932	1,517	1,932	813	1,030
0.6	0.70	0.7	0.1	0.1	0.1	389	498	375	476	366	468	366	468	203	258
0.6	0.80	0.7	0.1	0.1	0.1	88	114	94	119	83	108	83	108	49	62
0.6	0.90	0.7	0.1	0.1	0.1	34	44	42	53	32	42	32	42	19	24
0.6	0.95	0.7	0.1	0.1	0.1	22	29	31	39	21	28	21	28	12	16

SS-A&C-full is lower than SS-A&C-max in 32 (22.2%) cases and greater than SS-A&C-min in 20 (13.9%) cases (difference shown between brackets); SS-A&C-full is equal in the remaining 112 (77.8%) and 124 (86.1%) cases, respectively.

Table SIM.1A: Simulation study on 3x3 table. Descriptive statistics of bias, absolute bias and percent bias for k_0, k_A and coverage. The two sample sizes, calculated for a power of 0.80 and 0.90, are considered as a whole and for each power value.

Variable	Mean (SD)	Lower-Upper 95% CI	Median	Min; Q1; Q3;Max
k_0 Bias (All)	-0.006874 (0.003905)	-0.008007; -0.005739	-0.006766	-0.016063; -0.009346;-0.003202;-0.001130
k_0 Bias (power = 0.80)	-0.007588 (0.004099)	-0.009317; -0.005856	-0.007674	-0.016063; -0.010473; -0.003339; -0.001860
k_0 Bias (power = 0.90)	-0.006160 (0.003647)	-0.007699; -0.004620	-0.006151	-0.012819; -0.008971; -0.002511; -0.001130
%k_0 Bias (All)	-1.393691 (0.780862)	-1.620429; -1.166952	-1.415729	-3.485200; -1.960566; -0.606375; -0.205600
%k_0 Bias (power = 0.80)	-1.543668 (0.831957)	-1.894972; -1.192362	-1.470725	-3.485200; -2.098204; -0.752820; -0.465000
%k_0 Bias (power = 0.90)	-1.243714 (0.711973)	-1.544353; -0.943073	-1.357725	-2.649300; -1.859612; -0.495470; -0.205600
k_0 Abs.Bias (All)	0.006874 (0.003905)	0.005739; 0.008007	0.006766	0.001130; 0.003202; 0.009346; 0.016063
k_0 Abs.Bias (power = 0.80)	0.007587 (0.004099)	0.005856 0.009317	0.007674	0.001860; 0.003339; 0.010473; 0.016063
k_0 Abs.Bias (power = 0.90)	0.006160 (0.003646)	0.004620 0.007699	0.006151	0.001130; 0.002511; 0.008971; 0.012819
%k_0 Abs.Bias (All)	1.393691 (0.780862)	1.166952 1.620429	1.415729	0.205600; 0.606375; 1.960566; 3.485200
%k_0 Abs.Bias (power = 0.80)	1.543668 (0.831957)	1.192362 1.894972	1.470725	0.465000; 0.752820; 2.098204; 3.485200
%k_0 Abs.Bias (power = 0.90)	1.243714 (0.711973)	0.943073 1.544353	1.357725	0.205600; 0.495470; 1.859612; 2.649300
k_A Bias (All)	-0.003596 (0.001938)	-0.004158; -0.003033	-0.002821	-0.008565; -0.004654; -0.002192; -0.001079
k_A Bias (power = 0.80)	-0.003984 (0.002079)	-0.004861; -0.003105	-0.003153	-0.008565; -0.005608; -0.002518; -0.001328
k_A Bias (power = 0.90)	-0.003209 (0.001743)	-0.003944; -0.002473	-0.002394	-0.007202; -0.004616 -0.001885; -0.001079

%k_A Bias (All)	-0.464093 (0.256484)	-0.538568; -0.389617	-0.350172	-1.070687; -0.664878; -0.264617; -0.134887
%k_A Bias (power = 0.80)	-0.514539 (0.276899)	-0.631464; -0.397615	-0.416925	-1.070687; -0.746368; -0.297283; -0.145021
%k_A Bias (power = 0.90)	-0.413646 (0.229006)	-0.510346; -0.316945	-0.307175	-0.900262; -0.659500; -0.245342; -0.134887
k_A Abs.Bias (All)	0.003596 (0.001938)	0.003033; 0.004158	0.002821	0.001079; 0.002192; 0.004654; 0.008565
k_AAbs.Bias (power=0.80)	0.003984 (0.002079)	0.003105; 0.004861	0.003153	0.001328; 0.002518; 0.005608; 0.008565
k_A Abs.Bias (power=0.90)	0.003209 (0.001743)	0.002473; 0.003944	0.002394	0.001079; 0.001885; 0.004616; 0.007202
%k_A Abs.Bias (All)	0.464093 (0.256484)	0.389617; 0.538568	0.350172	0.134887; 0.264617; 0.664878; 1.070687
%k_A Abs.Bias (power=0.80)	0.514539 (0.276899)	0.397615; 0.631464	0.416925	0.145021; 0.297283; 0.746368; 1.070687
%k_A Abs.Bias (power=0.90)	0.413646 (0.229006)	0.316945; 0.510346	0.307175	0.134887; 0.245342; 0.659500; 0.900262
Coverage Bias (All)	-0.018017 (0.010990)	-0.021207; -0.014825	-0.017050	-0.043700; -0.025500; -0.008500; -0.003800
Coverage Bias (power=0.80)	-0.020550 (0.012121)	-0.025668; -0.015432	-0.020900	-0.043700; -0.028450; -0.010000; -0.004100
Coverage Bias (power=0.90)	-0.015483 (0.009301)	-0.019410; -0.011555	-0.016700	-0.036900;-0.018800;-0.008000;-0.003800
%Coverage Bias (All)	-1.896491 (1.156848)	-2.232405; -1.560577	-1.794736	-4.600000; -2.684210; -0.894736; -0.400000
%Coverage Bias (power=0.80)	-2.163158 (1.275840)	-2.701898; -1.624417	-2.200000	-4.600000; -2.994736; -1.052631; -0.431578
%Coverage Bias (power=0.90)	-1.629825 (0.979086)	-2.043256; -1.216392	-1.757894	-3.884210; -1.978947; -0.842105; -0.400000
Coverage Abs.Bias (All)	0.018017 (0.010990)	0.014825; 0.021207	0.017050	0.003800; 0.0085000; 0.025500; 0.043700
Coverage Abs.Bias (power=0.80)	0.020550 (0.012121)	0.015432; 0.025668	0.020900	0.004100; 0.010000; 0.028450; 0.043700

Coverage Abs.Bias (power=0.90)	0.015483 (0.009301)	0.011555; 0.019410	0.016700	0.003800; 0.008000; 0.018800; 0.036900
%Coverage Abs.Bias (All)	1.896491 (1.156848)	1.560577; 2.232405	1.794736	0.400000; 0.894736;2.684210;4.600000
%Coverage Abs.Bias (power=0.80)	2.163158 (1.275840)	1.624417; 2.701898	2.200000	0.4315789; 1.052631; 2.994736; 4.600000
%Coverage Abs.Bias (power=0.90)	1.629825 (0.979086)	1.216392; 2.043256	1.757894	0.400000; 0.842105; 1.978947; 3.884210

Table SIM.1B: Table3x3: descriptive statistics of the bias, absolute bias and percent bias for the power. The two power values of 0.80 and 0.90 are considered separately.

Variable	Mean(SD)	Lower-Upper 95%CI	Median	Min; Q1; Q3; Max
Bias Power0.80	0.077450 (0.047589)	0.057354; 0.097545	0.068200	0.010500; 0.042550; 0.118750; 0.167700
Bias Power0.90	0.032587 (0.023486)	0.022669; 0.042505	0.028500	-0.004800; 0.015150;0 .056300; 0.071900
%Bias Power0.80	9.681250 (5.948723)	7.169324; 12.193175	8.525000	1.312500; 5.318750; 14.843750; 20.962500
%Bias Power0.90	3.620833 (2.609632)	2.518882; 4.722784	3.166666	-0.533333; 1.683333; 6.255555; 7.988888
Abs.Bias (Power0.80)	0.077450 (0.047589)	0.057354; 0.097545	0.068200	0.010500; 0.042550; 0.118750; 0.167700
Abs.Bias (Power0.90)	0.033245 (0.022503)	0.023743; 0.042748	0.028500	0.001100; 0.015150; 0.056300; 0.071900
%Abs.Bias (Power0.80)	9.681250 (5.948723)	7.169324; 12.193175	8.525000	1.312500; 5.318750; 14.843750; 20.962500
%Abs.Bias (Power0.90)	3.693981 (2.500370)	2.638167; 4.749795	3.166666	0.122222; 1.683333; 6.255555; 7.988888

SD = Standard deviation,Q1 and Q3=first and third quartile, respectively.

Table SIM.2A: Simulation study on 4x4 table. Descriptive statistics of bias, absolute bias and percent bias for k_0, k_A and coverage. The two sample sizes, calculated for a power of 0.80 and 0.90, are considered as a whole and for each power value.

Variable	Mean(SD)	Lower-Upper 95%CI	Median	Min; Q1; Q3; Max
k_0 Bias (All)	-0.005907 (0.005105)	-0.007389; -0.004424	-0.004883	-0.019671; -0.009196; -0.001939; 0.002820
k_0 Bias (power=0.80)	-0.006622 (0.005937)	-0.009129; -0.004114	-0.006003	-0.019671; -0.010656; -0.002150; 0.002820
k_0 Bias (power=0.90)	-0.005192 (0.004116)	-0.006930; -0.003454	-0.004709	-0.013749; -0.007920; -0.001764; 0.001112
%k_0 Bias (All)	-1.211066 (1.120276)	-1.536360; -0.885771	-0.876641	-4.917850; -1.769454; -0.440987;0 .705025
%k_0 Bias (power=0.80)	-1.374657 (1.301142)	-1.924081; -0.825233	-1.131550	-4.917850; -2.228579; -0.537725; 0.705025
%k_0 Bias (power=0.90)	-1.047474 (0.903205)	-1.428864; -0.666084	-0.847708	-3.437450; -1.474237; -0.406254; 0.278150
k_0 Abs.Bias (All)	0.006071 (0.004905)	0.004646 0.007495	0.004883	0.0000776; 0.002150; 0.009196; 0.019671
k_0 Abs.Bias (power=0.80)	0.006857 (0.005652)	0.004470; 0.009243	0.006003	0.0001931; 0.002322; 0.010656; 0.019671
k_0 Abs.Bias (power=0.90)	0.005285 (0.003991)	0.003599; 0.006970	0.004709	0.0000776; 0.001764; 0.007920; 0.013749
%k_0 Abs.Bias (All)	1.252031 (1.073298)	0.940378; 1.563685	0.876641	0.019400; 0.484812; 1.769454; 4.917850
%k_0 Abs.Bias (power=0.80)	1.433409 (1.233215)	0.912668; 1.954150	1.131550;	0.032183; 0.560612; 2.228579; 4.917850
%k_0 Abs.Bias (power=0.90)	1.070653 (0.874384)	0.701433; 1.439874	0.847708	0.019400; 0.406254; 1.474237; 3.437450
k_A Bias (All)	0.007687 (0.029336)	-0.000830; 0.016206	-0.002044	-0.009640; -0.003351; -0.000652; 0.098868
k_A Bias (power=0.80)	0.007286 (0.029792)	-0.005293; 0.019866	-0.002420	-0.009640; -0.003681; -0.000656; 0.098868
k_A Bias (power=0.90)	0.008089 (0.029508)	-0.004371; 0.020549	-0.001792	-0.005858; -0.003211; -0.000620; 0.098160

%k_A Bias (All)	0.910636 (3.641532)	-0.146754; 1.968026	-0.248581	-1.227200; -0.486185; -0.078649; 12.358512
%k_A Bias (power=0.80)	0.849386 (3.701895)	-0.713787; 2.412559	-0.288198	-1.227200; -0.604650; -0.069131; 12.358512
%k_A Bias (power=0.90)	0.971886 (3.658716)	-0.573054; 2.516826	-0.225488	-0.836928; -0.451283; -0.086266; 12.270000
k_A Abs.Bias (All)	0.012590 (0.027551)	0.004590; 0.020590	0.002651	0.0000542; 0.001363; 0.005011; 0.098868
k_A Abs.Bias (power=0.80)	0.012998 (0.027688)	0.001306; 0.024690	0.002864	0.0002192; 0.001430; 0.007061; 0.098868
k_A Abs.Bias (power=0.90)	0.012183 (0.028002)	0.000358; 0.024007	0.002520	0.0000542; 0.001295; 0.004007; 0.098160
%k_A Abs.Bias (All)	1.568907 (3.404969)	0.580207; 2.557607	0.378128	0.006775; 0.170962; 0.691685; 12.358512
%k_A Abs.Bias (power=0.80)	1.630192 (3.418154)	0.186832; 3.073552	0.429830	0.023073; 0.169538; 0.934675; 12.358512
Variable	Mean (SD)	Lower-Upper 95% CI	Median	Min; Q1; Q3;Max
%k_A Abs.Bias (power=0.90)	1.507621 (3.464115)	0.044854; 2.970389	0.319257	0.006775; 0.170962; 0.523843; 12.270000
Coverage Bias (All)	-0.017562 (0.014708)	-0.021833; -0.013291	-0.014000	-0.049000; -0.031500; -0.006000; 0.010000
Coverage Bias (power=0.80)	-0.018291 (0.014822)	-0.024550; -0.012032	-0.015500	-0.049000; -0.030000; -0.006500; 0.005000
Coverage Bias (power=0.90)	-0.016833 (0.014875)	-0.023114; -0.010552	-0.013500	-0.048000; -0.032500; -0.005000; 0.010000
%Coverage Bias (All)	-1.848684 (1.548250)	-2.298249; -1.399119	-1.473684	-5.157894; -3.315789; -0.631578; 1.052631
%Coverage Bias (power=0.80)	-1.925438 (1.560218)	-2.584261; -1.266616	-1.631578	-5.157894; -3.157894; -0.684210; 0.526315
%Coverage Bias (power=0.90)	-1.771929 (1.565823)	-2.433119; -1.110740	-1.421052	-5.052631; -3.421052; -0.526315; 1.052631
Coverage Abs.Bias (All)	0.018437 (0.013570)	0.014497; 0.022378	0.014000	0.0; 0.006500; 0.031500; 0.04900

Coverage Abs.Bias (power=0.80)	0.019041 (0.013801)	0.013213; 0.024869	0.015500	0.001000; 0.006500; 0.030000; 0.049000
Coverage Abs.Bias (power=0.90)	0.017833 (0.013605)	0.012088; 0.023578	0.013500	0.0; 0.007000; 0.032500; 0.048000
%Coverage Abs.Bias(All)	1.940789 (1.428503)	1.525995; 2.355583	1.473684	0.0; 0.684210; 3.31578; 5.157894
%Coverage Abs.Bias (power=0.80)	2.004386 (1.452770)	1.390934; 2.617837	1.631578	0.105263; 0.684210; 3.157894; 5.157894
%Coverage Abs.Bias (power=0.90)	1.877193 (1.432126)	1.272459; 2.481926	1.421052	0.0; 0.736842; 3.421052; 5.052631

Table SIM.2B: Table4x4: descriptive statistics of the bias, absolute bias and percent bias for the power. The two power values of 0.80 and 0.90 are considered separately.

Variable	Mean(SD)	Lower-Upper 95%CI	Median	Min; Q1; Q3; Max
Bias Power0.80	0.087333 (0.061442)	0.061388; 0.113278	0.067500	0.004000; 0.038000; 0.148000; 0.198000
Bias Power0.90	0.041500 (0.028664)	0.029396; 0.053603	0.036500	0.001000; 0.017500; 0.066500; 0.099000
%Bias Power0.80	10.916666 (7.680320)	7.673551; 14.159781	8.437500	0.500000; 4.750000; 18.500000; 24.750000
%BiasPower0.90	4.611111 (3.184941)	3.266228; 5.955994	4.055555	0.111111; 1.944444; 7.3888889; 11.000000
Abs. Bias (Power0.80)	0.087333 (0.061442)	0.061388; 0.113278	0.067500	0.004000; 0.038000; 0.148000; 0.198000
Abs. Bias (Power0.90)	0.041500 (0.028664)	0.029396; 0.053603	0.036500	0.001000; 0.017500; 0.066500; 0.099000
%Abs. Bias (Power0.80)	10.916666 (7.680320)	7.673551; 14.159781	8.437500	0.500000; 4.750000; 18.500000; 24.750000
%Abs. Bias (Power0.90)	4.611111 (3.184941)	3.266228; 5.955994	4.055555	0.111111; 1.944444; 7.388888; 11.000000