Kappa statistics a method of measuring agreement in dental examinations

Farzan Madadizadeh  
Shahid Sadoughi University of Medical Sciences and Health Services

Hesam Ghafari  
Shahid Sadoughi University of Medical Sciences and Health Services

Sajjad Bahariniya (✉ sajjadbahari98@gmail.com)  
Shahid Sadoughi University of Medical Sciences and Health Services

Research Article

Keywords:

Posted Date: February 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2535291/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Statistical methods have always been the solution to medical problems. Due to the problem of inconsistency in the diagnosis of dentists, the 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 diagnosis is Kapa statistics that can be used in dental sciences. The present study examined different type of Kappa statistics for assessing agreement, including Cohen's kappa, Fleiss' kappa and Cohen's weighted kappa.

Introduction

Dentistry is one of the sciences that direct observation of the doctor and his personal opinion during the examination has a direct effect on the final diagnosis (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). But 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' diagnosis (3).

Statistical methods have always been the solution to medical problems. Regarding the problem of inconsistencies in the diagnosis of dentists, which both the World Health Organization and the Federation Dentaire Internationale are concerned about, statistical science is presented an indicator to measure the compatibility of diagnoses and the reliability of examining dentists (4). One of the most important statistical methods for checking the agreement between two tests or diagnosis is the Kappa statistic, which can be used in dental science. To the best of our knowledge, there were no tutorial study which mention all type of Kappa statistics and application in dentistry. So this tutorial study aimed to introduce all type of Kappa statistics and its application in dental science in simple way for dummies. Different types of Kappa and its applications for measuring agreement in dentistry are presented with examples in the following sections.

Different type of Kappa statistics for assessing agreement

1- Cohen's kappa

Cohen's kappa (CK) is a metric introduced by Jacob Cohen in 1960, it often used to assess the agreement between two raters. 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 simple percent agreement calculation (5, 6). In dentistry, this index can be defined as follows:

Suppose two dentists examine the same tooth, the degree of agreement in their diagnosis can be measured in the form of a statistical index called kappa. Its need to consider the final diagnosis of each dentist to be a dichotomous variable, such as healthy or decayed tooth). Accordingly, a table with two
rows for the first diagnosis and two columns for the second diagnosis is made as a 2 × 2 table such as one represented in Table 1.

Table 1
Summary of diagnosis of two dentists

<table>
<thead>
<tr>
<th></th>
<th>The diagnosis of dentist 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy Tooth</td>
</tr>
<tr>
<td>The diagnosis of dentist 1</td>
<td>Healthy Tooth</td>
</tr>
<tr>
<td></td>
<td>Decayed Tooth</td>
</tr>
</tbody>
</table>

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

I- Raw agreement (P0)

In Table 1, two dentists were agreeing in cells A and D, so calculating the raw agreement can be done as follows:

\[ P_0 = \frac{A+D}{N} \quad \text{(Eq. 1) (8)} \]

II- The expected agreement (\( \hat{P} \))

To calculate of 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 answers of the above multiplications are added together and divided by the square of the sample volume. Below is the formula for calculating Cohen's kappa:

\[ \hat{P} = \frac{[(A+B) \times (A+C)] + [(C+D) \times (B+D)]}{N^2} \quad \text{(Eq. 2) (9)} \]

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

Now, we can define kappa statistics using the formulas which were presented in section I and II as follows:

\[ \text{Kappa statistics} = \frac{P_0 - \hat{P}}{1 - \hat{P}} \quad \text{(Eq. 3) (7, 9-11)} \]
The range of values for this statistic for two dentists (two evaluators) is between −1 to 1 (12, 13). Values less than zero indicate disagreement and values between 0.6 and 0.8 indicate moderate agreement and values greater than 0.8 called the perfect agreement (4, 9). A generalized form of Cohen’s kappa statistic for more than two rater (dentist in our example) was introduced by Fleiss’ in 1970 and is known as Fleiss’ kappa (7, 13).

**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>
<thead>
<tr>
<th>The diagnosis of dentist 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Healthy Tooth</td>
</tr>
<tr>
<td>The diagnosis of dentist 1</td>
<td>Healthy Tooth</td>
</tr>
<tr>
<td></td>
<td>Decayed Tooth</td>
</tr>
</tbody>
</table>

For our example in Table 2 it value is 0.7. \( P_0 = \frac{40 + 30}{100} = 0.7 \)

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

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

So, in our example in Table 2 the kappa statistics is equal to 0.44.

\[ \text{Kappa statistics} = \frac{0.7 - 0.46}{1 - 0.46} = 0.44 \] (that shows the agreement between to dentist is low.

2- **Fleiss' kappa**

Fleiss' kappa was introduced by Joseph L. Fleiss et al. during the years 1970 to 2003. 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) (14, 15). Compared to Cohen's kappa, Fleiss' kappa works for any number of raters, Cohen's kappa only works for two raters; in addition, Fleiss' kappa allows for each rater to be rating different items, while Cohen's kappa assumes that both raters are rating identical items.

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

<table>
<thead>
<tr>
<th>Range of categorical score (K)</th>
<th>Patients (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>Patients (N)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>...</td>
<td></td>
<td>(P_i)</td>
</tr>
<tr>
<td>Patient 1</td>
<td>n11</td>
<td>n12</td>
<td>n13</td>
<td>n14</td>
<td></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>
<td></td>
</tr>
<tr>
<td>....</td>
<td>nij</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(P_j)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
P_i = \frac{1}{n(n-1)} \left[ \left( \sum_{j=1}^{K} n_{ij}^2 \right) - n \right]
\]

\[
P_{Raw} = \text{mean}(P_i)
\]

\[
P_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}
\]

\[
\hat{P} = \sum_{j=1}^{K} P_j^2
\]

\[
Kappa = \frac{P_{Raw} - \hat{P}}{1 - \hat{P}}
\]

Where \(n\) = number of raters; \(j\) = range of scores \((j = 1,..,k)\); \(i\) = number of patients \((i = 1,..,N)\).

**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 as bellow (Table 4).:
Table 4
Values for computing of the Fleiss’ kappa

<table>
<thead>
<tr>
<th>Categorical scoring</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Pᵢ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>1.000</td>
</tr>
<tr>
<td>Patient 2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>0.302</td>
</tr>
<tr>
<td>Patient 3</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0.324</td>
</tr>
<tr>
<td>Patient 4</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0.237</td>
</tr>
<tr>
<td>Patient 5</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.280</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>9</td>
<td>19</td>
<td>10</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

For our example in Table 3, P₁ is 0.157. 
\[ P₁ = \frac{0+0+2+3+6}{14 \times 5} = 0.157 \]

And taking the second row, 
\[ P₂ = \frac{1}{14(14-1)} \left( 0² + 0² + 3² + 5² + 6² - 14 \right) = 0.302 \]

In order to calculate \( \hat{P} \), 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, 
\[ = \frac{1}{5} \times 2.143 = 0.428 \]

\[ \hat{P} = 0.157² + 0.128² + 0.271² + 0.142² + 0.300² = 0.223 \]

Kappa = \[ \frac{0.428-0.223}{1-0.223} = 0.263 \]

3- Cohen’s weighted kappa

The weighted kappa allows disagreements to be weighted differently and is especially useful when rates are not ordered (nominal scales with no order structure). In this situation, three matrices are involved, such as the observed scores matrix, expected scores matrix based on agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent agreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. Often, cells one off the diagonal are weighted 1, those two off 2, etc. The equation for weighted kappa is:
\[
\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}
\]

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.

Landis and Koch (1977) gave the following table for interpreting \( k \) values \(^{(16, 17)}\) (Table 5).

<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>

**Conclusion**

In a nutshell, standard criteria for the diagnosis of dental complications are provided in dentistry, but the opinion of the examining dentist still has a direct impact on the final diagnosis. One of the statistical methods that can be used to measure the agreement in diagnosing the disease is kappa statistic. If the number of examiner is two, Cohen's kappa statistic and if there are more than two examiners, Fleiss' kappa statistic is used. For the special case that dentists determine the intensity of a tooth's pain and we want to measure the agreement between their opinions, another type of kappa statistic called weighted kappa is used. Therefore, diagnosis and application of kappa statistics in dental research is very necessary and useful.

**Declarations**

**ETHICS AND DISSEMINATION**

**Ethics approval and consent to participate:** Not applicable

**Consent for publication:** Not applicable
Availability of data and materials: All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing interests: None declared.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions: S.B. and F.M. conceived of the presented idea. S.B. wrote the manuscript with support from F.M. and H.G. All authors read the manuscript and verified it.

Acknowledgements: We hereby thank and appreciate all the colleagues who helped us in conducting this research.

References


