1 Introduction

Data where two raters evaluate the same phenomenon are common in clinical research. When ratings are categorical ratings Cohen's kappa coefficient is the standard method for evaluating agreement. The $\kappa$ coefficient is defined for nominal categorical variables, but extensions to weighted coefficients have also been proposed. Let $i$ and $j$ denote two raters and assume that the possible ratings are $\{0,1,\ldots,K\}$.

Let $X_i$ and $X_j$ be stochastic variables that denote the ratings given by rater $i$ and rater $j$, respectively and let

\[
(N_{ij}) = \begin{bmatrix}
N_{00} & N_{01} & \cdots & N_{0K} \\
N_{10} & N_{11} & \cdots & N_{1K} \\
\vdots & \vdots & \ddots & \vdots \\
N_{K0} & N_{K1} & \cdots & N_{KK}
\end{bmatrix}
\]

denote the contingency table of observed responses and let

\[
(p_{ij}) = \begin{bmatrix}
p_{00} & p_{01} & \cdots & p_{0K} \\
p_{10} & p_{11} & \cdots & p_{1K} \\
\vdots & \vdots & \ddots & \vdots \\
p_{K0} & p_{K1} & \cdots & p_{KK}
\end{bmatrix}
\]
denote the probabilities.

2 Cohens $\kappa$ coefficient

In the original formulation (for nominal categorical ratings) the diagonal corresponds to agreement ($A$):

$$\begin{bmatrix}
A & \cdots \\
\cdots & \ddots & \cdots \\
& & & A
\end{bmatrix}$$

(1)

Let $\Pr(a)$ denote the observed proportion of agreement, and let

$$\Pr(e) = \sum_k \Pr(X_i = k) \Pr(X_j = k)$$

denote the expected proportion of agreement under independence. The $\kappa$ coefficient is given by

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

(2)

this coefficient is widely used. The following features follow from the definition

(i) the value depends on the margins and thus on the sample
(ii) if ratings are ordinal this is not taken into account
(iii) there is no rational for saying that, e.g., $\kappa > 0.7$ is good.

2.1 Example

In a $2 \times 2$-table (the case $K = 1$) the $\kappa$ coefficient is given by

$$\kappa = \frac{[p_{00} + p_{11}] - [p_{0,0}p_{0,1} + p_{1,0}p_{1,1}]}{1 - [p_{0,0}p_{0,1} + p_{1,0}p_{1,1}]}.$$  

The three tables shown in Table 1 all feature agreement about $\frac{70}{100}$ of the subjects, but illustrate that $\kappa$ depends on the margins:
Tabel 1: Three tables with the same proportion of agreement. The value of the \( \kappa \) coefficient depends on margins.

\[
\begin{array}{c|cc}
X_j & 0 & 1 \\
\hline
X_i & 0 & 20 & 20 \\
& 1 & 10 & 50 \\
\kappa & 0.35 \\
\end{array}
\quad
\begin{array}{c|cc}
X_j & 0 & 1 \\
\hline
X_i & 0 & 10 & 20 \\
& 1 & 10 & 60 \\
\kappa & 0.21 \\
\end{array}
\quad
\begin{array}{c|cc}
X_j & 0 & 1 \\
\hline
X_i & 0 & 5 & 20 \\
& 1 & 10 & 65 \\
\kappa & 0.08 \\
\end{array}
\]

3 Weighted \( \kappa \) coefficients

The structure imposed by (1) may seem unreasonable for ordinal ratings and weighted \( \kappa \) coefficients are a remedy for this. For this a matrix of weights

\[
\begin{pmatrix}
\begin{array}{cccc}
    w_{00} & w_{01} & \cdots & w_{0K} \\
    w_{10} & w_{11} & \cdots & w_{1K} \\
    \vdots & \vdots & \ddots & \vdots \\
    w_{K0} & w_{K1} & \cdots & w_{KK} \\
\end{array}
\end{pmatrix}
\]

is specified and the weighted \( \kappa \) coefficient is given by

\[
\kappa = \frac{P_O - P_E}{1 - P_E} \quad (3)
\]

where \( P_O = \sum_i \sum_j w_{ij} p_{ij} \) is the observed agreement and \( P_E = \sum_i \sum_j w_{ij} p_i. p_j \) is the expected agreement under independence. Two standard choices of weights have been proposed: the Cichetti-Anderson weights

\[
w_{ij} = 1 - \frac{|i - j|}{K - 1}
\]

and the Fleiss-Cohen weights

\[
w_{ij} = 1 - \frac{(i - j)^2}{K - 1}
\]

the unweighted \( \kappa \) (2) of course corresponds to \( w_{ij} = 1_{(i=j)} \). User-defined weights, e.g.

\[
w_{ij} = \begin{bmatrix}
1.00 & 0.50 & 0.25 & 0 & 0 & 0 \\
0.50 & 1 & 0.50 & 0 & 0 & 0 \\
0.25 & 0.50 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0.75 & 0.50 \\
0 & 0 & 0 & 0.75 & 1 & 0.75 \\
0 & 0 & 0 & 0.50 & 0.75 & 1
\end{bmatrix}
\]
4 Computation in SAS

For most applications it is very easy to compute the $\kappa$ coefficient in SAS:

```sas
data rater;
  input rater1 rater2 count @@;
datalines;
  low low 10 low med 5 low high 1
  med low 5 med med 16 med high 3
  high low 8 high med 12 high high 28
;run;
proc freq data=rater;
  weight count;
  tables rater1*rater2 / agree norow nocol;
run;
```

If all frequencies of any row or any column of the table are 0, SAS will not calculate the $\kappa$ coefficient. In this case, the macro adds a data step that change the zero frequencies to negligible non-zero values:

```sas
data rater;
  set rater;
  if count=0 then count=0.0000000001;
run;
```

5 Example

We consider a data set with ratings of muscle tone for 31 patients with spinal cord injuries. The patients were rated using the modified Ashworth Scale an ordinal scale from 0 (no increase in muscle tone) to 5 (rigid joint). The code

```sas
proc freq data=ashworth;
  table rating1*rating2 / norow nocol nopercent;
run;
```

yields the contingency table
where one row (rating 2='5') has all zeros. This means that SAS will not recognize this as a square table and for this reason not compute the $\kappa$ coefficient. The macro is called using the statement

```sas
%weightedkappa( data=ashworth,
    k=5,
    rater1=rating1,
    rater2=rating2);
```

where $k=5$ indicates that ratings take the values \{0, 1, 2, 3, 4, 5\}. The macro generates an output data set `work.kappa` that is also printed in the output:

Weighted and unweighted kappa coefficients: data set ashworth

<table>
<thead>
<tr>
<th>type</th>
<th>kappa</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN</td>
<td>0.32117</td>
<td>0.08078</td>
<td>0.56156</td>
</tr>
<tr>
<td>CA</td>
<td>0.56338</td>
<td>0.35657</td>
<td>0.77019</td>
</tr>
<tr>
<td>FC</td>
<td>0.74590</td>
<td>0.56322</td>
<td>0.92858</td>
</tr>
</tbody>
</table>

the data set contains the unweighted (‘UN’) kappa coefficient and two weighted kappa coefficients using Fleiss-Cohen (‘FC’) and Cichetti-Anderson (‘CA’) weights.

6 References