Hacettepe Journal of Mathematics and Statistics
Volume 48 (2) (2019), 592 – 604
RESEARCH ARTICLE

# Agreement and adjusted degree of distinguishability for square contingency tables

Ayfer Ezgi Yilmaz*[†] and Tulay Saracbasi[‡]

## Abstract

In square contingency tables, analysis of agreement between the row and column classifications is of interest. In such tables, kappa or weighted kappa coefficients are used to summarize the degree of agreement between two raters. In addition to investigate the agreement between raters for square contingency tables, category distinguishability should be considered. Because the kappa coefficient is insufficient to measure the category distinguishability, the degree of distinguishability is suggested to use. In practice, some problems have occurred with regards to the use of the degree of distinguishability. The aim of this study is to assess the agreement coefficient and degree of distinguishability in square contingency tables together. In this study, the adjusted degree of distinguishability is suggested to solve the problem of calculating the degree of distinguishability falls outside the defined range. A simulation study is performed to compare the proposed adjusted degree of distinguishability and the classical degree of distinguishability. Furthermore, interpretation levels for the degree of distinguishability are determined based on a simulation study. The results are discussed over numerical examples and simulation.

**Keywords:** Agreement, Degree of distinguishability, Kappa coefficient, Square contingency table.

*Mathematics Subject Classification (2010):* 62 H 17

*Received* : 14.04.2017 *Accepted* : 10.09.2018 *Doi* : 10.15672/HJMS.2018.620

*Department of Statistics,Hacettepe University, Ankara, Turkey, Email: ezgiyilmaz@hacettepe.edu.tr

[†]Corresponding Author.

[‡]Department of Statistics,Hacettepe University, Ankara, Turkey, Email: toker@hacettepe.edu.tr

## 1. Introduction

Square contingency tables are often used in medical, sociology, and behavioral sciences. These tables may arise in different ways, such as: When a sample of individuals or subjects is cross-classified according to two essentially similar categorical variables; when samples of pairs of matched individuals or subjects are classified according to some categorical variable of interest; in panel studies where each individual or subject in a sample is classified according to the same criterion at two different points in time; in rating experiments in which a sample of individuals or subjects is rated independently by the same two raters into one of the categories [12].

In square contingency tables, analysis of agreement between the row and column classifications is of interest. Interrater agreement represents the extent to which different judges tend to assign exactly the same rating for each object [18]. The agreement between objects rated independently by two raters or in two different time points by the same rater is investigated with the agreement coefficients. The degree of agreement is assessed using Cohen's kappa coefficient [4].

Even though the raters rate the items independently, there occurs correlation between their decisions. There are two main components of agreement [6, 19]:

(1) Marginal homogeneity which corresponds to the differences in the marginal distributions of raters.
(2) The category distinguishability which is the ability for raters to distinguish the categories.

In the agreement studies, it is necessary to determine if the categories of the table are distinguishable from one to another [14]. If the categories are indistinguishable, then there could occur some differences between raters' perceptions. Different raters man understand the categories differently or the same rater may not distinguish the categories correctly. It is discussed that these two problems can occur because the raters may not be experts in their fields or it may be difficult to distinguish the categories. The measure to calculate the distinguishability level of the categories is called degree of distinguishability [6].

In practice, there occurs some problems to the use of the degree of distinguishability. The value of the measure falls outside the defined range in some tables. Furthermore, there is not any information about how to interpret the degree of distinguishability except the general one. In this article, the adjusted degree of distinguishability is suggested to solve the problem of calculating the degree of distinguishability falls outside the defined range. It is aimed to assess the agreement coefficient and the adjusted degree of distinguishability in square contingency tables together. A simulation study is performed to compare the proposed adjusted degree of distinguishability with the classical one. Furthermore, interpretation levels for the degree of distinguishability are determined based on a simulation study. The results are discussed over three numerical examples and simulation.

Agreement coefficients and degree of distinguishability are reviewed in Section 2. Section 3 presents the suggested adjusted degree of distinguishability. The simulation study results are summarized in Section 4. The illustrative examples are discussed in Section 5, followed by conclusion in Section 6.

## 2. Agreement Coefficients

There is a large literature on agreement coefficients. There are numerous agreement coefficients for each table structure or number of raters. The well-known agreement coefficient for nominal categories is Cohen's kappa coefficient [4]. When the categories are ordinal, instead of kappa, Cohen's weighted kappa coefficient is suggested for use [5]. Darroch and McCloud [6] recommend the degree of distinguishability to be used in place of kappa.

**2.1. Cohen's Kappa and Weighted Kappa Coefficients.** Consider two raters classify the objects from a population $n$ on a $R$ scale. Let $n_{ij}$ denote the number of objects $(i, j = 1, 2, \ldots, R)$. The cell probabilities are $p_{ij}$ and $p_{i.}$ indicates the $i$th row total probability, $p_{.j}$ indicates the $j$th column total probability of an $R \times R$ contingency table. The kappa coefficient $\kappa$ is calculated as

$$\kappa = \frac{\sum_{i=1}^{R} p_{ii} - \sum_{i=1}^{R} p_{i.}p_{.i}}{1 - \sum_{i=1}^{R} p_{i.}p_{.i}}. \tag{2.1}$$

For ordinal responses, instead of kappa, weighted kappa coefficient is suggested by Cohen (1968). The coefficient allows each $(i, j)$ cell to be weighted according to the degree of agreement between $i$th and $j$th categories [16]. The weighted kappa coefficient $\kappa_w$ is calculated as

$$\kappa_w = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i.}p_{.j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i.}p_{.j}} \tag{2.2}$$

where $w_{ij}$ is the weight ranges $0 \leq w_{ij} \leq 1$. The popular weights for weighted kappa are the linear and the quadratic weights shown in Equations (2.3) and (2.4), respectively [3, 7].

- Linear weights:

$$w_{ij} = 1 - |i - j|/(R - 1) \tag{2.3}$$

- Quadratic weights:

$$w_{ij} = 1 - (i - j)^2/(R - 1)^2. \tag{2.4}$$

In the literature, there are several interpretations of $\kappa$ coefficient. Landis and Koch [10] define the agreement levels of kappa coefficient as: "<0.00" poor, "0.00-0.20" slight, "0.21-0.40" fair, "0.41-0.60" moderate, "0.61-0.80" substantial, and "0.81-1.00" almost perfect.

**2.2. Degree of Distinguishability.** Degree of distinguishability ($DD$) is suggested to investigate the ability of the raters to distinguish between two categories [6]. The category distinguishability is defined in terms of the following odds ratio.

$$\tau_{ij} = \frac{n_{ii}n_{jj}}{n_{ij}n_{ji}}, \quad i < j. \tag{2.5}$$

The degree of distinguishability $(\delta_{ij})$ of $i$th and $j$th categories is

$$\delta_{ij} = 1 - \tau_{ij}^{-1}, \tag{2.6}$$

where $0 \leq \delta_{ij} \leq 1$. When $\delta_{ij} \cong 1$, then there is a perfect distinguishability between these two categories. When $\delta_{ij} \cong 0$, then it is impossible to distinguish between these two categories and this is not a preferred situation in the studies.

## 3. The Adjusted Degree of Distinguishability

Darroch and McCloud [6] defined the degree of distinguishability between two categories ranges from 0 to 1. In the applications, the degree of distinguishability may be calculated outside the defined range as negative.

Carcinoma in situ of uterine cervix data, one of the most common data in agreement studies, is an illustrative example where this problem observed. The data is discussed by Holmquist *et al.* [9], Landis and Koch [11], Becker and Agresti [2], and Saracbasi [15]. In order to investigate the variability in the classification of carcinoma in situ of the uterine cervix, seven pathologists classified 118 slides into the 5 categories: (1) Negative, (2) Atypical squamous hyperplasia, (3) Carcinoma in situ, (4) Squamous carcinoma with early stromal invasion, and (5) Invasive carcinoma. Two of the seven pathologists are chosen and given in Table 1.

**Table 1.** Independent classifications by two pathologists of most involved histological lesion

| | **Pathologist 2** | | | | |
|---|---|---|---|---|---|
| **Pathologist 1** | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** |
| **(1)** | 26 | 0 | 0 | 0 | 0 |
| **(2)** | 20 | 6 | 0 | 0 | 0 |
| **(3)** | 10 | 19 | 9 | 0 | 0 |
| **(4)** | 5 | 5 | 11 | 0 | 1 |
| **(5)** | 1 | 1 | 0 | 1 | 3 |

For Table 1, the degree of distinguishabilities are calculated and given in Table 2. The results in Table 2 show that the degree of distinguishability of (3) Carcinoma in situ and (4) Squamous carcinoma with early stromal invasion is $\delta_{34} = -0.21$ and the degree of distinguishability of (4) Squamous carcinoma with early stromal invasion and (5) Invasive carcinoma is $\delta_{45} = -4.11$ which fall outside of the defined range.

**Table 2.** The category distinguishabilities measures of carcinoma in situ of uterine cervix data

| | $\delta_{12}$ | $\delta_{13}$ | $\delta_{14}$ | $\delta_{15}$ | $\delta_{23}$ | $\delta_{24}$ | $\delta_{25}$ | $\delta_{34}$ | $\delta_{35}$ | $\delta_{45}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Estimate | 0.94 | 0.98 | 0.79 | 0.99 | 0.84 | 0.15 | 0.97 | -0.21 | 0.99 | -4.14 |

This well-known data illustrates the problem of the degree of distinguishability and shows the necessity of a new formulation. In this study, we proposed the adjusted degree of distinguishability ($ADD$) to calculate the distinguishabilities between adjacent categories.

We proposed the adjusted degree of distinguishability under two arguments. Firstly, when the category distinguishability is discussed, it should be considered that the distinguishability of $i$th and $j$th categories is equal to the distinguishability of $j$th and $i$th categories. Secondly, if categories (1) and (2) are distinguishable and categories (2) and (3) are distinguishable, then it is reasonable if categories (1) and (3) are distinguishable as well. For this reason, it is sufficient to calculate degree of distinguishability for only adjacent categories instead of all the pairs.

The adjusted degree of distinguishability ($ADD$) for $i$ and $i+1$ categories is calculated as

$$(3.1) \qquad ADD_{i,i+1} = \begin{cases} 1 - \tau_{i,i+1}^{-1} & \text{if } \tau_{i,i+1} \geq 1 \\ 1 - \tau_{i,i+1} & \text{if } \tau_{i,i+1} < 1 \end{cases}$$

where $0 \leq ADD_{i,i+1} \leq 1$, $i = 1, 2, \ldots, (R-1)$. The odds ratio for square contingency tables is

$$(3.2) \qquad \tau_{i,i+1} = \frac{n_{ii}\ n_{i+1,i+1}}{n_{i,i+1}\ n_{i+1,i}}.$$

For Table 1, the adjusted degree of distinguishabilities are calculated and given in Table 3. The results in Table 3 show that the adjusted degree of distinguishability of (3) Carcinoma in situ and (4) Squamous carcinoma with early stromal invasion is $ADD_{34} = 0.17$, (4) Squamous carcinoma with early stromal invasion and (5) Invasive carcinoma is $ADD_{45} = 0.81$.

**Table 3.** The adjusted degree of distinguishabilities of carcinoma in situ of uterine cervix data

| | $ADD_{12}$ | $ADD_{23}$ | $ADD_{34}$ | $ADD_{45}$ |
|---|---|---|---|---|
| Estimate | 0.94 | 0.84 | 0.17 | 0.81 |

If the table contains sampling zeros, then the odds ratio is

$$(3.3) \qquad \tau_{i,i+1} = \frac{(n_{ii} + c)(n_{i+1,i+1} + c)}{(n_{i,i+1} + c)(n_{i+1,i} + c)},$$

where $c$ is a constant value that can be 0.20, 0.50, or a minimum value which is different from zero [1].

## 4. Simulation Study

A simulation study is performed to compare the proposed adjusted degree of distinguishability with the classical one. It is also aimed to develop a table to interpret the adjusted degree of distinguishability.

To generate 2 × 2 contingency tables, we used the method presented by Goktas and Isci [8]. Bivariate standard normal distribution is used. At the first step, two identically independently distributed random variables ($X_1$ and $X_2$) are generated. Equations (4.1) and (4.2) is used to generate two random variables ($X$ and $Y$) from bivariate normal distribution with certain correlation ($\rho$).

$$(4.1) \qquad X = aX_1 + bX_2$$

$$(4.2) \qquad Y = bX_1 + aX_2$$

where

$$a = \frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2},$$

and

$$b = \frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}.$$

Then, $X$ and $Y$ variables are categorized into two equal intervals and crossed to have $2 \times 2$ tables. The sample sizes ($n$) of the table are considered as 30, 50, 70, 100, and 300. $\rho$ values are taken as 0.20, 0.50, and 0.80. The kappa coefficient, classical and adjusted degree of distinguishabilities are calculated for each table. All the results are based on 50,000 replications of each sample.

Table 4 shows the minimum, maximum values, median, mean, and standard errors of the classical and adjusted degree of distinguishabilities for different sample sizes and the different values of correlation. While some of the minimum values classical degrees of distinguishability are negative, $ADD$ lies between 0 and 1. In that case, ADD should be used instead of $DD$.

**Table 4.** The descriptive statistics of the classical and adjusted degree of distinguishabilities for different sample sizes and the different values of correlation

| | | *DD* | | | | | *ADD* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | $n$ | Min | Med | Max | Mean | S.E. | Min | Med | Max | Mean | S.E. |
| 0.20 | 30 | -29.18 | 0.4167 | 0.99 | 0.2043 | 0.0036 | 0.00 | 0.4857 | 0.99 | 0.4712 | 0.0011 |
| | 50 | -7.33 | 0.4000 | 0.96 | 0.2985 | 0.0021 | 0.00 | 0.4385 | 0.96 | 0.4268 | 0.0010 |
| | 70 | -3.55 | 0.4119 | 0.94 | 0.3336 | 0.0016 | 0.00 | 0.4271 | 0.94 | 0.4092 | 0.0010 |
| | 100 | -2.72 | 0.4023 | 0.92 | 0.3549 | 0.0012 | 0.00 | 0.4092 | 0.92 | 0.3937 | 0.0009 |
| | 300 | -0.63 | 0.4023 | 0.79 | 0.3885 | 0.0006 | 0.00 | 0.4023 | 0.79 | 0.3905 | 0.0006 |
| 0.50 | 30 | -8.00 | 0.7576 | 1.00 | 0.6843 | 0.0014 | 0.00 | 0.7576 | 1.00 | 0.7064 | 0.0010 |
| | 50 | -1.72 | 0.7656 | 0.99 | 0.7159 | 0.0009 | 0.00 | 0.7656 | 0.99 | 0.7189 | 0.0008 |
| | 70 | -0.81 | 0.7586 | 0.99 | 0.7266 | 0.0007 | 0.00 | 0.7586 | 0.99 | 0.7272 | 0.0007 |
| | 100 | -0.28 | 0.7567 | 0.97 | 0.7338 | 0.0005 | 0.00 | 0.7567 | 0.97 | 0.7338 | 0.0005 |
| | 300 | 0.34 | 0.7513 | 0.92 | 0.7453 | 0.0003 | 0.34 | 0.7513 | 0.92 | 0.7453 | 0.0003 |
| 0.90 | 30 | -0.91 | 0.9441 | 1.00 | 0.9190 | 0.0004 | 0.00 | 0.9441 | 1.00 | 0.9191 | 0.0004 |
| | 50 | 0.15 | 0.9394 | 1.00 | 0.9257 | 0.0003 | 0.15 | 0.9394 | 1.00 | 0.9257 | 0.0003 |
| | 70 | 0.41 | 0.9385 | 1.00 | 0.9285 | 0.0002 | 0.41 | 0.9385 | 1.00 | 0.9285 | 0.0002 |
| | 100 | 0.62 | 0.9372 | 1.00 | 0.9301 | 0.0002 | 0.62 | 0.9372 | 1.00 | 0.9301 | 0.0002 |
| | 300 | 0.78 | 0.9350 | 0.98 | 0.9320 | 0.0001 | 0.78 | 0.9350 | 0.98 | 0.9320 | 0.0001 |

While there is a medium correlation between raters where $n = 300$ and while there is high correlation between raters where $n > 30$, the classical and adjusted degrees of distinguishabilities are equal. The results in Table 4 show that when the correlation between raters increases, the classical and adjusted degrees of distinguishabilities also increase. While there is a low correlation and the sample size increases, the classical and adjusted degree of distinguishabilities decrease. However, while there is medium or high correlation, the classical and adjusted degree of distinguishabilities are not affected by the sample sizes. As expected, when the sample size increases, standard error decreases.

The scatter plots of the classical and adjusted degree of distinguishabilities for different sample sizes and the different values of correlation are given in Figure 1. Figure 1 shows that the negative values of degree of distinguishabilities are relocated to [0,1] interval. When the negative values of $DD$ diverge from 0, the values of $ADD$ converge to perfect agreement.

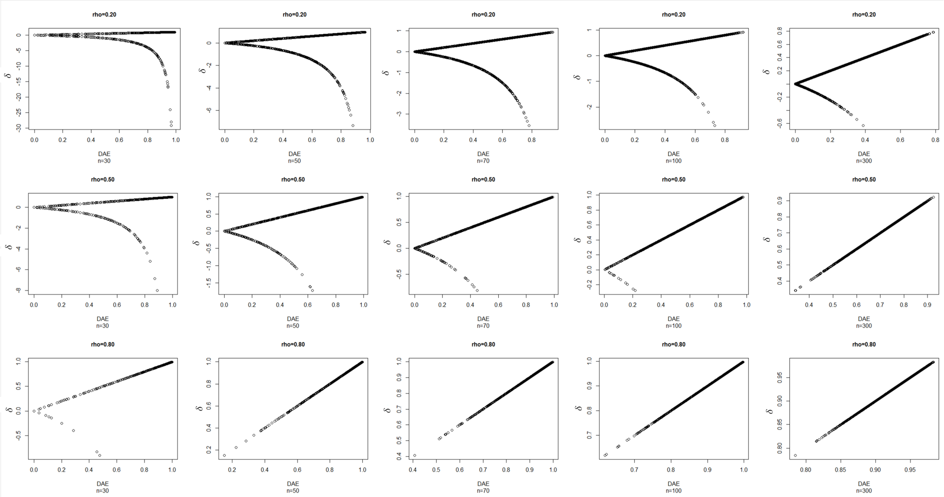

**Figure 1.** The scatter plots of the classical and adjusted degrees of distinguishabilities for different values of sample size and correlation

The kappa coefficients calculated for the tables generated randomly are classified into six categories considering Landis and Koch [10] intervals. Then, the minimum, maximum values, and medians of the adjusted degree of distinguishabilities are calculated for each kappa interval. The results are summarized in Table 5 and Figure 2. The aim of this study is to investigate the kappa coefficient and the adjusted category distinguishability together. Besides, it is purposed to investigate the $ADD$ intervals according to the kappa intervals.

**Table 5.** The minimum, maximum values, and medians of the adjusted degree of distinguishabilities for each kappa interval

| | $\kappa$ | <0.00 | | | 0.00-0.20 | | | 0.21-0.40 | | | 0.41-0.60 | | | 0.61-0.80 | | | 0.81-1.00 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho$ | Min | Med | Max | Min | Med | Max | Min | Med | Max | Min | Med | Max | Min | Med | Max | Min | Med | Max |
| 30 | 0.20 | 0.04 | 0.3330 | 0.97 | 0.00 | 0.3750 | 0.80 | 0.60 | 0.7270 | 0.95 | 0.84 | 0.8750 | 0.97 | 0.95 | 0.9650 | 0.99 | – | – | – |
| | 0.50 | 0.04 | 0.2380 | 0.89 | 0.00 | 0.4440 | 0.82 | 0.60 | 0.7500 | 0.95 | 0.84 | 0.9000 | 0.98 | 0.95 | 0.9730 | 0.99 | 0.99 | 0.9940 | 1.00 |
| | 0.80 | 0.04 | 0.1270 | 0.48 | 0.00 | 0.5320 | 0.78 | 0.60 | 0.7960 | 0.95 | 0.84 | 0.9210 | 0.98 | 0.95 | 0.9760 | 0.99 | 0.99 | 0.9950 | 1.00 |
| 50 | 0.20 | 0.03 | 0.2560 | 0.88 | 0.00 | 0.3740 | 0.71 | 0.57 | 0.6860 | 0.95 | 0.83 | 0.8610 | 0.96 | 0.95 | 0.9570 | 0.96 | – | – | – |
| | 0.50 | 0.03 | 0.1540 | 0.63 | 0.00 | 0.4710 | 0.71 | 0.57 | 0.7410 | 0.91 | 0.83 | 0.8800 | 0.99 | 0.95 | 0.9600 | 0.99 | – | – | – |
| | 0.80 | – | – | – | 0.15 | 0.5410 | 0.60 | 0.57 | 0.7970 | 0.91 | 0.83 | 0.9200 | 0.99 | 0.95 | 0.9670 | 0.99 | 0.99 | 0.9940 | 1.00 |
| 70 | 0.20 | 0.03 | 0.2060 | 0.78 | 0.00 | 0.3650 | 0.71 | 0.58 | 0.6670 | 0.95 | 0.83 | 0.8540 | 0.94 | – | – | – | – | – | – |
| | 0.50 | 0.03 | 0.1250 | 0.45 | 0.00 | 0.4930 | 0.64 | 0.58 | 0.7350 | 0.96 | 0.83 | 0.8700 | 0.97 | 0.94 | 0.9560 | 0.99 | – | – | – |
| | 0.80 | – | – | – | 0.41 | 0.5390 | 0.57 | 0.59 | 0.8040 | 0.88 | 0.83 | 0.9160 | 0.98 | 0.94 | 0.9650 | 0.99 | 0.99 | 0.9920 | 1.00 |
| 100 | 0.20 | 0.02 | 0.1490 | 0.73 | 0.00 | 0.3710 | 0.65 | 0.57 | 0.6500 | 0.85 | 0.83 | 0.8480 | 0.92 | – | – | – | – | – | – |
| | 0.50 | 0.03 | 0.0850 | 0.22 | 0.00 | 0.5100 | 0.62 | 0.57 | 0.7370 | 0.92 | 0.83 | 0.8630 | 0.96 | 0.94 | 0.9520 | 0.97 | – | – | – |
| | 0.80 | – | – | – | – | – | – | 0.62 | 0.8060 | 0.85 | 0.83 | 0.9190 | 0.97 | 0.94 | 0.9590 | 0.99 | 0.99 | 0.9910 | 1.00 |
| 300 | 0.20 | 0.02 | 0.0748 | 0.39 | 0.00 | 0.3888 | 0.58 | 0.57 | 0.6025 | 0.79 | – | – | – | – | – | – | – | – | – |
| | 0.50 | – | – | – | 0.34 | 0.5420 | 0.58 | 0.57 | 0.7450 | 0.84 | 0.82 | 0.8380 | 0.92 | – | – | – | – | – | – |
| | 0.80 | – | – | – | – | – | – | 0.78 | 0.8150 | 0.83 | 0.82 | 0.9240 | 0.95 | 0.94 | 0.9500 | 0.98 | – | – | – |

Table 5 shows that when there is a poor agreement and the correlation between raters decreases, *ADD* increases. Except for the poor agreement, when the correlation between raters increases, *ADD* also increases. When the agreement increases, the value of *ADD* also increases and converges to 1.

By means of the results in Table 5, it is possible to develop a mixed table of the kappa coefficient and adjusted degree of distinguishability. As the minimum, maximum values, and medians in Table 5 are considered, we suggested the interpretation levels for $ADD$. The results for 2 × 2 tables are summarized in Table 6.

**Table 6.** The interpretation levels of $ADD$

| $\kappa$ | $ADD$ | **Strength of** $ADD$ |
|---|---|---|
| 0.81-1.00 | >0.99 | Perfect |
| 0.61-0.80 | 0.94-0.99 | Substantial |
| 0.41-0.60 | 0.82-0.93 | Moderate |
| 0.21-0.40 | 0.57-0.81 | Fair |
| <0.20 | 0.00-0.56 | Poor |

In order to test the validity of the defined intervals, a simulation study is performed with 50,000 replications for different sample sizes and correlation levels. Then, the kappa coefficient and the adjusted degree of distinguishability are calculated for each replication. The percentages of correct classifications are calculated for each scenario and given in Table 7. The percentages of correct classifications in Table 7 change between 0.73 and 0.97.

**Table 7.** The percentages of correct classifications for different sample size and correlation

| | | $\rho$ | |
|---|---|---|---|
| $n$ | **0.20** | **0.50** | **0.80** |
| 30 | 0.85 | 0.78 | 0.73 |
| 50 | 0.94 | 0.85 | 0.78 |
| 70 | 0.96 | 0.87 | 0.81 |
| 100 | 0.97 | 0.90 | 0.82 |
| 300 | 0.97 | 0.93 | 0.75 |

## 5. Illustrative Examples

In this section, we revisit three examples that will be used to illustrate the kappa coefficient, classical and adjusted degrees of distinguishabilities.

**Example 1:** To illustrate the calculation of kappa coefficient and adjusted degrees of distinguishability, let us consider the $2 \times 2$ contingency tables in Table 8 and 9. The data is taken from Shoukri [16] who examined 197 patients with prostate cancer. A modified TNM (tumor, node, metastasis) staging system is used to categorized MRI (magnetic resonance imaging), ultrasound, and pathological finding.

**Table 8.** Ultrasonography vs. pathological analysis for prostate cancer differentiation

| | **Stage in pathological study** | | |
|---|---|---|---|
| **Stage in ultrasound** | **Advanced** | **Localized** | **Total** |
| Advanced | 45 | 50 | 95 |
| Localized | 60 | 90 | 150 |
| Total | 105 | 140 | 245 |

**Table 9.** MRI vs. pathological analysis for prostate cancer differentiation

| | **Stage in pathological study** | | |
|---|---|---|---|
| **Stage in MRI** | **Advanced** | **Localized** | **Total** |
| Advanced | 51 | 28 | 79 |
| Localized | 30 | 88 | 118 |
| Total | 81 | 116 | 197 |

For Table 8 and 9, kappa coefficients are calculated as 0.07 and 0.39, respectively. The adjusted degree of distinguishabilities are calculated as 0.26 and 0.81.
While it is possible to infer a slight agreement between pathological and ultrasound results, it can be said that the distinguishability of the advance and localized categories is also at a slight level. While it is possible to infer a fair agreement between pathological and MRI results, it can be said that the distinguishability of the advance and localized categories is also at a fair level. Furthermore, when the pathological analysis results accept as reference, it can be said that MRI is more able to distinguish the categories than ultrasound.

**Example 2:** The radiographs of each of 60 patients are shown to two groups of doctors (two trauma surgeons and two radiologists). The data is taken from Oh [13]. To illustrate the calculation of linearly weighted kappa coefficient, classical and adjusted degrees of distinguishabilities, we consider the $4 \times 4$ ordered square contingency table in Table 10.

**Table 10.** The ratings given by Trauma surgeons and radiologists

| | **Radiologists** | | | | |
|---|---|---|---|---|---|
| **Trauma surgeons** | **0** | **1** | **2** | **3** | **Total** |
| **0** | 3 | 15 | 1 | 2 | 21 |
| **1** | 1 | 11 | 13 | 1 | 26 |
| **2** | 1 | 5 | 4 | 2 | 12 |
| **3** | 0 | 0 | 1 | 0 | 1 |
| **Total** | 5 | 31 | 9 | 5 | 60 |

For Table 10, linearly weighted kappa coefficient is calculated as 0.11. The classical and adjusted degrees of distinguishabilities are:

| | 0-1 | 1-2 | 2-3 |
|---|---|---|---|
| $DD$ | 0.42 | -0.43 | -0.67 |
| $ADD$ | 0.42 | 0.30 | 0.40 |

While it is possible to infer a slight agreement between doctors' decisions, it is possible to infer poor distinguishabilities of all the pairs of categories. Furthermore, the distinguishabilities of the adjacent categories are homogenous.

**Example 3 :** 190 patients' slides with advanced and nonadvanced adenomas is identified in a case-control study of adenomatous polyps conducted in NYC colonoscopy-based practices. The slides were classified into 5 categories in 1988 and 1998 by a pathologist. The data taken from Terry *et al* [17] is given in Table 11.

**Table 11.** The ratings of case-control study

| | | | **1998** | | | |
|---|---|---|---|---|---|---|
| **1988** | **1** | **2** | **3** | **4** | **5** | **Total** |
| **1** | 8 | 13 | 4 | 1 | 1 | 27 |
| **2** | 9 | 16 | 12 | 2 | 0 | 39 |
| **3** | 1 | 13 | 8 | 1 | 1 | 24 |
| **4** | 2 | 19 | 12 | 9 | 6 | 48 |
| **5** | 2 | 6 | 11 | 6 | 27 | 52 |
| **Total** | 22 | 67 | 47 | 19 | 35 | 190 |

For Table 11, linearly weighted kappa coefficient is calculated as 0.38. The adjusted degrees of distinguishabilities are:

| | 1-2 | 2-3 | 3-4 | 4-5 |
|---|---|---|---|---|
| $ADD$ | 0.09 | 0.17 | 0.77 | 0.84 |

While it is possible to infer a fair agreement between doctors' decisions, it is possible to infer poor category distinguishability of categories (1) and (2), and categories (2) and (3). While categories (3) and (4) are fairly distinguishable, categories (4) and (5) are moderately distinguishable. Furthermore, the distinguishabilities of the adjacent categories are non-homogenous.

Because there are poor distinguishabilities of categories (1) and (2), and categories (2) and (3), it is suggested to be reclassified the categories. We can reclassify the categories as 3 alternatives. Linearly weighted kappa coefficients and the adjusted degrees of distinguishabilities are calculated for the reclassified tables.

**Alternative 1:** (1+2), (3), (4), (5)

| | $ADD_{1+2,3}$ | $ADD_{3,4}$ | $ADD_{4,5}$ | $\kappa_w$ |
|---|---|---|---|---|
| Estimate | 0.39 | 0.83 | 0.85 | 0.40 |
| Level | Poor | Moderate | Moderate | Fair |

**Alternative 2:** (1), (2+3), (4), (5)

| | $ADD_{1,2+3}$ | $ADD_{2+3,4}$ | $ADD_{4,5}$ | $\kappa_w$ |
|---|---|---|---|---|
| Estimate | 0.57 | 0.79 | 0.85 | 0.40 |
| Level | Fair | Fair | Moderate | Fair |

The linearly weighted kappas increase to 0.40 after the reclassifications 1 and 2.

**Alternative 3:** (1+2+3), (4), (5)

| | $ADD_{1+2+3,4}$ | $ADD_{4,5}$ | $\kappa_w$ |
|---|---|---|---|
| Estimate | 0.83 | 0.85 | 0.49 |
| Level | Moderate | Moderate | Moderate |

For the third alternative, the adjusted degree of distinguishabilities and linearly weighted kappa coefficient increase to moderate levels.

## 6. Conclusions

When working on square contingency tables, firstly the agreement between the row and column variables is investigated. The variables of a square contingency table can be possibly two different raters who rate the same subjects or two different time points which is rated by the same rater. In the agreement studies, it is expected that the decisions of the raters are correspond to each other. If the agreement between raters is not high enough, there could be many reasons. One of these reasons is that the raters cannot distinguish the categories and because of this cannot classify the subjects correctly. Incorrect classification may affect the level of the agreement. In that case, it will be useful to use the degree of distinguishability to detect if the categories are distinguishable or not.

In practice, there occurs some problems to the use of the degree of distinguishability. In this article, we purposed to solve the problem of calculating the degree of distinguishability falls outside the defined range and to discuss the terms agreement and category distinguishability together. We proposed to use adjusted degree of distinguishability instead of the classical one.

The simulation results show that the adjusted degrees of distinguishability increases when the the correlation between raters increases. Besides, when there is medium or high correlation, it is not affected by the sampling size changes. While there is a low correlation and the sample size increases, the adjusted degree of distinguishability decreases.

It is easy to interpret $ADD$ by use of the classification in Table 6. If the distinguishability between the categories is less than “moderate" level, then it is proposed to combine the categories. However, because of the definition of the categories, sometimes it is not logical to combine the categories. In that case, it is reasonable to reclassify the categories and repeat the study.

## References

[1] Agresti, A. *Categorical data analysis* (John Wiley and Sons, New York, 2002).

[2] Becker, M.P. and Agresti, A. *Log-linear modelling of pairwise interobserver agreement on a categorical scale*, Statistics in Medicine **11** (1), 101-114, 1992.

[3] Cicchetti, D. and Allison, T. *A new procedure for assessing reliability of scoring eeg sleep recordings*, American Journal EEG Technology **11**, 101-109, 1971.

[4] Cohen, J. *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement **20** (1), 37-46, 1960.

[5] Cohen, J. *Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit*, Psychological Bulletin **70** (4), 213-220, 1968.

[6] Darroch, J.N. and McCloud, P.I. *Category distinguishability and observer agreement*, Australian Journal of Statistics **28** (3), 371-388, 1986.

[7] Fleiss, J.L. and Cohen, J. *The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability*, Educational and Psychological Measurement **33**, 613-619, 1973.

[8] Goktas, A., Isci, O. *A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation*, Metodoloski Zvezki **8** (1), 17-37, 2011.

[9] Holmquist, N.D., McMahon, C.A., and Williams, O.D. *Variability in classification of carcinoma in situ of the uterine cervix*, Archives of pathology **84**, 334-345, 1967.

[10] Landis, J.R. and Koch, G.G. *The measurement of observed agreement for categorical data*, Biometrics **33** (1), 159-174, 1977a.

[11] Landis, J.R. and Koch, G.G. *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers*, Biometrics **33** (2), 363-374, 1977b.

[12] Lawal, B. *Categorical data analysis with SAS and SPSS applications* (Lawrence Erlbaum Associates, Publishers, Inc., New Jersey, 2003).

[13] Oh, M. *Inference on measurements of agreement using marginal association*, Journal of the Korean Statistical Society **38**, 41-46, 2009.

[14] Perkins, S.M. and Becker, M.P. *Assessing rater agreement using marginal association models*, Statistics in Medicine **21**, 1743-1760, 2002.

[15] Saracbasi, T. *Agreement models for multiraters*, Turkish Journal of Medical Sciences **41** (5), 939-944, 2011.

[16] Shoukri, M.M. *Measures of interrater agreement* (Chapman & Hall/CRC Press LLC., Florida, 2004).

[17] Terry, M.B., Neugut, A.I., Bostick, R.M., Potter, J.D., and Haile, R.W. *Reliability in the classification of advanced colorectal adenomas*, Cancer Epidemiol Biomarkers & Prevention **11**, 660-663, 2002.

[18] Tinsley, H.E.A. and Weiss, D.J. *Interrater reliability and agreement*, in: Handbook of applied multivariate statistics and mathematical modeling (Academic Press, New York, 2010).

[19] Valet, F. and Mary, J.-Y. *Power estimation of tests in log-linear nonuniform association models for ordinal agreement*, BMC Medical Research Methodology **11** (1), 70-80, 2011.