Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2015, Article ID 751803, 8 pages
http://dx.doi.org/10.1155/2015/751803

Research Article

Measurement of Interobserver Disagreement: Correction of Cohen's Kappa for Negative Values

Tarald O. Kvålseth

Departments of Mechanical Engineering and Industrial & Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Correspondence should be addressed to Tarald O. Kvålseth; kvals001@umn.edu

Received 24 June 2015; Accepted 6 September 2015

Academic Editor: Z. D. Bai

As measures of interobserver agreement for both nominal and ordinal categories, Cohen's kappa coefficients appear to be the most widely used with simple and meaningful interpretations. However, for negative coefficient values when (the probability of) observed disagreement exceeds chance-expected disagreement, no fixed lower bounds exist for the kappa coefficients and their interpretations are no longer meaningful and may be entirely misleading. In this paper, alternative measures of disagreement (or negative agreement) are proposed as simple corrections or modifications of Cohen's kappa coefficients. The new coefficients have a fixed lower bound of −1 that can be attained irrespective of the marginal distributions. A coefficient is formulated for the case when the classification categories are nominal and a weighted coefficient is proposed for ordinal categories. Besides coefficients for the overall disagreement across categories, disagreement coefficients for individual categories are presented. Statistical inference procedures are developed and numerical examples are provided.

1. Introduction

When two (or more) observers are independently classifying observations or items (objects) into the same set of k mutually exclusive and exhaustive categories, it may be of interest to have a summary description of the extent to which the observers agreed in their classifications. The total probability (proportion) of agreement is one such obvious summary measure. However, since some agreement is to be expected purely by chance, Cohen [1] introduced the *kappa coefficient of agreement* as one that corrects for the chance-expected agreement. Cohen's kappa has since become widely used in a variety of situations and discussed extensively in various textbooks (e.g., [2–5]) and a wide variety of journal publications (e.g., [6–10]).

In order to define the kappa coefficient in terms of probabilities (proportions), let p_{ij} be the probability that a random observation is assigned to category i by Observer 1 and to category j by Observer 2 for $i = 1, \ldots, k$ and $j = 1, \ldots, k$. Furthermore, let p_{i+} denote the probability that a randomly chosen observation is assigned to category i by Observer 1 and p_{+j} the probability that a randomly chosen observation is assigned to category j by Observer 2 ($i, j = 1, \ldots, k$). If these probabilities are represented in terms of a two-way contingency table with rows $i = 1, \ldots, k$ and columns $j = 1, \ldots, k$, then p_{ij} becomes the probability in cell (i, j) and $\{p_{i+}\}$ becomes the marginal row distribution and $\{p_{+j}\}$ becomes the marginal column distribution. With the row categories and the column categories being the same, $\sum_{i=1}^{k} p_{ii}$ is the total probability of agreement between the two observers. Cohen [1] used the overall statistical independence as the condition for chance agreement and defined K as

$$K = \frac{P_{\mathrm{AO}} - P_{\mathrm{AC}}}{1 - P_{\mathrm{AC}}}, \quad P_{\mathrm{AO}} = \sum_{i=1}^{k} p_{ii}, \; P_{\mathrm{AC}} = \sum_{i=1}^{k} p_{i+} p_{+i} \quad (1)$$

with P_{AO} and P_{AC} being the observed agreement probability and the chance-expected agreement probability, respectively. In terms of the observed and chance-expected disagreement probabilities P_{DO} and P_{DC}, K can alternatively be expressed as

$$K = 1 - \frac{P_{\mathrm{DO}}}{P_{\mathrm{DC}}},$$
$$P_{\mathrm{DO}} = \sum_{i=1,i\neq j}^{k} \sum_{j=1,i\neq j}^{k} p_{ij}, \quad P_{\mathrm{DC}} = \sum_{i=1,i\neq j}^{k} \sum_{j=1,i\neq j}^{k} p_{i+}p_{+j}. \tag{2}$$

It is clear from (1)-(2) that $K = 1$ if the interobserver agreement is perfect, that is, if $P_{\mathrm{AO}} = 1$ ($P_{\mathrm{DO}} = 0$), $K = 0$ if $P_{\mathrm{AO}} = P_{\mathrm{AC}}$ ($P_{\mathrm{DO}} = P_{\mathrm{DC}}$), and $K < 0$ if $P_{\mathrm{AO}} < P_{\mathrm{AC}}$ ($P_{\mathrm{DO}} > P_{\mathrm{DC}}$). The case of negative K-values will be discussed further in the next section.

In addition to measuring the overall agreement between two observers, it may be of interest to assess their level of agreement for specific categories. Spitzer et al. [11] first proposed such a measure by collapsing the original $k \times k$ table into a 2×2 table, one such 2×2 table for each category $i = 1, \ldots, k$, and then computing K in (1)-(2) for each such 2×2 table (see also [2, Chapter 18]). As a simpler procedure yielding the same numerical results, Kvålseth [12] proposed the following form of kappa for the specific category i ($i = 1, \ldots, k$):

$$K_i = \frac{p_{ii} - p_{i+}p_{+i}}{\overline{p}_i - p_{i+}p_{+i}}, \quad \overline{p}_i = \frac{p_{i+} + p_{+i}}{2} \tag{3}$$

$$= 1 - \frac{\sum_{D_i} \sum_{D_i} p_{ij}}{\sum_{D_i} \sum_{D_i} p_{i+}p_{+j}}, \tag{4}$$

where $\sum_{D_i} \sum_{D_i}$ denotes the summation over all disagreement cells for category i. With, say, $k = 3$, D_2 consists of cells $(2,1)$, $(2,3)$, $(1,2)$, and $(3,2)$. For complete agreement with respect to category i, $K_i = 1$ when $p_{ii} = p_{i+} = p_{+i}$, $K_i = 0$ for the independence $p_{ii} = p_{i+}p_{+i}$, and $K_i < 0$ when observed disagreement exceeds chance disagreement.

To account for the potential fact that some disagreements may be more serious than others, as when the k categories have a natural order, Cohen [13] and Cicchetti and Allison [14] independently introduced the *weighted kappa* K_w, which can be expressed as

$$K_w = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} p_{ij} - \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} p_{i+} p_{+j}}{1 - \sum_{i=1}^{k} \sum_{j=1}^{k} v_{ij} p_{i+} p_{+j}} \tag{5}$$

$$= 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}}, \tag{6}$$

where each weight $w_{ij} \in [0,1]$, with $w_{ii} = 0$ and $v_{ij} = 1 - w_{ij}$ for all i and j and with the following logical weight choices (e.g., [2, page 609]):

$$w_{ij} = \frac{|i-j|}{k-1}$$
$$\text{or } w_{ij} = \left(\frac{i-j}{k-1}\right)^2; \tag{7}$$
$$i, j = 1, \ldots, k.$$

For a specific category i, Kvålseth [15] proposed the following measure as an extension of (4):

$$K_{wi} = 1 - \frac{\sum_{D_i} \sum_{D_i} w_{ij} p_{ij}}{\sum_{D_i} \sum_{D_i} w_{ij} p_{i+} p_{+j}} \tag{8}$$

with D_i denoting the set of all disagreement cells for category i. The values of these weighted measures equal 1 for perfect agreement and 0 if observed agreement equals chance agreement, with negative values if observed agreement is less than chance agreement.

The kappa coefficients in (1)–(8) may be appropriate measures of agreement when their values are nonnegative, but not when their values are negative as discussed in the next section. From a theoretical point of view at least, it is certainly troublesome that their negative values lack appropriate meaning and validity. This paper presents simple corrections or modifications of the kappa coefficients in (1)–(8) such that the negative values of the corrected coefficients provide appropriate representation of the extent to which the observers disagree. Statistical inference procedures for the new coefficients or measures are developed. Numerical examples are also given.

2. Comments on Kappa

One of the most appealing properties of kappa, and undoubtedly a reason for its popularity, is its simplicity and transparency. All the kappa coefficients in (1)–(8) have intuitively appealing and meaningful interpretations. In the case of K in (1)-(2), for example, it seems most meaningful to interpret any K-value in terms of (2) as the proportional difference between P_{DC} and P_{DO}, that is, the relative extent to which the observed disagreement probability P_{DO} is less than the disagreement probability P_{DC} attributable to chance. By comparison, the norming used in (1) is not unique, with any number of different potential denominators d such that $(P_{\mathrm{AO}} - P_{\mathrm{AC}})/d \leq 1$ [16].

Complete statistical independence, that is, $p_{ij} = p_{i+}p_{+j}$ for $i, j = 1, \ldots, k$, is a sufficient, but not a necessary, condition for the kappa coefficients in (1)–(8) to take on value 0. In fact, for $K = 0$ in (1) and $k > 2$, it is not necessary that $p_{ii} = p_{i+}p_{+i}$ for $i = 1, \ldots, k$. It is possible that $K = 0$ even if $p_{ij} \neq p_{i+}p_{+j}$ for all i and j when $k > 2$. As a simple example, consider

$$p_{11} = \frac{1}{6}, \quad p_{12} = \frac{1}{6}, \quad p_{13} = 0,$$
$$p_{21} = 0, \quad p_{22} = \frac{1}{12}, \quad p_{23} = \frac{3}{12}\left(= \frac{1}{4}\right), \tag{9}$$
$$p_{31} = \frac{1}{6}, \quad p_{32} = \frac{1}{12}, \quad p_{33} = \frac{1}{12},$$

where all marginal probabilities equal 1/3. And $p_{ij} \neq p_{i+}p_{+j}$ for all i and j, but $K = 0$. In this case, from (3), $K_1 = 1/4$, $K_2 = K_3 = -1/8$, and, from (6) and (7), $K_w = 0.06$ for $w_{ij} = |i-j|/(k-1)$.

Note that the two expressions for K in (1)-(2) are weighted arithmetic means of the expressions for K_i in (3)-(4). Thus, from (1) and (3), for instance, it is seen that

$$K = \sum_{i=1}^{k} u_i K_i, \quad u_i = \frac{\overline{p}_i - p_{i+}p_{+i}}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}}, \ i = 1, \ldots, k. \tag{10}$$

Similarly, for the weighted measures in (6) and (8),

$$K_w = \sum_{i=1}^{k} u_i K_{wi}, \qquad u_i = \frac{\sum_{D_i}\sum_{D_i} w_{ij} p_{i+}p_{+j}}{\sum_{i=1}^{k}\sum_{D_i}\sum_{D_i} w_{ij} p_{i+}p_{+j}}, \ i = 1, \ldots, k. \tag{11}$$

In order to show that the interobserver agreement for a specific category i can be determined directly from (3)-(4), without the need to collapse the original $k \times k$ table as suggested by Spitzer et al. [11], consider that the original $k \times k$ table with probability components p_{ii}, p_{i+}, and p_{+i} for category i is collapsed into the following 2×2 table:

$$p_{11}^{(2)} = p_{ii}, \qquad p_{12}^{(2)} = p_{i+} - p_{ii}, \qquad p_{21}^{(2)} = p_{+i} - p_{ii}, \quad p_{22}^{(2)} = 1 - p_{i+} - p_{+i} + p_{ii}. \tag{12}$$

When (12) is substituted into (1), K_i in (3) results immediately. However, no such corresponding procedure applies to K_w in (6) and K_{wi} in (8). Note that, for $k = 2$, $K_1 = K_2 = K$ and $K_{w1} = K_{w2} = K_w$.

In spite of its wide appeal, kappa is not without some criticism or controversy, especially related to its dependence on the marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$ (see, e.g., [4, pages 168–173]). The chance agreement (disagreement) for all the kappa coefficients in (1)–(8) is based on the marginal distributions. If those distributions are highly uneven (nonuniform) and nearly symmetric, the values of the kappa coefficients may become unreasonably small due to the relatively large chance agreements.

A clear limitation of the kappa coefficients relates to situations when the values of those coefficients become negative and lack meaningful interpretations. This limitation has generally been ignored in published studies, partly perhaps because such studies using kappa have typically involved positive kappa values. Negative kappa values could, however, lead to incorrect interpretations, results, and conclusions. Also, if, for instance, $K > 0$ in (1)-(2), it is possible that some $K_i < 0$ in (3)-(4).

For the overall kappa in (1)-(2), when $P_{\text{AO}} < P_{\text{AC}}$ so that $K < 0$, K has no reasonable meaning in terms of (1), but $-K$ does in terms of (2); that is, $-K$ is the relative extent to which P_{DO} exceeds P_{DC}. The same argument applies to K_i in (3)-(4). However, two serious limitations of all the kappa coefficients are that, for negative values, (a) the coefficients have no fixed lower bounds, making it impossible to appropriately assess the size or magnitude of coefficient values, and (b) the coefficients take on negative values that do not appear reasonable as discussed below.

The minimum values $-P_{\text{AC}}/(1 - P_{\text{AC}})$ of K in (1) and $-p_{i+}p_{+i}/(\overline{p}_i - p_{i+}p_{+i})$ of K_i in (3) depend exclusively on the marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$. Values such as $K = -0.4$ or $K_i = -0.2$ are uninformative since they cannot be related to any fixed lower bounds on K or K_i such as -1, irrespective of the marginal distributions. There is no basis for making any interpretation or statement such as $K = -0.5$ indicating a "moderate," "low," or "high" level of disagreement between the two observers.

There is also some confusion in the literature about the minimum value of K, with some stating that the minimum value is $-\infty$ or $-P_{\text{AO}}/(1 - P_{\text{AO}})$ [5, page 4] and others stating that it is -1 when $p_{i+} = p_{+j} = 1/k$ for all i and j [17, page 113]. Such statements are clearly incorrect. In fact, the minimum value $K = -P_{\text{AC}}/(1 - P_{\text{AC}})$ equals -1 if, and only if, $P_{\text{AC}} = 0.5$. Similarly, the minimum value of K_i in (3) equals -1 only when the harmonic mean $2p_{i+}p_{+i}/(p_{i+} + p_{+i})$ of p_{i+} and p_{+i} equals 0.5.

What is needed are chance-corrected measures of disagreement, both weighted and unweighted, which have fixed lower bounds of -1 and which are attainable irrespective of the marginal distributions. This requirement has also been clearly emphasized by others [18]. Such measures will be introduced in the next section as simple corrections or modifications of the existing kappa coefficients.

3. Proposed Kappa Coefficients of Disagreement

3.1. Overall Coefficients. When $P_{\text{AO}} < P_{\text{AC}}$ and hence $P_{\text{DO}} > P_{\text{DC}}$, it seems most logical and intuitive to define negative overall kappa as

$$K^- = -\left(\frac{P_{\text{DO}} - P_{\text{DC}}}{1 - P_{\text{DC}}}\right) = -\left(1 - \frac{P_{\text{AO}}}{P_{\text{AC}}}\right), \tag{13}$$

where P_{AO}, P_{AC}, P_{DO}, and P_{DC} are the probabilities defined in (1)-(2). Consequently,

$$K_{\text{corrected}} = \begin{cases} K \text{ in (1)-(2)}, & \text{if } P_{\text{AO}} \geq P_{\text{AC}}, \\ K^- \text{ in (13)}, & \text{if } P_{\text{AO}} \leq P_{\text{AC}}, \end{cases} \tag{14}$$

where, of course, $K = K^- = 0$ for $P_{\text{AO}} = P_{\text{AC}}$. Except for the minus sign, K^- in (13) follows from K in (1)-(2) by simply substituting disagreement probabilities for the corresponding agreement probabilities.

The properties of K^- can be summarized as follows:

(P1) K^- is well defined if at least two cells of the contingency table contain nonzero probabilities.

(P2) $K^- \in [0, -1]$, with $K^- = 0$, if observed agreement (disagreement) equals chance agreement (disagreement) (i.e., $P_{\text{AO}} = P_{\text{AC}}$ or $P_{\text{DO}} = P_{\text{DC}}$) and $K^- = -1$ if, and only if, $P_{\text{AO}} = 0$ ($P_{\text{DO}} = 1$).

(P3) K^- can take on value -1 for any marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$.

(P4) $|K^-|$ has a meaningful interpretation as the relative extent to which the observed agreement probability is less than that expected by chance alone.

Table 1: A 2×2 contingency table with marginal probabilities r and $1 - r$ and with the entries in each cell as follows: first entry corresponds to $P_{\text{AD}} = 0$, second entry corresponds to $P_{\text{AO}} = P_{\text{AC}}$, and third entry is the weighted mean of the other two entries where $0 \leq \lambda \leq 1$.

Observer 1	Observer 2		Total
	Category 1	Category 2	
Category 1	0 $r(1-r)$ $(1-\lambda)r(1-r)$	r r^2 $\lambda r + (1-\lambda)r^2$	r
Category 2	$1-r$ $(1-r)^2$ $\lambda(1-r) + (1-\lambda)(1-r)^2$	0 $r(1-r)$ $(1-\lambda)r(1-r)$	$1-r$
Total	$1-r$	r	1.00

(P5) K^- takes on values that appear reasonable throughout its 0 to −1 range.

While Properties (P1)–(P4) are immediately apparent from the definition in (13), Property (P5) needs an explanation. This can most simply be done for the $k = 2$ category case and without undue loss of generality since, for any data set with $k > 2$, there exists an equivalent 2×2 table with the same K^--value. Therefore, one may consider a 2×2 table such as the one in Table 1 with the marginal probabilities r and $1-r$ $(0 \leq r \leq 1)$. The first two entries in each cell correspond to the cases when $K^- = -1$ and 0, respectively, while the third entry equals the weighted arithmetic mean of the other two entries with weights λ and $1 - \lambda$ $(0 \leq \lambda \leq 1)$.

In order for the values of K^- to be considered reasonable throughout the $[-1, 0]$-interval, the only logical condition would clearly seem to be that the value of K^- for the weighted mean cell probabilities should equal the weighted mean value of K^- for the other cell probabilities with the same weights λ and $1 - \lambda$; that is,

$$K^-\left(\{\text{mean probabilities}\}\right) = \lambda\left(K^- = -1\right) + (1-\lambda)\left(K^- = 0\right) = -\lambda. \tag{15}$$

By substituting the expressions for the mean cell probabilities from Table 1 into K^- in (13), it is seen that K^- does meet the condition in (15), irrespective of the marginal probabilities r and $1 - r$. This assumes, of course, as with Cohen's K, that chance agreement (disagreement) based on the marginal probabilities is reasonable.

By contrast, substituting the mean probabilities from Table 1 into K in (1)-(2) gives

$$K\left(\{\text{mean probabilities}\}\right) = -\left(\frac{2r(1-r)}{1-2r(1-r)}\right)\lambda \tag{16}$$

showing the strong dependence of K on the marginal probabilities. The parenthetical term in (16) equals 1 if $r = 0.5$ and approaches 0 as the marginal distributions become highly uneven or nonuniform (i.e., as r approaches 0 or 1).

When $K_w < 0$ in (5)-(6) and hence $\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij} < \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}$ and $\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij} > \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}$ and with the sets of weights $\{v_{ij}\}$ and $\{w_{ij}\}$ as defined in (7), the following weighted negative kappa is proposed:

$$\begin{aligned} K_w^- &= -\left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}\right) \\ &= -\left(1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}}\right) \end{aligned} \tag{17}$$

and hence

$$K_{w(\text{corrected})} = \begin{cases} K_w \text{ in (5)-(6)}, & \text{if } \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij} \geq \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}, \\ K_w^- \text{ in (17)}, & \text{if } \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij} \leq \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}. \end{cases} \tag{18}$$

Except for the minus sign, K_w^- in (17) follows from (5)-(6) by simply substituting $\{w_{ij}\}$ for $\{v_{ij}\}$ in (5) and $\{v_{ij}\}$ for $\{w_{ij}\}$ in (6).

K_w^- is well defined if at least two cells of the $k \times k$ table contain nonzero probabilities. It is also apparent from (17) that K_w^- takes on values between 0 and −1, inclusive, with $K_w^- = 0$ if $p_{ij} = p_{i+}p_{+j}$ for all i and j (as a sufficient but not necessary condition). Also $K_w^- = -1$ if, and only if, $p_{ij} = 0$ for all i and j except for $i = 1$ and $j = k$ and $i = k$ and $j = 1$, that is, when the only nonzero probabilities occur in the corner cells $(1, k)$ and $(k, 1)$ and the weights are of the type of form as in (7). These properties of K_w^- all appear to be reasonable.

By contrast, if $p_{1k} \neq 0$, $p_{k1} \neq 0$, and all other $p_{ij} = 0$, K_w in (5)-(6) becomes $K_w = -2p_{1k}p_{k1}/(1 - 2p_{1k}p_{k1})$, which equals −1 only if $p_{1k} = p_{k1} = 0.5$. Otherwise, the value of K_w increases as p_{1k} and p_{k1} become increasingly different, approaching 0 as $|p_{1k} - p_{k1}|$ approaches 1. Such behavior of $K_w < 0$ makes any reasonable interpretation of negative K_w-values impossible and meaningless.

3.2. Specific Category Coefficients. Just as the K and K_w coefficients are inappropriate for negative values, so is the category-specific coefficient K_i in (3)-(4) as pointed out in Section 2. Therefore, for $K_i < 0$, another coefficient is needed that satisfies the reasonable requirements that its value equals 0 when $p_{ii} = p_{i+}p_{+i}$ and equals −1 when $p_{ii} = 0$. The following proposition seems most reasonable:

$$K_{i(\text{corrected})} = \begin{cases} K_i \text{ in (3)-(4)}, & \text{if } p_{ii} \geq p_{i+}p_{+i}, \\ K_i^- = -\left(1 - \dfrac{p_{ii}}{p_{i+}p_{+i}}\right), & \text{if } p_{ii} \leq p_{i+}p_{+i}. \end{cases} \tag{19}$$

K_i^- is well defined unless either $p_{i+} = 0$ or $p_{+i} = 0$ (and hence $p_{ii} = 0$). $K_i^- = -1$ if, and only if, $p_{ii} = 0$ (and $p_{i+} \neq 0$ and

$p_{+i} \neq 0$). For the weighted mean cell probabilities in Table 1, $K_1^- = K_2^- = -\lambda$. Also, analogous to (10),

$$K^- = \sum_{i=1}^{k} u_i K_i^-, \quad u_i = \frac{p_{i+}p_{+i}}{\sum_{i=1}^{k} p_{i+}p_{+i}} \tag{20}$$

with K^- defined in (13).

In terms of weights $v_{ij} = 1 - w_{ij}$, with the types of w_{ij} as in (7), the proposed specific-category weighted kappa coefficient may be defined as

$$K_{wi(\text{corrected})} = \begin{cases} K_{wi} \text{ in } (8), & \text{if } K_{wi} \geq 0, \\ K_{wi}^- = -\left(1 - \dfrac{\sum_{h=1}^{k} v_{hi}p_{hi} + \sum_{j=1}^{k} v_{ij}p_{ij}}{\sum_{h=1}^{k} v_{hi}p_{h+}p_{+i} + \sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}}\right), & \text{if } K_{wi} \leq 0, \end{cases} \tag{21}$$

where, as always, the first subscript refers to the table row and the second subscript to the table column. Note that the component for cell (i, i) appears twice in K_{wi}^-. Note also that, analogous to (20), K_w^- in (17) is the weighted arithmetic mean of $K_{w1}^-, \ldots, K_{wk}^-$ in (21) with weights based on the denominator in (21) for $i = 1, \ldots, k$. It is apparent from (21) that, for the weights in (7) and with $v_{ij} = 1 - w_{ij}$, $K_{wi}^- = -1$ if $p_{i+} = p_{+i} = 0$ for all i, but also $K_{w1}^- = K_{wk}^- = -1$ if p_{1k} and p_{k1} are the only nonzero probabilities in the table.

4. Statistical Inferences

Consider now that the coefficients (measures) discussed above are all sample estimates (and estimators) based on the sample probabilities $p_{ij} = n_{ij}/N$ $(i, j = 1, \ldots, k)$ with frequencies (counts) n_{ij} and sample size $N = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}$. p_{ij}'s are maximum likelihood estimates (and estimators) of the unknown population probabilities π_{ij} $(i, j = 1, \ldots, k)$ on which the corresponding population coefficients are based such as the population coefficient $K_w(\{\pi_{ij}\}) = 1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{ij} / \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{i+}\pi_{+j}$ corresponding to the sample coefficient K_w in (6). It may then be of interest to make statistical inferences about the population coefficients corresponding to the sample coefficients discussed above.

Such statistical inferences would probably be most meaningful in terms of the construction of confidence intervals for the overall kappa coefficients in (14) and (18). The inference procedure needs necessarily to be approximated for reasonably large sample size N and be based on the *delta method* (e.g., [19, Chapter 14]) or resampling methods such as the *bootstrap* and the *jackknife* (e.g., [20, 21]). The delta method is chosen in this paper. By developing the procedure based on the K_w expression in (6), the procedures for K_w^- in (17), K^- in (13), and K in (1) follow as special cases by the appropriate selection of the set of weights $\{w_{ij}\}$. Fleiss et al. [2] gave the estimated large sample variance of K_w based on the expression in (5) without presenting any intermediate steps. Instead, the expression in (6) will be used here as being more convenient and some of the important intermediate steps will be presented.

Then, letting K_w in (6) denote both the sample estimate and estimator of the corresponding population coefficient $K_w(\{\pi_{ij}\})$ (based on population probabilities π_{ij}, π_{i+}, and π_{+j} for $i = 1, \ldots, k$ and $j = 1, \ldots, k$), it follows from the delta method that, under multinomial sampling (when the k categories and the sample size N are *a priori* fixed), the estimator K_w is approximately normally distributed with mean $K_w(\{\pi_{ij}\})$ and estimated variance $\widehat{\text{Var}}(K_w)$ if N is reasonably large.

In order to derive the estimated variance of K_w, express $K_w(\{\pi_{ij}\})$ as $K_w(\{\pi_{ij}\}) = 1 - a/b = 1 - d$ and let a_{ij}', b_{ij}', and d_{ij}' denote the partial derivatives of these quantities with respect to π_{ij}, with π_{ij} then being replaced with the estimated probabilities p_{ij} for all i and j. Then,

$$\begin{aligned} \widehat{\text{Var}}(K_w) &= \widehat{\text{Var}}(d) \\ &= N^{-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left(d_{ij}'\right)^2 - \left(\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}d_{ij}'\right)^2\right], \end{aligned} \tag{22}$$

where

$$d_{ij}' = b^{-1}w_{ij} - ab^{-2}b_{ij}' \tag{23}$$

for all i and j. It is found that

$$b_{ij}' = \sum_{j=1}^{k} w_{ij}p_{+j} + \sum_{i=1}^{k} w_{ij}p_{i+} = \overline{w}_{i+} + \overline{w}_{+j} \tag{24}$$

so that, from (23)-(24),

$$\begin{aligned} d_{ij}' &= b^{-1}w_{ij} - ab^{-2}\left(\overline{w}_{i+} + \overline{w}_{+j}\right) \\ &= b^{-1}\left[w_{ij} - \left(1 - K_w\right)\left(\overline{w}_{i+} + \overline{w}_{+j}\right)\right] \end{aligned} \tag{25}$$

from which one gets

$$\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}d_{ij}' = -ab^{-1} = -\left(1 - K_w\right). \tag{26}$$

When (25) and (26) are substituted into (22), one obtains

$$\begin{aligned}\widehat{\mathrm{Var}}\left(K_w\right) \\ &= N^{-1}\left\{b^{-2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[w_{ij}-\left(1-K_w\right)\left(\overline{w}_{i+}+\overline{w}_{+j}\right)\right]^2 \right. \\ &\left. -\left(1-K_w\right)^2\right\},\end{aligned} \tag{27}$$

where $b=\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}$ and $\overline{w}_{i+}$ and $\overline{w}_{+j}$ are defined in (24). This variance formula, which gives the same numerical results as the formula given by Fleiss et al. [22], can then be used for interval estimation.

By comparing $|K_w^-|$ in (17) with the expression for K_w in (6), it follows from (27) that the estimated variance of K_w^- is given by

$$\begin{aligned}\widehat{\mathrm{Var}}\left(K_w^-\right) \\ &= N^{-1}\left\{b^{-2}\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left[v_{ij}-\left(1-\left|K_w^-\right|\right)\left(\overline{v}_{i+}+\overline{v}_{+j}\right)\right]^2 \right. \\ &\left. -\left(1-\left|K_w^-\right|\right)^2\right\},\end{aligned} \tag{28}$$

where $b=\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}$ is the denominator in (17) and

$$\begin{aligned}\overline{v}_{i+} &= \sum_{j=1}^{k} v_{ij}p_{+j}, \\ \overline{v}_{+j} &= \sum_{i=1}^{k} v_{ij}p_{i+}.\end{aligned} \tag{29}$$

By setting $w_{ij}=1$ for all $i\neq j$ and $w_{ij}=0$ for all $i=j$, K_w in (6) reduces to K in (2) and, furthermore, $\overline{w}_{i+}=1-p_{+i}$ and $\overline{w}_{+j}=1-p_{j+}$ in (24) so that, from (27), it is found that

$$\begin{aligned}\widehat{\mathrm{Var}}(K) = N^{-1}P_{\mathrm{DC}}^{-2}\Bigg[&P_{\mathrm{DO}}\left(1-P_{\mathrm{DO}}\right) \\ &+(1-K)^2\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left(2-p_{+i}-p_{j+}\right)^2 \\ &-2(1-K)\sum_{i=1,i\neq j}^{k}\;\sum_{j=1,i\neq j}^{k} p_{ij}\left(2-p_{+i}-p_{j+}\right)\Bigg],\end{aligned} \tag{30}$$

where P_{DO} and P_{DC} are defined in (2). Similarly, by setting $v_{ij}=1$ for all $i=j$ and $v_{ij}=0$ for all $i\neq j$, K_w^- in (17) reduces to K^- in (13) and, furthermore, $\overline{v}_{i+}=p_{+i}$ and $\overline{v}_{+j}=p_{j+}$ in (29) so that, from (28), the following result is obtained:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(K^-\right) = N^{-1}P_{\mathrm{AC}}^{-2}\Bigg[&P_{\mathrm{AO}}\left(1-P_{\mathrm{AO}}\right) \\ &+\left(\frac{P_{\mathrm{AO}}}{P_{\mathrm{AC}}}\right)^2\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}\left(p_{+i}+p_{j+}\right)^2 \\ &-2\left(\frac{P_{\mathrm{AO}}}{P_{\mathrm{AC}}}\right)\sum_{i=1}^{k} p_{ii}\left(p_{+i}+p_{i+}\right)\Bigg],\end{aligned} \tag{31}$$

where P_{AO} and P_{AC} are defined in (1). The expression in (30) is somewhat different from that given by Fleiss et al. [22], but they are found to give exactly the same numerical results.

If it should be of interest to test the null hypothesis that the population equivalent to one of the new coefficients is equal to zero, then the same procedure as proposed by Fleiss et al. [22] for the case of Cohen's K and K_w would involve replacing p_{ij} with $p_{i+}p_{+j}$ for all i and j in the variance expressions in (28) or (30). However, a simpler method would be to use the chi-square goodness-of-fit statistics χ^2 or G^2 to test for independence (noting again that independence is a sufficient but not necessary condition for the coefficients to equal zero).

5. Numerical Examples

5.1. Example 1: K^-. Instead of one pair of observers assigning each of N items (observations) to one of k categories, consider the statistically equivalent situation in which each of N pairs of observers assigns one item to one of k categories. For example, among $N = 100$ randomly selected couples, each spouse answers a question with $k = 3$ choice categories C_1, C_2, and C_3. The (fictitious) data are given in Table 2.

With $P_{\mathrm{AO}} = 0.12$ and $P_{\mathrm{AC}} = 0.3410$ in Table 2, it follows from (13) that $K^- = -0.65$, indicating a substantial disagreement between husbands and wives. By contrast, the corresponding value of Cohen's K in (1) is found to be $K = -0.34$, which could have been interpreted as indicating a much lower level of disagreement. However, since K has no fixed lower bound as discussed above, any interpretation or conclusion based on $K = -0.34$ would be invalid and misleading.

The next question may then be, how do the disagreements on the individual categories contribute to the overall disagreement of $K^- = -0.65$? The answer from Table 2 and (19) is found to be $K_1^- = -0.80$, $K_2^- = -0.79$, and $K_3 = 0.10$. Therefore, the substantial overall disagreement $K^- = -0.65$ is attributable to the high disagreement on each of the categories C_1 and C_2, whereas category C_3 involves a very low level of agreement. By comparison, the negative values from (3) or (4) would have been substantially different, with $K_1 = -0.62$ and $K_2 = -0.34$.

In order to construct a confidence interval for the population equivalent $K^-(\{\pi_{ij}\})$ of K^- based on the data in Table 2, it is found from (31) that $\widehat{\mathrm{Var}}(K^-) = 0.0115$. Then, since the estimator K^-, with the sample estimate of -0.65, is

Table 2: Results from $N = 100$ couples answering a multiple-choice question with three choice categories (fictitious data).

Wives	Husbands			Total
	C_1	C_2	C_3	
C_1	0.04	0.35	0.21	0.60
C_2	0.22	0.02	0.01	0.25
C_3	0.08	0.01	0.06	0.15
Total	0.34	0.38	0.28	1.00

approximately normally distributed with mean $K^-(\{\pi_{ij}\})$ and estimated variance of 0.0115, an approximate 95% confidence interval for $K^-(\{\pi_{ij}\})$ becomes $-0.65 \pm 1.96\sqrt{0.0115}$ or $[-0.86, -0.44]$.

5.2. Example 2: K_w^-. Consider now that the categories in Table 2 are ordinal so that the weighted kappa coefficients would be appropriate. Then, with the weights $w_{ij} = |i-j|/(k-1)$ in (7) and $v_{ij} = 1 - w_{ij}$ for all i and j, it is found from Table 2 that $\sum_{i=1}^{3}\sum_{j=1}^{3} v_{ij}p_{ij} = 0.4150$ and $\sum_{i=1}^{3}\sum_{j=1}^{3} v_{ij}p_{i+}p_{+j} = 0.5610$ so that, from (17)-(18), $K_w^- = -0.2602 = -0.26$. This weighted disagreement value differs considerably from the above $K^- = -0.65$ value when the three categories are considered to be nominal.

In terms of the disagreements for the individual categories, it is found from (21) and Table 2 that $K_{w1}^- = -0.35$, $K_{w2}^- = -0.18$, and $K_{w3}^- = -0.12$. Again, these results differ considerably from those in the nominal case considered in Example 1. Note that the arithmetic mean -0.22 of K_{wi}^-'s does not differ greatly from the overall $K_w^- = -0.26$.

An interval estimate for the population measure $K_w^-(\{\pi_{ij}\})$ can be derived from (28) by first computing $\overline{v}_{i+}$ and $\overline{v}_{+j}$ for each of i and j from (29) and Table 2, giving $\overline{v}_{i+} = 0.5300, 0.6900$, and 0.4700 for $i = 1, 2$, and 3, respectively, and $\overline{v}_{+j} = 0.7250, 0.6250$, and 0.2750 for $j = 1, 2$, and 3, respectively. Then, from (28), with $b = 0.5610$ and $K_w^- = -0.2602$, it is found that $\widehat{\text{Var}}(K_w^-) = 0.0028$. Consequently, a 95% confidence interval for $K_w^-(\{\pi_{ij}\})$ is given by $-0.2602 \pm 1.96\sqrt{0.0028}$ or $[-0.36, -0.16]$.

5.3. Logistic Transformation. Instead of making statistical inferences about the kappa coefficients directly, as done above, it is likely advantageous to do so indirectly via the logistic transformation. Therefore, in the case of K^- in (13), consider the following logistic transformation of $1 + K^-$ and its inverse:

$$L = \log\left(\frac{1+K^-}{-K^-}\right), \quad K^- = -\frac{1}{1+e^L}. \tag{32}$$

Since the derivative $dL/dK^- = -1/K^-(1+K^-)$, the estimated variance of L becomes

$$\widehat{\text{Var}}(L) = \left(\frac{1}{K^-(1+K^-)}\right)^2 \widehat{\text{Var}}(K^-), \tag{33}$$

where $\widehat{\text{Var}}(K^-)$ is given in (31). An approximate confidence interval for the population equivalent of L can then be constructed based on (33), with the corresponding confidence interval for $K^-(\{\pi_{ij}\})$ resulting from the inverse transform in (32).

In the case of K_w^- in (17), K^- in (32)-(33) is simply replaced with K_w^-. For K and K_w in (1) and (5)-(6), the transformation becomes $\log[K/(1-K)]$ and $\log[K_w/(1-K_w)]$. With such transformations, the lower end of a confidence interval for K^- or K_w^- cannot be less than -1 and the upper end of a confidence interval K or K_w cannot exceed 1. Most importantly, the normal distribution approximation is likely to be improved with the above logistic transforms. Unless the sample size N is very large, the distributions of the kappa coefficients are likely to be skewed, especially when a coefficient is near -1 or 1. For instance, when, say, the population coefficient $K^-(\{\pi_{ij}\}) = -0.9$, the estimator K^- cannot be much smaller than $K^-(\{\pi_{ij}\})$, but it could be much larger with nonnegligible probability. The logistic transformation to the $(-\infty, \infty)$-interval tends to correct for such skewness and provide for a more rapid convergence to normality.

In Example 1, with $K^- = -0.6481$ and $\widehat{\text{Var}}(K^-) = 0.0115$, it follows from (32)-(33) that $L = -0.6107$ and $\widehat{\text{Var}}(L) = 0.2211$. An approximate 95% confidence interval for the population equivalent of L is then given by $-0.6107 \pm 1.96\sqrt{0.2211}$ or $[-1.5323, 0.3109]$. Then, from the inverse transform in (32), it follows that an approximate 95% confidence interval for $K^-(\{\pi_{ij}\})$ has the limits $-1/(1+e^{-1.5323}) = -0.82$ and $-1/(1+e^{0.3109}) = -0.42$, that is, the interval $[-0.82, -0.42]$. This confidence interval is slightly shorter than the interval $[-0.86, -0.44]$ determined above. Similarly, with K_w^- substituted for K^- in (32)-(33) and with $K_w^- = -0.2602$ and $\widehat{\text{Var}}(K^-) = 0.0028$ from Example 2, it is found that an approximate 95% confidence interval for the population coefficient $K_w^-(\{\pi_{ij}\})$ is $[-0.38, -0.17]$. This interval differs little from the interval $[-0.36, -0.16]$ determined above when applying the inference procedure directly to K_w^-.

6. Conclusion

If Cohen's kappa is accepted as an appropriate measure of interobserver agreement, as many do judging by its widespread use, then the corrections proposed here for negative kappa values should be equally acceptable. Of course, since the chance-expected disagreement (or agreement) terms in the new coefficients also depend exclusively on the marginal distributions, the criticism by some that Cohen's coefficients depend too much on the marginal distributions would similarly apply to the new coefficients. Such concern is particularly important in cases of highly uneven (nonuniform or "skewed") marginal distributions. If, however, those distributions are fairly even (uniform), Cohen's kappa and hence the measures proposed in this paper for interobserver disagreement (negative agreement) would seem to be reasonably acceptable agreement-disagreement measures.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The author gratefully acknowledges the helpful recommendations from a reviewer.

References

[1] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[2] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, Wiley, Hoboken, NJ, USA, 3rd edition, 2003.

[3] K. L. Gwet, *Handbook of Inter-Rater Reliability*, Advanced Analytics, Gaithersburg, Md, USA, 4th edition, 2014.

[4] M. M. Shoukri, *Measures of Interobserver Agreement and Reliability*, CRC Press, Boca Raton, Fla, USA, 2nd edition, 2011.

[5] A. Von Eye and E. Y. Mun, *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2005.

[6] M. J. Warrens, "New interpretations of Cohen's kappa," *Journal of Mathematics*, vol. 2014, Article ID 203907, 9 pages, 2014.

[7] M. J. Warrens, "Five ways to look at Cohen's kappa," *Journal of Psychology & Psychotherapy*, vol. 5, no. 4, pp. 1–4, 2015.

[8] Z. Yang and M. Zhou, "Kappa statistic for clustered matched-pair data," *Statistics in Medicine*, vol. 33, no. 15, pp. 2612–2633, 2014.

[9] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[10] H. L. Kundel and M. Polansky, "Measurement of observer agreement," *Radiology*, vol. 228, no. 2, pp. 303–308, 2003.

[11] R. L. Spitzer, J. Cohen, J. L. Fleiss, and J. Endicott, "Quantification of agreement in psychiatric diagnosis," *Archives of General Psychiatry*, vol. 17, no. 1, pp. 83–87, 1967.

[12] T. O. Kvålseth, "Note on Cohen's kappa," *Psychological Reports*, vol. 65, no. 1, pp. 223–226, 1989.

[13] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[14] D. V. Cicchetti and T. Allison, "A new procedure for assessing reliability of scoring EEG sleep recordings," *American Journal of EEG Technology*, vol. 11, no. 3, pp. 101–109, 1971.

[15] T. O. Kvålseth, "Weighted specific-category Kappa measure of interobserver agreement," *Psychological Reports*, vol. 93, no. 3, pp. 1283–1290, 2003.

[16] T. O. Kvålseth, "Kappa coefficients of agreement," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., part 11, pp. 710–713, Springer, Berlin, Germany, 2011.

[17] H. E. A. Tinsley and D. J. Weiss, "Interrater reliability and agreement," in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, H. E. A. Tinsley and S. D. Brown, Eds., pp. 95–124, Academic Press, San Diego, Calif, USA, 2000.

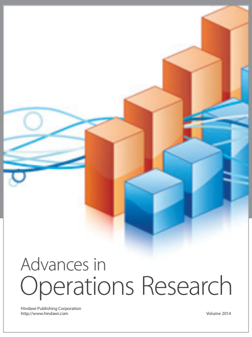
[18] K. F. Hirji and M. H. Rosove, "A note on interrater agreement," *Statistics in Medicine*, vol. 9, no. 7, pp. 835–839, 1990.

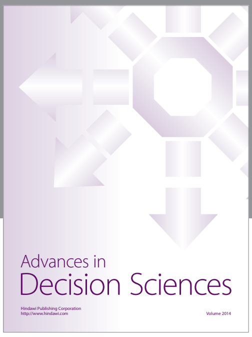
[19] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2002.

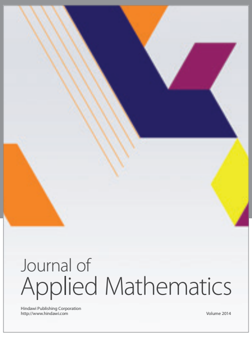
[20] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, UK, 1997.

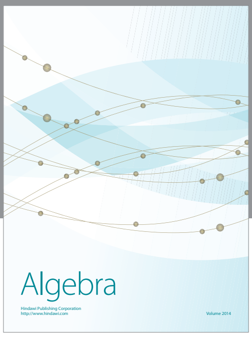
[21] W. C. Parr and H. D. Tolley, "Jackknifing in categorical data analysis," *The Australian Journal of Statistics*, vol. 24, no. 1, pp. 67–79, 1982.

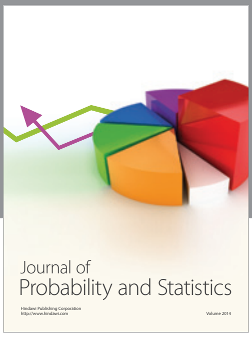
[22] J. L. Fleiss, J. Cohen, and B. S. Everitt, "Large sample standard errors of kappa and weighted kappa," *Psychological Bulletin*, vol. 72, no. 5, pp. 323–327, 1969.


Advances in
Operations Research

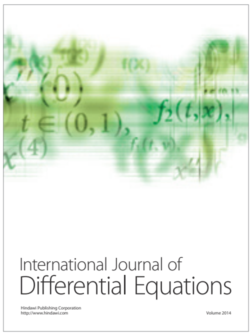

Advances in
Decision Sciences


Journal of
Applied Mathematics

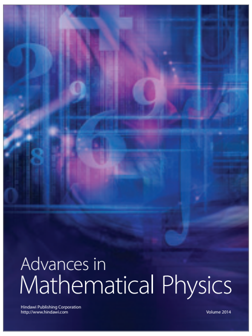

Algebra

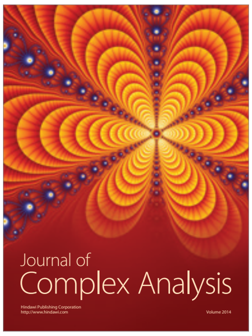

Journal of
Probability and Statistics

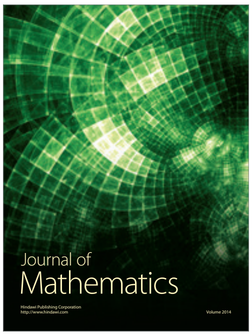
The Scientific
World Journal

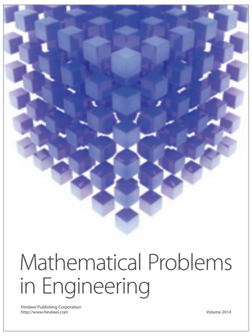

International Journal of
Differential Equations

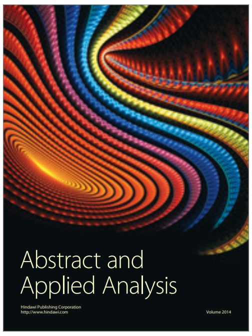
International Journal of
Combinatorics

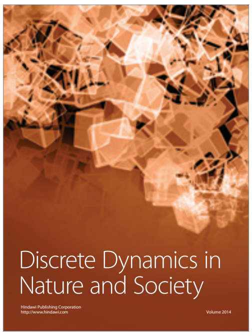

Advances in
Mathematical Physics

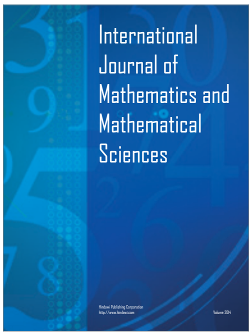

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

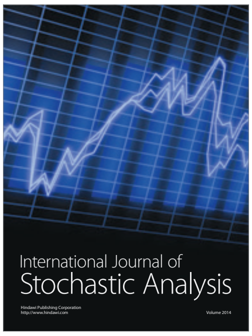

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization