Kappa Statistics: A Method of Measuring Agreement in Dental Examinations

Farzan Madadizadeh¹, Hesam Ghafari² and Sajjad Bahariniya³,*

¹Departments of Biostatistics and Epidemiology, Center for Healthcare Data modeling, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
²School of Dentistry, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
³Department of Health Services Management, School of Public Health, Shahid Sadoughi, Iran

Abstract: Statistical methods have always been the solution to medical problems. Due to the problem of inconsistency in the diagnosis of dentists, statistical science has been provided for measuring the compatibility of diagnosis and reliability of dentists. One of the most important statistical methods for examining the agreement between the two experiments or diagnoses is Kappa statistics which can be used in dental sciences. The present study reviewed different types of Kappa statistics for assessing agreement, including Cohen's kappa, Fleiss' kappa, and Cohen's weighted kappa.

Keywords: Kappa statistics, Dental examinations, Compatibility, Cohen's kappa, Fleiss' kappa, Cohen's weighted kappa.

Article History
Received: May 09, 2023
Revised: July 04, 2023
Accepted: July 13, 2023

1. INTRODUCTION

Dentistry is one of the sciences in that direct observation of the doctor and his personal opinion during the examination has a direct effect on the final diagnoses [1]. Due to the multiplicity of dentists, the Decayed Missing Filled (DMF) Index was introduced to integrate diagnoses and standardize diagnostic criteria in the field of caries [2]. However, later it became clear that the accuracy of the index is not a guarantee for the consistency of dentists' diagnosis. Even a dentist can make a heterogeneous diagnosis on two examinations. Factors such as prior knowledge, experience, and aptitude directly affect dentists' diagnoses.

Statistical methods can always be used to solve medical problems. Considering the fact that there is inconsistency in dental diagnosis, which both the World Health Organization and the International Dental Federation are concerned about, statistics is an indicator for measuring the consistency of diagnoses in dental examinations [3]. One of the most widely used statistical methods to check the agreement between two diagnoses used in dental sciences is the Kappa statistic. To the best of our knowledge, there were no tutorial studies that mention all types of Kappa statistics and applications in dentistry. So this tutorial study aimed to introduce all types of Kappa statistics and their application in dental science in a simple way for dummies. Different types of Kappa and their applications for measuring agreement in dentistry are presented with examples in the following sections.

2. DIFFERENT TYPES OF KAPPA STATISTICS

2.1. Cohen's Kappa

Cohen's kappa (CK) was introduced by Jacob Cohen in 1960 and is often used to assess concordance between two raters [4]. CK is a statistic that is used to measure inter-rater reliability for qualitative items. Also, it is generally thought to be a more robust measure than a simple percent agreement calculation [5, 6]. In dentistry, this index can be defined as follows:

Suppose two dentists examine the same tooth, the rate of concordance in their diagnoses can be measured in the form of a statistical index called kappa. It requires the final diagnosis of each dentist to be a dichotomous variable, such as a healthy or decayed tooth. Accordingly, a table with two rows for the first diagnosis and two columns for the second diagnosis is made as shown in Table 1.

After creating a 2 x 2 table, Cohen's kappa (CK) uses the numbers in the table to evaluate the agreement of two dentists in examining patients and diagnosing decayed teeth [7]. To define this statistic, we first need to introduce the observed proportion of agreement and expected probability (Table 1).

2.1.1. Raw Agreement (P0)

In Table 1, agreement between the two dentists is
The Open Public Health Journal, 2023, Volume 16

Madadizadeh et al.

represented in cells A and D, so calculating the raw agreement can be done as follows: (Eq 1).

\[ P_0 = \frac{A+D}{N} \]  \hspace{1cm} (1)

2.1.2. The Expected Agreement ( \hat{P} )

To calculate the expected agreement ( \hat{P} ), in each row and column, the sum of rows and columns is calculated, then the sum of the first row is multiplied by the sum of the first column and the sum of the second column is also multiplied by the sum of the second row. The resulting products are added together and divided by the square of the sample size. Below is the formula for calculating the expected agreement and Cohen's kappa: (Eq 2)

\[ \text{Expected Agreement} \hat{P} = \frac{[(A+B) \times (A+C)] + [(C+D) \times (B+D)]}{N^2} \]  \hspace{1cm} (2)

\( N = \text{total sample size} \)

Now, one can define kappa statistics using the formulae presented in sections I and II as follows: (Eq 3).

\[ \text{Kappa statistics} = \frac{\text{Raw agreement (P0)} - \text{expected agreement (\hat{P})}}{1 - \text{expected agreement (\hat{P})}} \]  \hspace{1cm} (3)

\[ \text{Example 1:} \] Suppose two dentists examine 100 teeth (N=100) at the end of a working day so both of them diagnose 40 decayed and 30 healthy teeth. Then, the first dentist will diagnose 20 cases of them as rotten, while the second dentist will diagnose them as healthy. Also if dentist 1 diagnoses 10 cases as healthy while the second dentist has diagnosed them as decayed at the same time, the data are given in the 2 × 2 table as follows (Table 2).

Table 1. Summary of diagnosis of two dentists

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Healthy Tooth</th>
<th>Decayed Tooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy Tooth</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Decayed Tooth</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

The range of values for Kappa statistics is between -1 to 1 [12, 13]. If the values are less than 0, it indicates no agreement, values between 0.6 and 0.8 indicate moderate agreement and values greater than 0.8 mean nearly complete agreement [9]. A generalized form of Cohen's kappa statistic for more than two raters was introduced by Fleiss in 1970, known as Fleiss' kappa [7, 13].

\[ \text{Fleiss' kappa} \] was introduced by Fleiss et al. between 1970 and 2003 [14]. This index is used to check the agreement of the results diagnosed by two or more evaluators. It is necessary for the diagnosis result to be a qualitative variable (good or bad, sick or healthy, normal or moderate or severe) [15, 16].

In calculating Fleiss' kappa, one must first create a table where each row of the table corresponds to a patient and the columns represent the possible categorical scores so that there is one column for each score. The values in the cells of the table reflect the number of raters (dentists in the previous example) who chose that score (n_i). Then a probability is calculated for the row and column according to the following relations. Row probabilities for each patient are calculated from (Eq. 4) and column probabilities are calculated from equation (Eq. 6). Then, the raw agreement probability is calculated based on raw probabilities (P_i) as (Eq. 5). The amount of expected agreement is also calculated through relation. Finally, like the previous formula that we had in Cohen's kappa, the calculations are obtained (Table 3).

Fleiss' kappa is not a multi-rater extension of Cohen's kappa [17].

Table 2. Results of diagnosis of two dentists

<table>
<thead>
<tr>
<th>Patient 1</th>
<th>Healthy Tooth</th>
<th>Decayed Tooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy Tooth</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>Decayed Tooth</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>

For our example in Table 2, the raw agreement in diagnosis between the two dentists is 0.7.

\[ P_0 = \frac{40+30}{100} = 0.7 \]

So, for our example in Table 2, its value is 0.46.

\[ \hat{P} = \frac{[50 \times 60] + [40 \times 40]}{100^2} = 0.46 \]

Therefore, in our example in Table 2, the kappa statistic is equal to 0.44.

\[ (\text{Kappa statistics} = 0.7 - 0.46 = 0.44) \]

It shows the agreement between to dentist is low.

2.2. Fleiss' Kappa

Fleiss' kappa was introduced by Fleiss et al. between 1970 and 2003 [14]. This index is used to check the agreement of the results diagnosed by two or more evaluators. It is necessary for the diagnosis result to be a qualitative variable (good or bad, sick or healthy, normal or moderate or severe) [15, 16].

In calculating Fleiss' kappa, one must first create a table where each row of the table corresponds to a patient and the columns represent the possible categorical scores so that there is one column for each score. The values in the cells of the table reflect the number of raters (dentists in the previous example) who chose that score (n_i). Then a probability is calculated for the row and column according to the following relations. Row probabilities for each patient are calculated from (Eq. 4) and column probabilities are calculated from equation (Eq. 6). Then, the raw agreement probability is calculated based on raw probabilities (P_i) as (Eq. 5). The amount of expected agreement is also calculated through relation. Finally, like the previous formula that we had in Cohen's kappa, the calculations are obtained (Table 3).

Fleiss' kappa is not a multi-rater extension of Cohen's kappa [17].

Table 3. Formulation of fleiss' kappa.

<table>
<thead>
<tr>
<th>Patient (N)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>( P_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>n1</td>
<td>n12</td>
<td>n13</td>
<td>n14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Patient 2</td>
<td>n21</td>
<td>n22</td>
<td>n23</td>
<td>n24</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mij</td>
<td>-</td>
</tr>
<tr>
<td>( P_j )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

\[ \text{Raw Probability} \ P_i = \frac{1}{n(n-1)} \left( \sum_{j=1}^{n} n_j^2 - n \right) \]  \hspace{1cm} (4)

\[ \text{Raw Agreement Probability} \ P_{raw} = \text{mean}(P_i) \]  \hspace{1cm} (5)

\[ \text{Column probability} \ P_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \]  \hspace{1cm} (6)

\[ \text{The expected agreement} \ \hat{P} = \sum_{j=1}^{K} P_j^2 \]  \hspace{1cm} (7)

\[ \text{Fleiss' kappa} = \frac{\text{Raw Agreement Probability} - \text{Expected Agreement Probability}}{1 - \hat{P}} \]  \hspace{1cm} (8)
Where \( n \) = number of raters; \( j \) = range of scores \((j=1,\ldots,k)\); \( i \) = number of patients \((i=1,\ldots,N)\).

**Example 2:** Fourteen dentists give grades from 1 to 5 about the severity of a tooth's damage. If 5 patients are examined, the statistical value of Fleiss' kappa is calculated below (Table 4).

For our example in Table 3, \( P_1 \) is 0.157.

\[
P_1 = \frac{0 + 0 + 2 + 3 + 6}{14 \times 5} = 0.157
\]

And taking the second row,

\[
P_2 = \frac{1}{14 (14-1)} \left( (0^2 + 0^2 + 3^2 + 5^2 + 6^2 - 14) \right) = 0.302
\]

In order to calculate \( P_k \), we need to know the sum of \( P_i \).

\[
\sum_{i=1}^{N} P_i = 1.000 + 0.302 + 0.324 + 0.237 + 0.280 = 2.143
\]

Over the whole sheet,

\[
\bar{P} = \frac{1}{5} \times 2.143 = 0.428
\]

\[
\hat{P} 0.157^2 + 0.128^2 + 0.271^2 + 0.142^2 + 0.306^2 = 0.223
\]

Fleiss' kappa

\[
\text{Fleiss' kappa} = \frac{0.428 - 0.223}{1 - 0.223} = 0.263
\]

It shows a weak agreement.

### Table 4. Values for computing of the Fleiss' kappa

<table>
<thead>
<tr>
<th>-</th>
<th>Categorical Scoring</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Patient 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient 3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Patient 4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Patient 5</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

2.3. Cohen's Weighted Kappa

The weighted kappa penalizes disagreements in terms of their seriousness whereas the unweighted kappa treats all differences equally. Consequently, Cohen's weighted kappa ought to be employed when data is in the form of a graded ordinal scale. In this situation, three matrices such as observed score matrix, the expected score matrix based on agreement, and the weight matrix are involved. Weight matrix cells located on the diagonal indicate agreement and off-diagonal cells indicate disagreement. The equation for weighted kappa is: (Eq 9).

\[
\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}} \tag{9}
\]

where \( k \) = number of codes and \( W_{ij}, X_{ij} \) and \( M_{ij} \) are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above (Figs. 1 and 2).
Table 5. Interpretation of different values of all types of kappa statistics.

<table>
<thead>
<tr>
<th>Kappa Value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>NO agreement</td>
</tr>
<tr>
<td>0.01 – 0.20</td>
<td>Slight agreement</td>
</tr>
<tr>
<td>0.21 – 0.40</td>
<td>Fair agreement</td>
</tr>
<tr>
<td>0.41 – 0.60</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.61 – 0.80</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>0.81 – 1.00</td>
<td>Almost perfect agreement</td>
</tr>
</tbody>
</table>

Table 6. Summary of different Types of Kappa statistics.

<table>
<thead>
<tr>
<th>Types of Kappa</th>
<th>Calculation Formulas</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s kappa</td>
<td>( k = \frac{P_{\text{raw}} - \hat{P}}{1 - \hat{P}} )</td>
<td>Determine the level of agreement among two independent raters</td>
</tr>
<tr>
<td></td>
<td>- ratings are based solely on nominal scales</td>
<td>- lacks the ability to measure inter-rater reliability with more</td>
</tr>
<tr>
<td></td>
<td>- it does not consider or account for the inherent order or</td>
<td>than two independent raters. In such cases, a suitable alternative</td>
</tr>
<tr>
<td></td>
<td>hierarchy of categories within an ordinal variable.</td>
<td>is to utilize Krippendorff’s alpha.</td>
</tr>
<tr>
<td>Fleiss’ kappa</td>
<td>( \hat{k} = \frac{P_{\text{raw}} - \hat{P}}{1 - \hat{P}} )</td>
<td>Determine the level of agreement among multiple independent</td>
</tr>
<tr>
<td></td>
<td>- it does not consider or account for the inherent order or</td>
<td>raters, it can be used with binary or nominal-scale or ordinal or</td>
</tr>
<tr>
<td></td>
<td>hierarchy of categories within an ordinal variable.</td>
<td>ranked data</td>
</tr>
<tr>
<td></td>
<td>- it requires the random selection of patients and raters.</td>
<td></td>
</tr>
<tr>
<td>Cohen’s weighted</td>
<td>( \kappa = 1 - \frac{\sum_{i=1}^{h} \sum_{j=1}^{k} w_{ij} \hat{p}<em>{ij}}{\sum</em>{i=1}^{h} \sum_{j=1}^{k} w_{ij} \hat{p}_{ij}} )</td>
<td>Determine the level of agreement among two independent raters</td>
</tr>
<tr>
<td>kappa</td>
<td>- ratings are based on ordinal scales</td>
<td>- it is used to compare the agreement among two independent raters</td>
</tr>
</tbody>
</table>

Landis and Koch (1977) gave the following table for interpreting Kappa values [18 - 20] (Table 5).

In short, the types of Kappa calculation methods are presented in (Table 6). Using this table, researchers will be able to choose the type of kappa suitable for their research (Table 6).
Cohen's weighted kappa:

The way to perform weighted kappa is the same as Cohen's kappa, only before going to the above analysis, weighting should be done in the following way:

Data $\rightarrow$ weight cases
Fleiss' kappa: Analyze $\rightarrow$ scales $\rightarrow$ reliability analysis

R:

• Cohen's / Cohen's weighted kappa: the function 'kappa2' from the package 'irr',

• Other functions:
  Cohen's unweight kappa: kappa Cohen (data, weight="unweighted")
  Cohen's weighted kappa: kappa Cohen (data, weight="weighted")
Fleiss' kappa: function 'kappa m. fleiss' from the package 'irr'

SATA:
  Cohen's unweight kappa: cap rater1 rater2, tab
  Cohen's weighted kappa: cap rater1 rater2 [fweight=wvar], tab
Fleiss' kappa: cap rater1 rater2 rater3

CONCLUSION

Cohen's kappa coefficient ($\kappa$) is a statistical measure utilized to assess the reliability between raters for qualitative items, both in terms of inter-rater and intra-rater reliability. In cross-classification, Cohen's weighted kappa is commonly employed as an agreement measure among raters (dentists). It serves as a suitable indicator for agreement when ratings are based solely on nominal scales without any inherent order structure. The original method of calculating Cohen's kappa lacks the ability to measure inter-rater reliability with more than two raters. In such cases, a suitable alternative is to utilize Krippendorff's alpha.

On the other hand, Cohen's weighted kappa is preferable when dealing with categorical data that possess an ordinal structure, such as rating systems with categories like high, medium, or low presence of a specific attribute. This weighted kappa places more emphasis on the seriousness of disagreements, while the unweighted kappa treats all disagreements as equally important. Therefore, the unweighted kappa is not suitable for ordinal scales.

Fleiss' kappa ($\kappa$), is another measure of inter-rater agreement, specifically employed to determine the level of agreement among multiple raters, it can be used with binary or nominal-scale data and can also be applied to ordinal or ranked data. However, in cases involving ordinal ratings, such as defect severity ratings on a scale of 1–5, Kendall's coefficients, which account for the ordering of categories, are usually more appropriate for determining association than kappa alone. Fleiss' kappa does not consider or account for the inherent order or hierarchy of categories within an ordinal variable.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[16] Chicco D, Warrens MJ, Jurman G. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier


