Generating Kappa Statistics and Testing Useful Hypotheses with PROC CATMOD

Robert A. Terry, The University of North Carolina at Chapel Hill

Abstract

This paper demonstrates how PROC CATMOD may be used in the construction of generalized kappa statistics. Furthermore, it is shown how relevant hypotheses concerning interrater agreement may be tested using PROC CATMOD, within the conceptual framework of the GSK method. A general form of the response function transformation matrix, \( F \), is given. Hypotheses considered concern overall and category specific measures of agreement and the equality of kappa values in multiple populations. An example is given to illustrate the flexibility of the CATMOD procedure.

Introduction

A standard measure of interrater agreement for categorical data used in psychological studies is the kappa statistic (Cohen, 1960). Since kappa has an interpretation as an intraclass correlation coefficient, kappa has proved a useful tool for determining interrater reliability. Cohen (1960) generalized kappa to a partial-credit model that quantifies the relative seriousness of the disagreements. This weighted kappa statistic has usefulness when pure agreement is impractical to expect, and when near-misses can be considered to be agreements.

Kappa statistics prove to be useful for other purposes as well. For example, kappa can be used as a measure of similarity between two or more objects categorized on dichotomous items. Another potential use is as a measure of concordance between methods of classification or clustering. Finally, kappa may be used as an index for selecting optimal test items for constructing mastery tests.

The usage of kappa, especially weighted kappa, has been somewhat curtailed by the lack of available computer packages which can compute kappa. Those packages that can compute kappa typically do not have a provision for handling weighted kappa, nor do they provide a method for testing hypotheses using the resulting kappa estimates.

This paper uses the general methodology of Grizzle, Starmer, and Koch (1969), hereafter abbreviated GSK, to construct and test various hypotheses concerning kappa-type measures. Landis and Koch (1977) demonstrated that the GSK method could be used as a conceptual framework for the consideration of rater agreement for categorical data. The computational tool which now facilitates this usage is the Version 5 SASI procedure, PROC CATMOD. This paper outlines the use of PROC CATMOD in testing various hypotheses concerning kappa-type statistics. Examples will be presented which demonstrate the implementation of the procedures.

Models

The kappa statistic incorporates the level of chance agreement into the assessment of interrater agreement. For example, if two raters employ completely different and independent criteria for distinguishing between the occurrence or non-occurrence of an event, then all agreement is due to chance. Kappa considers the level of chance agreement in its formulation.

Let \( p_{ij} \) be elements of a probability matrix \( P \) associated with the \( I \times J \) contingency table \( V \), with \( I=J \), where the rows correspond to the observations of rater \( 1 \), and the columns correspond to rater \( 2 \). Let \( P_e \) denote the observed agreement between the raters, where \( P_e \) is the sum of the diagonal elements of \( P \).

Let \( P_r \) denote the expected cell probabilities based upon the observed category usage of the two raters. \( P_e \) may be computed as in the standard \( \chi^2 \) test of multiplicative independence. Then, kappa is defined as

\[
(1) \quad k = \frac{P_e - P_r}{1 - P_r},
\]

which is the ratio of excess agreement beyond chance to maximum possible excess agreement beyond chance.

Kappa has desirable properties in that the range of kappa lies in the interval \([0,1]\) under most conditions, with \( k = 1 \) indicating complete agreement. Landis and Koch (1977) give suggestions as to what particular values of kappa indicate about the level of rater agreement. Basically, values greater than .75 indicate excellent agreement, whereas values less than .40 represent poor agreement.

Weighted kappa is a simple generalization from unweighted kappa. Suppose a set of agreement weights, \( w_{ij} \), can be assigned on rational (theoretical) grounds. Let \( k_w \) range in the interval \([0,1]\), where a weight of \( 1 \) indicates a complete "hit", \( 0 \) indicates a complete "miss", and all other values indicate some partial-credit. Then, weighted kappa can be written as

\[
(3) \quad k_w = \frac{E_1 E_j w_{1j} + P_1 - E_1 E_j w_{ij} - P_j}{1 - E_1 E_j w_{ij} + P_1 + P_j},
\]

where \( p_{1*} \), \( p_{*j} \) indicate row and column sums. Note that unweighted kappa is simply a special case of the weighted version, where the weight matrix \( W \) is an identity matrix.

To calculate weighted kappa, then, is to simply compute the quantities \( P_{1*} \) and \( P_{*j} \), which are functions of the cell probabilities. The use of the GSK method with PROC CATMOD will calculate the population parameter estimate of kappa.
as well as provide a conceptual framework for testing hypotheses about the parameters.

**GSK Method**

Let the elements of a \((1 \times 1)\) column vector \(p\) consist of the observed cell probabilities of \(P\). We assume that the observed cell frequencies of the contingency table \(Y\) generating \(P\) follow a product multinomial distribution with corresponding parameter matrix \(w\). GSK methods allow for transformation of \(p\) into the transformed \((u \times 1)\) column vector \(f = f(p)\), where \(f\) is called the response function. Under certain asymptotic conditions, we can then assume that the covariance matrix of \(f\) is consistently estimated (via a linearized Taylor series approximation) by the \((u \times u)\) matrix

\[ a_u = N_u w H_u H_u' \]

where \(H = \{df(x)/dx \mid x = p\}\).

We then assume that the vector of transformed true probabilities, \(f(x)\), can be represented by a linear model

\[ f(x) = f(a) = xa \]

where \(E\) denotes asymptotic expectation, \(X\) is the \((u \times t, t \times u)\) design matrix for the specification of effects, and \(a\) is a \((t \times 1)\) vector of unknown parameters to be estimated. Once the parameters have been estimated through the method of weighted least squares, hypotheses of the form

\[ H_0: Ca = 0 \]

can be constructed and tested with the \(Q_e\) statistic

\[ Q_e = b' C' (C w^{-1} C')^{-1} C a \]

which approximately follows a chi-square distribution with \(\text{Rank}(C)\) degrees of freedom. Since kappa statistics are functions of the cell probabilities, by providing the proper transformation matrix \(f\) we can generate response functions which are estimates of kappa statistics, and use the \(Q_e\) test statistic for evaluating hypotheses of interest.

**Generating the Response Functions**

The method of GSK (and PROC CATMOD) restricts the type of function which may be applied to \(p\). The functions are restricted to any combination of the following four operations:

1. Linear combination \((A_1)\)
2. Natural logarithm \((\text{LOG})\)
3. Exponential \((\text{EXP})\)
4. Adding constant \((+K)\)

The \(F\) matrix which constructs the relevant kappa statistics has the following general form:

\[ F(p) = \text{EXP}(A_1 \text{LOG}(A_2 \text{EXP}(A_3 \text{LOG}(A_4 + K))) \]

The first \(1-J\) rows of \(A_1\) compute the row sums \(p_1, p_2, \ldots\) respectively, while the next row computes the quantity \(\text{log}(p)\). The rows of \(A_2\) compute the quantities \(\text{log}(p_i p_j)\), for use in computing cell expected values. \(A_3\) computes the numerator and denominator of the kappa statistic given in (3). Finally, \(A_4\) computes the quantity \(\text{log}(k_w)\). The final exponentiation produces the desired statistic.

Several usage notes concerning the function \(F(k)\) must be noted. First, since unweighted kappa is simply a special case of the weighted kappa, only the formulation for the construction of \(k_w\) is given. By setting \(M = 1\), we can obtain an unweighted estimate if desired.

Second, many different \(W\) matrices can be constructed, depending on the hypotheses of interest. Computation of multiple \(k\), statistics is possible with one transformation statement by simply adding rows to the various linear operator matrices \(A_1\), in accordance with the desired weight matrices.

Third, the above formulation assumes only one population is of interest. However, provisions for handling generalized kappa models for multiple populations can be constructed from \(F(k)\) by allowing the linear operators \(A_1\) to be applied in block operator form to the respective populations. Thus, for \(s\) distinct populations, \(p\) is \((s t \times 1)\), \(f(x)\) is \((s u \times 1)\), \(V_r\) is \((s u \times s u)\), \(X\) is \((t \times su)\), and so on.

To illustrate the flexibility of PROC CATMOD, an example which tests three different hypotheses will be given. The three hypotheses are:

1. The overall measure of agreement as measured by unweighted kappa is not significantly different from zero, or some constant value \(K\).
2. The individual measure of agreement for various categories of the agreement matrix are not significantly different from zero, or some constant value.
3. The difference in concordance as measured by kappa statistics for three distinct populations is not significantly different from zero.

**Example**

The example comes from a study by Coie, Terry, Finn, and Krehbiel (1997, in press). Two methods are used to assign children to one of three social status groups with respect to peer relationships. The first method is based upon a two dimensional finite choice social preference model, whereas the second method is based upon a unidimensional rating scale model.

The three resulting groups are (1) popular \((P)\) kids, who are extremely well liked by their peer group, (2) average \((A)\) kids, who are neither extremely liked or extremely disliked by their peers, and (3) rejected \((R)\) kids, who are extremely disliked by their peer group.
The questions of interest relate to the interchangeability of the two methods. That is, could the rating scale model be used in place of the choice model, since the rating scale is both easier to administer and avoids possible ethical dilemmas.

Furthermore, the concordance between the two methods with respect to each specific category is of particular interest. For instance, if groups are identified solely for the purpose of clinical intervention with rejected kids, then it is only the concordance for the rejected category which is of paramount interest.

Finally, the timing of identification of these children may be of some importance. Therefore, the issue of interchangeability as a function of grade level will also be examined.

The data were collected for three different grades and are given in Table 2.

To test the interchangeability of the two methods, kappa statistics were constructed for both the overall table and for each specific category of popular, average, and rejected. Furthermore, the concordance between the two methods as a function of grade will be tested by considering each grade to be a distinct population.

This analysis, therefore, involves \( s = 3 \) populations and \( u = 4 \) different response functions, resulting in a \((12 \times 1)\) column vector \( f \) to be analyzed.

Conceptually, computing kappa statistics for each of the specific categories involves collapsing the \( P \) matrices into \( 2 \times 2 \) matrices of hits and misses on that category. However, proper use of weight matrices can accomplish the same goal with less computation. The weight matrices for estimating both the overall and specific categories of agreement are given in Table 1.

The data setup for using PROC CATMOD to compute these kappa statistics is in the following table input format.

<table>
<thead>
<tr>
<th>Table 2. DATA FROM COLE, TERRY, FINN, KREHBIEEL (1987)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>O</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>METHOD 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>T</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>O</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Once the CATMOD procedure has been invoked, the first required statement is the WEIGHT statement, since the data are in the form of frequency counts. Specifying the statement

\[ \text{WEIGHT COUNT;} \]

identifies the variable \text{COUNT} as containing the frequency counts.

Second, since we will specify the design matrix \( X \) to equal an identity matrix \( I \), we need to use the POPULATION statement to induce the
proper analysis. Specifying

**POPULATION GRADE**

is sufficient to induce Kronecker product operations on the \( A_i \) matrices in the formulation of \( F \). This allows for identical response functions to be applied to each population of interest.

The most crucial aspect to the analysis is the specification of the **RESPONSE** statement. By specifying the formulation given in (8), the kappa statistics will be estimated properly.

The most crucial aspect to the analysis is the specification of the RESPONSE statement. By specifying the formulation given in (8), the kappa statistics will be estimated properly.

\[
A_i = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

\[
A_e = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

The first 3 rows of \( A_i \) compute the column sums, the second 3 rows of \( A_i \) compute the row sums, and the last 4 rows compute the \( P_s \) according to the weight matrices for overall, average, popular, and rejected, successively.

Finally, for classification with respect to clinical intervention with rejected children, it is clear that for grade 3, the choice of method might be extremely important, since the kappa value for the rejected category at the third grade level is extremely low.

It should be noted that these analyses in no way imply that one method is to be preferred over the other. The correct conclusion is to simply note that the two methods are not interchangeable, especially at lower grade levels.
Table 3.

SELECTED CATMOD OUTPUT

---

<table>
<thead>
<tr>
<th>PARAM</th>
<th>EST.</th>
<th>S.E.</th>
<th>SQUARE</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL 1</td>
<td>0.207</td>
<td>0.039</td>
<td>27.19</td>
<td>0.0001</td>
</tr>
<tr>
<td>2</td>
<td>0.146</td>
<td>0.043</td>
<td>11.50</td>
<td>0.0007</td>
</tr>
<tr>
<td>3</td>
<td>0.327</td>
<td>0.055</td>
<td>35.32</td>
<td>0.0001</td>
</tr>
<tr>
<td>4</td>
<td>0.191</td>
<td>0.047</td>
<td>13.80</td>
<td>0.0002</td>
</tr>
<tr>
<td>5</td>
<td>0.314</td>
<td>0.051</td>
<td>61.16</td>
<td>0.0001</td>
</tr>
<tr>
<td>6</td>
<td>0.224</td>
<td>0.043</td>
<td>32.01</td>
<td>0.0001</td>
</tr>
<tr>
<td>7</td>
<td>0.461</td>
<td>0.049</td>
<td>86.23</td>
<td>0.0001</td>
</tr>
<tr>
<td>8</td>
<td>0.423</td>
<td>0.042</td>
<td>97.62</td>
<td>0.0001</td>
</tr>
<tr>
<td>9</td>
<td>0.368</td>
<td>0.046</td>
<td>64.14</td>
<td>0.0001</td>
</tr>
<tr>
<td>10</td>
<td>0.440</td>
<td>0.058</td>
<td>56.64</td>
<td>0.0001</td>
</tr>
<tr>
<td>11</td>
<td>0.496</td>
<td>0.055</td>
<td>79.33</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

ANALYSIS OF CONTRASTS

<table>
<thead>
<tr>
<th>CONTRAST</th>
<th>DF</th>
<th>CHI-SQ</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(K)=P(K)=P3(K)</td>
<td>2</td>
<td>13.67</td>
<td>0.0011</td>
</tr>
<tr>
<td>P1(K)=P2(K)=P3(K)</td>
<td>2</td>
<td>12.42</td>
<td>0.0008</td>
</tr>
<tr>
<td>P1(K)+P2(K)=P3(K)</td>
<td>2</td>
<td>13.80</td>
<td>0.0001</td>
</tr>
<tr>
<td>OVERALL LINEAR TREND</td>
<td>1</td>
<td>13.67</td>
<td>0.0002</td>
</tr>
<tr>
<td>A LINEAR TREND</td>
<td>1</td>
<td>12.41</td>
<td>0.0004</td>
</tr>
<tr>
<td>P LINEAR TREND</td>
<td>1</td>
<td>1.98</td>
<td>0.1595</td>
</tr>
<tr>
<td>R LINEAR TREND</td>
<td>1</td>
<td>16.02</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Discussion

We have attempted to show how PROC CATMOD can be used to construct and estimate a variety of kappa statistics. Furthermore, we have shown how various tests of hypotheses can be undertaken with PROC CATMOD. This provides a powerful tool for the sophisticated use of the kappa statistic.

There are several considerations which must be addressed concerning the usage of PROC CATMOD and the GSK method for the present purposes. First, these procedures depend upon asymptotic considerations. Therefore, a sample size of at least 30 is recommended for theory to hold. Failure to heed this requirement will result in inflated standard errors, since linearized Taylor series approximations are used to compute the covariance matrix of \( \hat{f} \).

Second, it is clear that some values of zero will occasionally be observed in the off-diagonal cells of the probability matrix. CATMOD will automatically treat these observed zeros as structural zeros, provided they are observed consistently across populations. This will result in a smaller \( f \) vector than expected. Since these are by nature sampling zeros, it is necessary to substitute some arbitrarily small value (1E-20) to induce the proper structure on the data.

Finally, it is possible to test kappa statistics against some value other than zero, since this can often be a meaningless test. Use of the constant operator in the RESPONSE statement will enable the subtraction of the constant value one wishes to test against, allowing a more reasonable test to be made.

REFERENCES


NOTES

1. SAS is a registered trademark of SAS Institute Inc., Cary, NC.

2. The author can be contacted at the following address:

   Davie Hall 813A
   University of North Carolina
   Chapel Hill, NC 27514
   (919) - 962 - 7043