Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 10

5-1-2009

Beyond Kappa: Estimating Inter-Rater Agreement with Nominal Classifications

Nol Bendermacher
Radboud University, Nijmegen, The Netherlands, Bendermacher@hotmail.com

Pierre Souren
Radboud University, Nijmegen, The Netherlands, p.souren@socsci.ru.nl

Recommended Citation
Bendermacher, Nol and Souren, Pierre (2009) "Beyond Kappa: Estimating Inter-Rater Agreement with Nominal Classifications," *Journal of Modern Applied Statistical Methods*: Vol. 8 : Iss. 1 , Article 10.
DOI: 10.22237/jmasm/1241136540

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 110-121

1538 – 9472/09/$95.00

Beyond Kappa: Estimating Inter-Rater Agreement with Nominal Classifications

Nol Bendermacher Pierre Souren
Radboud University
Nijmegen, The Netherlands

Cohen's Kappa and a number of related measures can all be criticized for their definition of correction for chance agreement. A measure is introduced that derives the corrected proportion of agreement directly from the data, thereby overcoming objections to Kappa and its related measures.

Key words: Interrater agreement, Cohen's Kappa, nominal data, reliability.

Introduction

The most popular measure of inter-rater agreement in the case of nominal classification is Cohen's kappa (Cohen, 1960). Kappa is a member of a family of measures that are all defined by the same basic formula (Zwick, 1988):

$$A = \frac{f - p_c(A)}{1 - p_c(A)} \tag{1.1}$$

where f = the observed proportion of agreement and $p_c(A)$ = the definition of chance agreement according to measure A. The measures of this family differ only in their definitions of chance agreement $p_c(A)$.

Methodology

A General Model

Starting with n cases classified by two raters into c exhaustive and mutually exclusive categories, the population distribution of the c categories is given by the vector V. The joint distribution of the ratings is given by the c by c population matrix X. The model distinguishes three types of classifications: (1) a correct observation, (2) a correct guess, and (3) a wrong guess. The second type is a correct classification, but not a correct observation. The model assumes a fixed probability p_r that rater r makes a correct observation, i.e., a classification of type (1). Fixed means that p_i is independent of the true category V_i of the case and of its classification by the other rater. Rater agreement, as far as it is not based on chance, arises if both raters make a correct observation. Assuming that raters act independently, the probability of such non-chance-agreement is p_1p_2. Therefore a measure of inter-rater agreement is defined as: $s = p_1p_2$.

If rater r performs a correct observation, the probabilities of the categories are given by the population distribution V. However, if the rater does not, the classifications follow an error distribution W_r. The error distributions may differ from V and from each other. It is assumed that W_r is independent of the true category of the case. The model parameters are p_1, p_2, V, W_1 and W_2 as defined above. In order to simplify the formulas $q_r = 1-p_r$ and $D_r = W_r-V$ are also defined. This article will show that s and V can be estimated directly from the observed sample of classifications by the raters, without any assumptions regarding the error distributions W_1 and W_2.

Nol Bendermacher is a retired member of the Research Technical Support Group of the Faculty of the Social Sciences. Pierre Souren is a current member of this group. Email them at Bendermacher@hotmail.com and p.souren@socsci.ru.nl

Some Measures for Inter-rater Agreement

In formula (1.1), f is the proportion of cases classified in the same way by both raters, and $p_c(A)$ is the correction for chance agreement according to measure A. The denominator is a scaling factor restricting the measure to a maximum of 1.

Bennett, Alpert and Goldstein (1954) assumed that a rater who does not recognize the true category of a case draws from a uniform distribution, thus giving each category an equal chance. In terms of the general model, they assume $W_1 = W_2 = \left(\frac{1}{c},\frac{1}{c},...,\frac{1}{c},\frac{1}{c}\right)^T$. If both raters draw from this common error distribution, the probability of chance agreement on one specific category is $\frac{1}{c^2}$ and the overall expected proportion of chance agreement is $\frac{c}{c^2} = \frac{1}{c}$. Therefore, Bennett, Alpert and Goldstein (1954) defined the correction for chance agreement as $p_c(A) = \frac{1}{c}$ for their measure S.

At least two objections to this choice exist:

1. In many situations it is plausible that the true distribution V of the cases deviates from uniformity and that the raters, knowing so, adjust their guessing distributions accordingly.
2. Scott (1955) objected that if W_1 and/or W_2 deviate from uniformity, the proportion of agreement by chance will always be greater than $\frac{1}{c}$. In other words, $\frac{1}{c}$ is a lower limit for the proportion of agreement by chance, meaning that S is an upper bound for inter-rater agreement.

S has been presented several times under different names and different notations. For the case of two categories S is equal to the random error RE (Maxwell, 1977). With only two categories, this measure is equal to the difference between the proportion of agreement and the proportion of disagreement: $\frac{f}{1-f}$. For the general case, Brennan and Prediger (1981) reported the measure as κ_n, Zwick (1988) mentioned Guilford's G, for the two categories case, and Janson and Vegelius' C for the general case.

Scott (1955) tried to overcome the second objection by introducing the assumption that both raters, when guessing, follow the true distribution. In terms of the general model, Scott assumed that $W_1 = W_2 = V$. Therefore he estimated the distribution by the average of the two marginal distributions. His measure is called π and $p_c(\pi)$ is defined as $\sum_{i=1}^{c}\left(\frac{M_{1i}+M_{2i}}{2}\right)^2$, where M_{1i} and M_{2i} are the two observed marginal proportions of category i.

Cohen (1960) objected to Scott that one source of disagreement is precisely the tendency of the raters to spread their ratings differently over the categories: "one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments differently over the categories." Therefore, Cohen dropped the assumption of equal marginal distributions and defined the proportion of chance agreement as

$$\sum_{i=1}^{c} M_{1i}M_{2i} .$$

It can be seen, however, that the marginal distributions are a mix of the true distribution V and the error distributions W_1 and W_2, more precisely, $M_r = p_rV + q_rW_r$, so Cohen's estimation of chance agreement is only correct under the null hypothesis that p_1 and p_2 are both zero, or under the assumption that $W_1 = W_2 = V$. The latter assumption would mean that the two marginal distributions are equal, so Scott's π could be used as well. As Brennan and Prediger (1981) stated: "For *descriptive* purposes, therefore, when marginals are free it seems questionable to reduce observed agreement by $\sum P_{i.}P_{.i}$, which is directly dependent on agreement in the marginals" (p. 692). Other objections and alternatives to Kappa have also been brought forward. For details, readers are referred to Perreault and Leigh (1989) and Brennan and Prediger (1981).

The next section will elaborate on the formal model and investigate possibilities to identify and estimate the model parameters.

What is special to this approach is that the inter-rater agreement is estimated without any assumptions regarding the rater distributions W_1 and W_2. In addition, a short outline of an algorithm that performs the required calculations is provided and an extension for the case of three simultaneous raters is introduced. Two computer programs, called Raters2 and Raters3, that implement these ideas are available at http://www.ru.nl/socialewetenschappen/rtog /software/statistische/kunst/.

Table 1 shows the two-way frequency distribution and the corresponding proportions, Cohen (1960, p. 45) used as an illustration. The proportion of joint judgments is the sum of the diagonal cells, here called f. In this example f = 0.70. Cohen defined chance agreement as $\sum_{i=1}^{c} M_{1i}M_{2i}$. In the example the correction is 0.30 + 0.09 + 0.02 = 0.41, so the corrected proportion of joint judgments is:

$$f - \sum_{i=1}^{c} M_{1i}M_{2i} = 0.29.$$

If this value is rescaled by dividing it by its maximum, Cohen's Kappa results:

$$\text{Kappa} = \frac{f - \sum_{i=1}^{c} M_{1i}M_{2i}}{1 - \sum_{i=1}^{c} M_{1i}M_{2i}} = 0.4915 \tag{1.2}$$

Table 1: Cohen's Example Data

Frequencies				Proportions			
88	14	18	120	0.44	0.07	0.09	0.60
10	40	10	60	0.05	0.20	0.05	0.30
2	6	12	20	0.01	0.03	0.06	0.10
100	60	40	200	0.50	0.30	0.20	1.00

The General Model in Detail

From the model parameters, the population distribution X of the simultaneous classifications can be derived. Any cell X(i,j) of X defines the probability of a joint classification in category i by rater 1 and category j by rater 2. X can be estimated from the two-way frequency matrix of the ratings in the sample, which will be indicated as $\hat{X}$. X can be interpreted as a weighted sum of four c by c matrices, corresponding to the behavior of the raters:

X_1: Both raters perform a correct observation. The probability of a score in a diagonal cell X_{1ii} is the product of: (a) the probability V_i that the case belongs to category i, (b) the probability p_1 that rater 1 performs a correct observation and (c) the probability p_2 that rater 2 performs a correct observation. Thus, $X_{1ii}=p_1p_2V_i$. The probability of a score in an off-diagonal cell is zero, so X_1 is a diagonal matrix.

X_2: Only rater 1 performs a correct observation. The probability of a score in a cell X_{2ij} is the product of: (a) the probability V_i that the case belongs to category i, (b) the probability p_1 that rater 1 performs a correct observation, (c) the probability q_2 that rater 2 guesses and (d) the probability W_{2j} that rater 2 guesses category j. Thus, $X_{2ij} = p_1V_iq_2W_{2j}$.

X_3: Only rater 2 performs a correct observation. The probability of a score in a cell X_{3ij} is the product of: (a) the probability V_j that the case belongs to category j, (b) the probability p_2 that rater 2 performs a correct observation, (c) the probability q_1 that rater 1 guesses and (d) the probability W_{1i} that rater 1 guesses category i. Thus, $X_{3ij} = p_2V_jq_1W_{1i}$.

X_4: Both raters are guessing. The probability of a score in a cell X_{4ij} is the product of: (a) the probability q_1 that rater 1 is guessing, (b) the probability q_2 that rater 2 is guessing, (c) the probability W_{1i} that rater 1 guesses category i and (d) the probability W_{2j} that rater 2 guesses category j. Thus, $X_{4ij} = q_1q_2W_{1i}W_{2j}$.

The matrix X is the sum of these 4 matrices and its content can be summarized as follows:

For $i \neq j$:

$$\begin{aligned} X_{ij} &= p_1q_2V_iW_{2j}+q_1p_2W_{1i}V_j+q_1q_2W_{1i}W_{2j} \\ &= (1-p_1p_2)V_iV_j+q_1V_jD_{1i}+q_2V_iD_{2j}+q_1q_2D_{1i}D_{2j}, \end{aligned} \tag{2}$$

and, for i = j:

$$X_{ii} = p_1p_2V_i+p_1q_2V_iW_{2i}+q_1p_2V_iW_{1i}+q_1q_2W_{1i}W_{2i} = p_1p_2V_i+(1-p_1p_2)V_i^2+q_1V_iD_{1i}+q_2V_iD_{2i}+q_1q_2D_{1i}D_{2i}. \quad (3)$$

The marginal distributions M_1 and M_2 of X are given by:

$$M_r = p_r.V + q_rW_r = V + q_rD_r, \text{ for } r = 1, 2. \quad (4)$$

A similar model is given by Klauer and Batchelder (1996).

Comparing s to Cohen's Kappa

The measure s and Cohen's Kappa can be compared based on the following derivation: from (3) it is evident that

$$X_{ii} = sV_i(1–V_i)+V_i^2+q_1V_iD_{1i}+q_2V_iD_{1i}+ q_1q_2D_{1i}D_{2i}$$

and, from (4),

$$M_{1i}M_{2i} = V_i^2+q_1V_iD_{1i}+q_2V_iD_{2i}+q_1q_2D_{1i}D_{2i}$$

thus,

$$X_{ii}–M_{1i}M_{2i} = sV_i(1-V_i), \quad (5)$$

and,

$$\text{s} = \frac{X_{ii}-M_{1i}M_{2i}}{V_i\left(1-V_i\right)} = \frac{f-\sum_{i=1}^{c}M_{1i}M_{2i}}{\sum_{i=1}^{c}V_i\left(1-V_i\right)} = \frac{f-\sum_{i=1}^{c}M_{1i}M_{2i}}{1-\sum_{i=1}^{c}V_i^2}.$$

Comparing this result with the formula for Kappa in (1.2) it follows that Kappa and s are only equivalent if $\sum_{i=1}^{c} M_{1i}M_{2i} = \sum_{i=1}^{c} V_i^2$. From (4) it becomes clear that such is the case only if $p_1 = p_2 = 1$, or if $W_1 = W_2 = V$. The $p_1 = p_2 = 1$ assumption is very unrealistic. The assumption that both W-vectors equal the true distribution implies that the two marginal distributions M_1 and M_2 are equal. This is a severe and unnecessary restriction that Cohen rejected when he introduced Kappa. In his example, as shown in Table 1, the two marginal distributions differ significantly ($\chi^2 = 34.6959$, df = 3, p = 0.0000). Table 2 shows Kappa as well as the results of an analysis of Cohen's example according to the model presented herein.

Table 2: Parameter Estimates According to Proposed Model for Cohen's Example

V	W_1	W_2	Parameter Estimates
0.6861	0.0000	0.0000	s = 0.6280
0.2347	0.7620	0.4683	p_1 = 0.8696
0.0792	0.2380	0.5317	p_2 = 0.7221
			kappa = 0.4915

model fit: χ^2 = 2.0325, df = 1, p = 0.1540

Identifiability of Model Parameters

By the identifiability of the parameters is meant that their values can be uniquely derived from the joint distribution matrix X. If $B_i = X_{ii} – M_{1i}M_{2i}$, then from (5):

$$B_i = sV_i(1 – V_i) \quad (6)$$

With at least 3 non-zero entries in V, the largest entry is the one closest to 0.5. Therefore, it corresponds to the largest entry in B. In other words: if B_m is (one of) the largest entry(s) in B, V_m is (one of) the largest entry(s) in V. From (6):

$$\frac{V_j\left(1-V_j\right)}{V_m\left(1-V_m\right)} = \frac{B_j}{B_m}$$

and, as a consequence, for all j ≠ m,

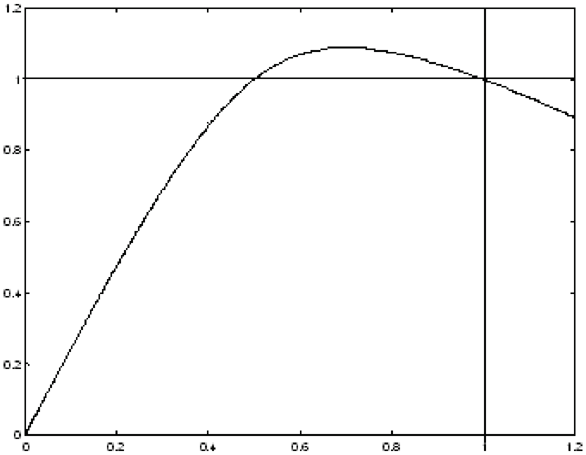
$$V_j = 0.5 \pm \sqrt{0.25 - V_m\left(1-V_m\right)\frac{B_j}{B_m}}$$

Because there can be only one entry in V greater than 0.5, the sign before the square root must be negative for all j ≠ m:

$$\mathrm{V_j} = 0.5 - \sqrt{0.25 - V_m\left(1 - V_m\right)\frac{B_j}{B_m}} \qquad (7)$$

It can be proved that there is only one value V_m for which the sum of elements in V according to (7) becomes 1, provided that: X obeys to the model, c > 2 (and consequently Vm < 1), s > 0, and, by definition, the sum of the elements in V equals 1. Figure 1 shows an example of the sum $g(V_m) = V_m + \sum_{j \neq m} V_j$ as a function of V_m and with V_j defined by (7).

Figure 1: Example of the Function $g(V_m)$



Thus, there is only one vector V for which equation (7) holds and for which the elements of V add up to 1. So V can be identified.

Once V has been identified, s can also be derived from (6):

$$\mathrm{s} = \frac{B_i}{V_i\left(1 - V_i\right)} \qquad (8)$$

for any i, except those for which $V_i = 0$.

Although the product p_1p_2 (i.e., s) can be identified, it is generally impossible to identify its components p_1 and p_2. From (4) it is known that $q_rD_r = M_r - V$, but looking at formulas (2) and (3) for the cells in X a multiplication of D_r by a constant h can be compensated by dividing q_r by the same h. Thus, neither W_1 and W_2, nor p_1 and p_2 can be identified.

The good news is that boundaries can be identified, within which these parameters are enclosed. The boundaries follow from the facts that: all cells of V, W_1 and W_2 represent probabilities and therefore must be in the range [0,1], and that V, W_1 and W_2 must add up to 1. Therefore, the following series of restrictions can be derived:

1. $s \leq p_1 \leq 1$ and $s \leq p_2 \leq 1$.

2. From (4) it is known that

$$s\frac{V_i}{M_{2i}} = p_1\frac{p_2V_i}{p_2V_i + q_2W_{2i}}, \text{ thus,}$$

$$s\frac{V_i}{M_{2i}} \leq p_1 \text{ and } p_2 \leq \frac{M_{2i}}{V_i}.$$

3. Similarly, it is known that:

$$s\frac{V_i}{M_{1i}} \leq p_2 \text{ and } p_1 \leq \frac{M_{1i}}{V_i}$$

4. Since all values of W_1 are between 0 and 1, it is known that:
$q_1(1-W_{1i}) \geq 0$
$\Rightarrow q_1(1-V_i) - q_1(W_{1i} - V_i) \geq 0$
$\Rightarrow p_1(1-V_i) + q_1(1-V_i) - q_1(W_{1i} - V_i) \geq p_1(1-V_i)$
$\Rightarrow 1-V_i - q_1(W_{1i} - V_i) \geq p_1(1-V_i)$
$$\Rightarrow \frac{1 - V_i - q_1\left(W_{1i} - V_i\right)}{1 - V_i} \geq p_1$$
$$\Rightarrow \frac{1 - M_{1i}}{1 - V_i} \geq p_1, \text{ and consequently:}$$
$$s\frac{1 - V_i}{1 - M_{1i}} \leq p_1.$$

5. In the same way the following may be derived:

$$\frac{1 - M_{1i}}{1 - V_i} \geq p_1 \text{ and } s\frac{1 - V_i}{1 - M_{1i}} \leq p_2.$$

These restrictions can be summarized by the following boundaries for all i and k:

$$p_1.p_2 \leq p_1 \leq 1 \qquad (9)$$

$$p_1.p_2 \leq p_2 \leq 1 \qquad (10)$$

If $M_{2i} < 1$ (and $V_i < 1$), then:

$$s\frac{V_i}{M_{2i}} \leq p_1 \leq \frac{M_{1k}}{V_k} \qquad (11)$$

If $M_{1i} < 1$ (and $V_i < 1$), then:

$$s\frac{V_k}{M_{1k}} \leq p_2 \leq \frac{M_{2i}}{V_i} \qquad (12)$$

If $M_{2i} < 1$ (and $V_i < 1$), then:

$$s\frac{1-V_i}{1-M_{2i}} \leq p_1 \leq \frac{1-M_{1k}}{1-V_k} \qquad (13)$$

If $M_{1i} < 1$ (and $V_i < 1$), then:

$$s\frac{1-V_k}{1-M_{1k}} \leq p_2 \leq \frac{1-M_{2i}}{1-V_i} \qquad (14)$$

These formulas are cross-linked: the minimum for p_1 in (11) goes together with the maximum p_2 in (12) and the maximum for p_1 in (11) corresponds to the minimum for p_2 in (12). The link comes from the fact that their product must be s. The formulas in (13) and (14) are connected in a similar way.

The limits from (11) through (14) all hinge upon the differences between the true distribution V and the rater error distributions W_1 and W_2. If $W_2 = V$ the lower limit for p_1 is s and the maximum for p_2 is 1. If $W_1 = V$ the lower limit for p_2 is s and the maximum for p_1 is 1. From these formulas it is also observed that, for categories with V-values close to 0 or 1, even small differences between W_1 or W_2 and V will impose strong restrictions.

It must be noted that this model cannot be applied if the number of categories is only 2.

Reparametrization

The parameters, as defined to this point, are neither all identifiable, nor are they independent. W_r and p_r cannot be identified, only the combination $q_rD_r = (1-p_r)(W_r-V)$. Therefore, a reparametrization from the original set of parameters to the set [s, V, q_1D_1 and q_2D_2] is used in the estimations. Moreover, the vector V adds up to 1 and the vectors q_1D_1 and q_2D_2 add up to 0, which means that their elements cannot be estimated independent. Therefore, following Klauer and Batchelder (1996) the model is reparametrized again as follows:

$$A_i^* = \frac{A_i}{1.1 - \sum_{j=1}^{i-1} A_j} \text{ for i = 1, c,}$$

where A = V, q_1D_1 and q_2D_2 respectively. The last element A_c is dropped. The back-translation to the original parameters is performed by the formula:

$$A_i = A_i^*\left(1.1 - \sum_{j=1}^{i-1} A_j\right), \text{for i = 1, c}$$

Initial Parameter Estimations: V and s

For the parameter estimations from an observed matrix $\hat{X}$ one may proceed in two steps. The first step is a procedure directly derived from the model and uses only information from the diagonal and the marginal frequencies of the observed matrix. In the second step a general minimization algorithm is applied to minimize a criterion (for instance, the negative of the likelihood) based on all cells of $\hat{X}$. This algorithm starts from the estimations produced by the first step.

For the first step define:

$$g(x) = x + \sum_{j \neq m}^{c}\left(0.5 - \sqrt{0.25 - x(1-x)\frac{\hat{B}_j}{\hat{B}_m}}\right)$$

with $\hat{B}_i = \hat{X}_{ii} - \hat{M}_{1i}\hat{M}_{2i}$ and $\hat{B}_m$ = the largest value in $\hat{B}$. (Figure 1 shows an example of this function.) From (7) it is clear that V_m can be estimated by the value of x for which g(x) = 1 with 0 < x < 1. Starting with evaluations of g at 1/c and a suitable maximum (for instance f), an estimate of V_m can be found by a simple iteration process using, for example, the bisection method. The remaining elements of V can be estimated by:

$$\hat{V}_j = 0.5 - \sqrt{0.25 - \hat{V}_m\left(1 - \hat{V}_m\right)\frac{\hat{B}_j}{\hat{B}_m}} \quad (15)$$

Once estimates of V are obtained, s can be estimated on the base of (8) as:

$$\hat{s} = \frac{\hat{B}_i}{\hat{V}_i\left(1 - \hat{V}_i\right)} \text{ for any i (unless } \hat{V}_i = 0\text{),}$$

or for a combination of the estimates for different i.

However, with sampled data this direct method may easily fail. Therefore a numerically more robust algorithm to find the same initial estimates of V and s was designed, called the *ping-pong algorithm*, a detailed description of which is provided later.

Initial Estimation of W_1 and W_2

Although the parameters W_1, W_2, p_1 and p_2 are not identifiable, the inequalities (9) through (14) offer the possibility to set boundaries around them. These boundaries may define a very narrow area, especially for categories that are very frequent or very rare. But unfortunately it is the infrequent categories for which the analysis produces the least reliable estimations. The problem becomes most serious if there are many categories and relatively few observations, i.e., if n/c is small. If the limits given by (9) through (14) restrict the estimates $\hat{p}_1$ and $\hat{p}_2$ to single values, as occurred when Cohen's example was analyzed, W_1 and W_2 can also be estimated using (4) as:

$$\hat{W}_r = \hat{V} + \frac{1}{\hat{q}_r}\left(M_r - \hat{V}\right)$$

Final Estimations and Model Test

The initial parameter estimates based on the considerations above are based completely on the diagonal and marginal distributions of $\hat{X}$ disregarding any information in the off-diagonal cells. In the final estimation procedure information from all cells will be used. A criterion is defined for the dissimilarity between the reconstruction X^* of X from the parameter estimates and the observed matrix $\hat{X}$, and a powerful minimization technique like the Davidon-Fletcher-Powell algorithm is used to improve the initial parameter estimates. An attractive criterion is based on the negative of the likelihood ratio with a small adjustment, defined as:

$$e_{ij} = \mathrm{Max}(X^*_{ij}, \varepsilon)$$

$$\mathrm{Crit} = \sum_{i=1}^{c}\sum_{j=1}^{c}\hat{X}_{ij}.\mathrm{LN}\left(\frac{\hat{X}_{ij}}{e_{ij}}\right) + \text{penalty}$$

The term ε is a small value to prevent division by zero and to avoid too exotic values of $\frac{\hat{X}_{ij}}{e_{ij}}$, for instance ε = 1.0e-20. The *penalty* serves to force the parameters within the restrictions of the model (for instance $0 \le s \le 1$).

The estimation procedure as designed starts with the ping-pong algorithm resulting in estimates $\hat{V}$ and $\hat{s}$, after which the reparametrizations and the minimization procedure are applied. When the final parameter estimates are obtained, a model test can be performed based on the test statistic for the likelihood ratio:

$$\chi^2 = 2.n.\left(\sum_{i=1}^{c}\sum_{j=1}^{c}\hat{X}_{ij}.\mathrm{LN}\left(\frac{\hat{X}_{ij}}{e_{ij}}\right)\right)$$

The associated number of degrees of freedom is c^2-3.c + 1.

The whole model as described above is based on the assumption that s is greater than zero. If $p_1 = 0$ or $p_2 = 0$, the value of any cell X_{ij} is equal to the product of the corresponding marginal probabilities M_{1i} and M_{2j}, even if X_{ij} is a diagonal cell. This assumption that s > 0 may be tested by the statistic $t = f - \sum_{i=1}^{c} M_{1i}.M_{2i}$, which is (approximately) distributed as Student's t with 1 degree of freedom. Confidence intervals for the parameters may be constructed by the use of the information matrix or, if the Hessian matrix is singular, by bootstrapping methods.

The Ping-Pong Algorithm

The ping-pong algorithm is designed to simultaneously estimate $s = p_1.p_2$, and the largest element V_m in V. Once V_m is estimated, the entire vector V can be estimated according to (15). In order to grasp the basic idea of the algorithm, assume that the exact values for B and f are known. Then the logic is as follows. Define: t_i = upper boundary for s in the i^{th} iteration, and u_i = lower boundary for V_m in the i^{th} iteration.

1. From (3) $s \le f = \sum_{i=1}^{c} X_{ii}$, so choose $t_0 = f$.
2. From (7):

   $1 = \sum_{i=1}^{c} V_i =$

   $$\frac{1}{2}(c-1) + V_m - \sum_{j\neq m} \sqrt{0.25 - V_m(1-V_m)\frac{B_j}{B_m}}$$

   so, using (8)

   $$V_m = 1 - \frac{1}{2}(c-1) + \sum_{j\neq m}\sqrt{0.25 - \frac{B_j}{s}}$$

   and as a consequence:

   $V_m \le 1 - \frac{1}{2}(c-1) + \sum_{j\neq m}\sqrt{0.25 - \frac{B_j}{t_i}}$ for any step i
3. From (8): $s = \frac{B_m}{V_m(1-V_m)}$, thus

   $$s \le \frac{B_m}{u_i(1-u_i)}$$

Now the following procedure is applied:

1) $t_0 = f$

2) $u_i = 1 - \frac{1}{2}(c-1) + \sum_{j\neq m}\sqrt{0.25 - \frac{B_j}{t_i}}$

3) $t_i = \frac{B_m}{u_{i-1}(1-u_{i-1})}$

4) Repeat from 2) until convergence is reached.

This algorithm converges to $t_i = \hat{s}$ and $u_i = \hat{V}_m$.

When working with the sample estimators $\hat{B}$ and $\hat{f}$ it may be necessary to make some corrections during the iteration process:

1. In the iteration process t_i may exceed the value $t_0 = \hat{f}$. In that case, force $\hat{B}_m$ to $\hat{f}.u_{i-1}(1-u_{i-1})$ and set t_i equal to $\hat{f}$. In order to keep the sum of $\hat{B}$ unchanged, replace the other elements of $\hat{B}$ according to the following rule:

   $$\hat{B}_i \leftarrow \overline{B} + \left(\hat{B}_i - \overline{B}\right)\frac{\hat{B}_m^* - \overline{B}}{\hat{B}_m - \overline{B}}$$

   where $\overline{B}$ is the mean of the $\hat{B}$-values, $\hat{B}_m$ the original estimate of B_m and $\hat{B}_m^*$ the corrected estimate.

2. If the estimate u_i becomes less than 1/c force it back to 1/c and adjust the B-values accordingly:

   $$u_i < \frac{1}{c},$$

   so

   $$1 - \frac{1}{2}(c-1) + \sum_{j\neq m}\sqrt{0.25 - \frac{B_j}{t_i}} < \frac{1}{c}$$

Adjust the B-vector by a vector B^*, such that

$$1 - \frac{1}{2}(c-1) + \sum_{j\neq m}\sqrt{0.25 - \frac{B_j^*}{t_i}} = \frac{1}{c},$$

which means that

$$\sum_{j\neq m}\sqrt{0.25 - \frac{B_j^*}{t_i}} = 0.5c + \frac{1}{c} - 1.5$$

Make the adjustment by taking B^* such that each term in the summation, except B_m, is multiplied by:

$$a = \frac{0.5c + \frac{1}{c} - 1.5}{\sum_{j\neq m}\sqrt{0.25 - \frac{B_j}{t_i}}} = \frac{0.5c - 1.5 + \frac{1}{c}}{0.5c - 1.5 + u_i}$$

This is realized by replacing each B_j, except B_m, by $B^*_j = 0.25t_i(1\text{-}a^2) + a^2B_j$.

Three Raters

Under the given model, the expansion to three simultaneous raters is straightforward. Moreover, with three simultaneous raters, all parameters are identifiable if there are at least three categories. The notation must be extended to three p-values p_1, p_2 and p_3, three q-values q_1, q_2 and q_3, three W-vectors W_1, W_2 and W_3, and three marginal distributions M_1, M_2 and M_3. In addition the matrix X will now have three dimensions. The formulas for the probabilities in the cells of X are more complicated: X_{ijk} is the sum of the corresponding cells in eight submatrices as shown in Tables 3a through 3c.

Table 3a: Formulas for Two Parts of the Matrix X in Case of Three Raters

Raters i, j, k	123	i = j = k	i = k ≠ j
		Xijk	Xijk
X_1	ccc	p1p2p3Vi,	0
X_2	cci	p1p2Viq3W3k	0
X_3	cic	p1p3Viq2W2j	p1p3Viq2W2j
X_4	cii	p1Viq2W2jq3W3k	p1Viq2W2jq3W3k
X_5	icc	p2p3Vjq1W1i	0
X_6	ici	p2Vjq1W1iq3W3k	p2Vjq1W1iq3W3k
X_7	iic	p3Vkq1W1iq2W2j	p3Vkq1W1iq2W2j
X_8	iii	q1W1iq2W2jq3W3k	q1W1iq2W2jq3W3k

Table 3b: Formulas for Two Parts of the Matrix X in Case of Three Raters

Raters i, j, k	123	i = j ≠ k	i ≠ j = k
		Xijk	Xijk
X_1	ccc	0	0
X_2	cci	p1.p2Viq3W3k	0
X_3	cic	0	0
X_4	cii	p1Viq2W2jq3W3k	p1Viq2W2jq3W3k
X_5	icc	0	p2p3Vjq1W1i
X_6	ici	p2Vjq1W1iq3W3k	p2Vjq1W1iq3W3k
X_7	iic	p3Vkq1W1iq2W2j	p3Vkq1W1iq2W2j
X_8	iii	q1W1iq2W2jq3W3k	q1W1iq2W2jq3W3k

Table 3c: Formulas for One Part of the Matrix X in Case of Three Raters

Raters i, j, k	123	i ≠ j ≠ k
		Xijk
X_1	ccc	0
X_2	cci	0
X_3	cic	0
X_4	cii	p1Viq2W2jq3W3k
X_5	icc	0
X_6	ici	p2Vjq1W1iq3W3k
X_7	iic	p3Vkq1W1iq2W2j
X_8	iii	q1W1iq2W2jq3W3k

Submatrix X_1 contains those ratings for which all three raters make a correct observation, as indicated by the code ccc, which means correct-correct-correct. The value in cell i, j, k depends on the equality of the three indices as indicated by the column headings. The other submatrices are organized in the same way: X_2 contains ratings where raters 1 and 2 made correct observations, but rater three did not (he guessed, correctly or not), indicated by the label cci (correct-correct-incorrect).

Table 4: Frequency Matrix with Three Categories and Three Raters

	Rater 3 = 1			Rater 3 = 2			Rater 3 = 3		
	Rater 2			Rater 2			Rater 2		
Rater 1	37	16	19	32	21	13	0	2	7
	19	11	7	30	103	38	9	11	16
	5	7	2	10	22	11	11	13	28

Table 4 shows an example of the three-way distribution in a sample with size 500. The data were generated by random sampling from a theoretical distribution based on the probabilities given in tables 3a-3c, with the following parameters: $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.6$, $V^T = (0.3,0.5,0.2)$, $W_1^T = (0.2,0.5,0.3)$, $W_2^T = (0.3,0.4,0.3)$, and $W_3^T = (0.1,0.7,0.2)$. Initial estimations for p_1, p_2, p_3, V, W_1, W_2 and W_3 can be derived from the three marginal planes, which can be computed from $\hat{X}$ by summation over the categories of one rater:

$$X^{12}_{ij} = \sum_{k=1}^{c} X_{ijk}$$

$$X^{13}_{ij} = \sum_{k=1}^{c} X_{ikj}$$

$$X^{23}_{ij} = \sum_{k=1}^{c} X_{kij}$$

Tables 5, 6 and 7 show these planes for the example in table 4.

Table 5: The Marginal Planes for Raters 1 and 2 in the Example

$\hat{X}^{12}$:	Rater 2			$\hat{M}_1$
Rater 1	0.138	0.078	0.078	0.294
	0.116	0.250	0.122	0.488
	0.052	0.084	0.082	0.218
$\hat{M}_2$	0.306	0.412	0.282	1.000

Table 6: The Marginal Planes for Raters 1 and 3 in the Example

$\hat{X}^{13}$:	Rater 3			$\hat{M}_1$
Rater 1	0.144	0.132	0.018	0.294
	0.074	0.342	0.072	0.488
	0.028	0.086	0.104	0.218
$\hat{M}_3$	0.246	0.560	0.194	1.000

Table 7: The Marginal Planes for Raters 2 and 3 in the Example

$\hat{X}^{23}$:	Rater 3			$\hat{M}_2$
	0.122	0.144	0.040	0.306
Rater 2	0.068	0.292	0.052	0.412
	0.056	0.124	0.102	0.282
$\hat{M}_3$	0.246	0.560	0.194	1.000

Define:

$$B^{12}_i = X^{12}_{ii} - M_{1i}.M_{2i}$$
$$B^{13}_i = X^{13}_{ii} - M_{1i}.M_{3i}$$
$$B^{23}_i = X^{23}_{ii} - M_{2i}.M_{3i}$$

In the example above these values are estimated by:

$$\hat{B}^{12T} = [0.038036, 0.048944, 0.020524]$$
$$\hat{B}^{13T} = [0.071676, 0.068720, 0.061708]$$
$$\hat{B}^{23T} = [0.046724, 0.061280, 0.047292]$$

Because $B^{12}_i = p_1p_2V_i(1 - V_i)$ and analogously $B^{13}_i = p_1p_3V_i(1 - V_i)$ and $B^{23}_i = p_2p_3V_i(1 - V_i)$:

$$\frac{V_j(1-V_j)}{V_i(1-V_i)} = \frac{B_j^{12}}{B_i^{12}} = \frac{B_j^{13}}{B_i^{13}} = \frac{B_j^{23}}{B_i^{23}} = \frac{B_j^{12}+B_j^{13}+B_j^{23}}{B_i^{12}+B_i^{13}+B_i^{23}} = \frac{1}{3}\left(\frac{B_j^{12}}{B_i^{12}}+\frac{B_j^{13}}{B_i^{13}}+\frac{B_j^{23}}{B_i^{23}}\right) \tag{16}$$

The largest value V_m in V can be estimated by setting it to the value of x for which the function $g(x) = 1$, where g is defined as:

$$g(x) = x + \sum_{j \neq m}^{c}\left(0.5 - \sqrt{0.25 - x(1-x)\left(\frac{\hat{B}_j^{12}+\hat{B}_j^{13}+\hat{B}_j^{23}}{\hat{B}_m^{12}+\hat{B}_m^{13}+\hat{B}_m^{23}}\right)}\right)$$

or as

$$g(x) = x + \sum_{j \neq m}^{c}\left(0.5 - \sqrt{0.25 - x(1-x)\frac{1}{3}\left(\frac{\hat{B}_j^{12}}{\hat{B}_m^{12}}+\frac{\hat{B}_j^{13}}{\hat{B}_m^{13}}+\frac{\hat{B}_j^{23}}{\hat{B}_m^{23}}\right)}\right)$$

In this function the index m refers to the largest value (or one of the largest values) in the B-vectors. Because the three estimated B-vectors in a sample may have different orders, choose m as the index for which

$$\hat{B}_m^{12} + \hat{B}_m^{13} + \hat{B}_m^{23} \geq \hat{B}_i^{12} + \hat{B}_i^{13} + \hat{B}_i^{23}$$

for all i. In the example, from (15), $\hat{V}_m$ = 0.422659.

From (14) follows that, for the other elements of V,

$$V_j = 0.5 - \sqrt{0.25 - V_m(1 - V_m)\left(\frac{\hat{B}_j^{12} + \hat{B}_j^{13} + \hat{B}_j^{23}}{\hat{B}_m^{12} + \hat{B}_m^{13} + \hat{B}_m^{23}}\right)}$$

and it is found that:

$$\hat{V}^T = [0.348216, 0.422659, 0.229124]$$

Once the initial estimate of V is made, the parameters p_1, p_2 and p_3 can be estimated in the following way: from (8) it is known that, for all i,

$$s^{12} = p_1 p_2 = \frac{B_i^{12}}{V_i(1 - V_i)},$$

so the product can be estimated by averaging over i-values:

$$\hat{s}^{12} = \frac{1}{c}\sum_{i=1}^{c} \frac{\hat{B}_i^{12}}{\hat{V}_i\left(1 - \hat{V}_i\right)}.$$

In the same way s^{13} and s^{23} can be estimated and estimates of the parameters p_1, p_2 and p_3 can be found by combining the three estimated s-values. For any triple (i, j, k) raters:

$$\frac{s^{ij}s^{ik}}{s^{jk}} = \frac{p_i p_j p_i p_k}{p_j p_k} = p_i^2,$$

so the p-values can be estimated from their estimated products:

$$\hat{p}_i = \sqrt{\frac{\hat{s}^{ij}\hat{s}^{ik}}{\hat{s}^{jk}}}.$$

In the example: $\hat{s}^{12}$ = 0.176141, $\hat{s}^{13}$ = 0.315598, $\hat{s}^{23}$ = 0.241583 and $\hat{p}_1$ = 0.479694, $\hat{p}_2$ = 0.367195, $\hat{p}_3$ = 0.657915. Once initial estimates for V and the p-parameters are obtained, the estimation of the W-vectors is straightforward. From (4), it is known that, for rater r, $M_r = p_r V + (1-p_r)W_r$, so W_r can be estimated by:

$$W_r = \frac{1}{1 - \hat{p}_r}\left(M_r - \hat{p}_r \hat{V}\right).$$

In the example this results in the following initial estimates:

$$\hat{W}_1^T = [0.244016, 0.548241, 0.207744],$$
$$\hat{W}_2^T = [0.281504, 0.405815, 0.312682],$$
$$\hat{W}_3^T = [0.049413, 0.824141, 0.126448],$$

but with sample data these formulas may lead to negative entries in the estimated W-vectors. If that occurs the initial estimate for the W-vector at hand can be set equal to the estimated V.

Final estimates, using information from all cells in $\hat{X}$, can be computed by methods analogous to those described, minimizing the adjusted likelihood ratio.

Conclusion

When Cohen (1960) introduced his measure Kappa, he provided a good index to estimate inter-rater agreement in the case of a nominal category system that could be easily computed by hand. Cohen argued that differences in the marginal distributions must be taken into account, but, as shown, his measure Kappa does so correctly only if the marginal distributions are equal. For practical reasons, especially the fact that computers were mostly unavailable in 1960, Kappa could be considered the best available instrument at the time, but with modern computers advancements can be made. A model based on Cohen's ideas and a procedure to correctly estimate its parameters was presented herein. The model allows - to a certain extent - to separately estimate the qualities of two raters by giving two measures p_1 and p_2. It also breaks apart the rater characteristics (W_1 and W_2) on one hand and the true distribution of the categories (V) on the other.

If the estimates p_r and W_r are truly independent from the distribution V, it becomes possible first to assess these statistics for one rater (using a second rater) in a pilot study, and then to use them in order to find boundaries for the V-values in the main study without the need for a second rater. The formula to be used follows from (4): $\hat{V}_i = \dfrac{M_{ri} - \hat{q}_r \hat{W}_{ri}}{\hat{p}_r}$.

References

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

Brennan, R. L., & Dale J. Prediger, D. J. (1981). Coefficient kappa: some uses, misuses and alternatives, *Educational and Psychological Measurement*, *41*, 687-699.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46

Klauer, K.C., & Batchelder, W.H. (1996). Structural analysis of subjective categorical data. *Psychometrika*, *61*, 199-240.

Maxwell, A.E. (1977), Coefficients of agreement between Observers and their interpretation. *British Journal of Psychiatry*, *130*, 79-83.

Perreault, W.D., & Leigh, E. (1989). Reliability of nominal data based on qualitative judgements. *Journal of Marketing Research*, *26*, 135-148.

Scott, W.A. (1955), Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374 378.