Introduction

When two observers independently categorize responses according to the same nominal variable, a statistic is often needed which measures the agreement between them. Agreement must be distinguished from association in that the former requires a pair of responses to fall into an identical category, whereas the latter requires only the predictability of one response from the other. Indeed, a table of responses may exhibit perfect association with little or no agreement so that many standard measures associated with the analysis of contingency tables provide little insight when assessing agreement.

The need for measures of agreement has been especially critical in psychiatry for evaluating the reliability of both diagnostic instruments and symptom rating scales. In general, there is a lack of laboratory tests or other external validators that may be used as standards in evaluating psychiatric instruments. Accordingly, investigators must rely on the strength of agreement between different raters to obtain information about error. Even when an instrument of "known" reliability is used in an investigation, reliability studies are necessary to ensure that the training of raters has been adequate.

Hall (1974) examined Volumes 120 and 121 (1972) of the British Journal of Psychiatry, and found that only 7 out of the 25 papers which used rating scales included an assessment of inter-rater reliability (4 using a correlational method and 3 an agreement method). However, acceptance of the Kappa and Weighted Kappa statistics was rapid after their introduction into the psychiatric literature (Fleiss et al., 1972; Hall, 1974) and Kappa is currently the statistic of choice for assessing inter-rater reliability in psychiatry. In addition, Kappa is used in follow-up studies to assess diagnostic consistency across time.

There is a corresponding need for the ability to compute Kappa and related statistics directly from a SAS data base. Currently, investigators must rely either on the availability of non-SAS special purpose programs or a hand calculator. Often Kappa values are reported with inappropriate standard errors, or with none at all. The Macro described here permits computation of some of these statistics on a routine basis.

The Kappa Statistic

Cohen (1960) introduced Kappa as an index of agreement to take chance agreement into account. Measures such as overall agreement or pairwise positive agreement are limited in that they do not incorporate information on marginal frequencies (Cloninger et al., 1979).

Consider the following cross-classification of responses by two observers for a variable with n categories,

\[
\begin{array}{cccc}
& 1 & 2 & \ldots & n \\
1 & p_{11} & p_{12} & \ldots & p_{1n} \\
2 & p_{21} & p_{22} & \ldots & p_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
 n & p_{n1} & p_{n2} & \ldots & p_{nn} \\
\end{array}
\]

where \( p_{ij} \) is the joint probability that the first observer places an observation in the \( i \)th category and the second observer places that observation in the \( j \)th category. The probabilities \( p_{ij} \) and \( p_{i.} \) are the marginal distributions for the first and second observer, respectively, where \( \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot
\[ \hat{\kappa} = \frac{\hat{P}_{ij} - \hat{P}_{ji}}{\hat{P}_{ij} + \hat{P}_{ji}} \]

Under the null hypothesis that \( \kappa = 0 \), this formula reduces to that given in Fleiss (1973):

\[ \hat{\kappa} = \frac{\hat{P}_{ij} - \hat{P}_{ji}}{\hat{P}_{ij} + \hat{P}_{ji}} \]

However, (2) rather (3) is appropriate when computing confidence limits for \( \kappa \) or when testing the equality of Kappas in two independent samples. Indeed, in most settings, investigators are interested in the precision of their estimates and the test that \( \kappa = 0 \) is not seriously considered. Also, when equality of Kappas in two samples is tested, the appropriate (large sample) statistic under the null hypothesis that \( \kappa_1 = \kappa_2 \) is given by

\[ \hat{\kappa} = \frac{\hat{P}_{ij} - \hat{P}_{ji}}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}'^2}} \]

where \( \hat{\sigma}^2 \) and \( \hat{\sigma}'^2 \) are computed using (2). Under the assumption that \( \kappa_1 = \kappa_2 \), (4) is approximately a standard normal random variable.

Agreement for Ordinal Data

Often ratings are made on a scale with a natural ordering and it is desired to give 'partial credit' when responses are close to one another. For example, diagnoses are often rated as absent, probable or definite, and Kappa as defined above would not distinguish between absent-definite or probable-definite disagreements. Symptom rating scales are often defined by a 5-point or 7-point severity scale rather than simply as a dichotomous present-absent variable, so that the degree of disagreement becomes important in assessing reliability.

Cohen (1968) developed the Weighted Kappa Statistic, \( \kappa_w \), to take into account the relative level of disagreement between two observers by assigning a priori weights to the levels of disagreement. Cohen (1968) pointed out that if quadratic weights are assigned to levels of disagreement then \( \kappa_w \) is equal to the product moment correlation coefficient.

Krippendorff (1970) advocates the use of the product moment correlation coefficient since it leads to several desired properties and has familiar interpretations in measuring agreement. The Macro described here gives the value of the product moment correlation coefficient with categories coded as 1, 2, ..., \( n \).

McNemar's Test of Symmetry

Let \( P_{ij} = \frac{1}{n} P_{i+} P_{+j} \) in the above table, and suppose that B observations fall above the diagonal and C observations fall below. The null hypothesis that \( P_{ij} = P_{ji} \) may be tested using the McNemar-like statistic (Bishop et al., 1975)

\[ \chi^2 = \frac{(B-C)^2}{B+C} \]

which is approximately a chi-square variable with 1 degree of freedom. For a 2x2 table this is the usual McNemar test without continuity correction for testing marginal homogeneity using paired data. For ordinal data with \( n^2 \) this statistic tests whether one rater tends to give higher ratings than the other. Equation (5) is simply a test that the B+C off-diagonal elements are distributed as a binomial random variable with probability \( \frac{1}{2} \) of lying above the diagonal and \( \frac{1}{2} \) of lying below. Accordingly, an exact test is recommended if \( B+C \leq 20 \).

The Use of KAPPAAL

The Macro KAPPAAL uses PROC MATRIX to read in two data sets, KAP_DAT and KAP_N, and outputs the value of \( \kappa \), the standard error of \( \kappa \), 95% confidence limits for \( \kappa \), the product moment correlation coefficient and McNemar's Test of symmetry. The data set KAP_DAT contains the variables for which the above statistics are to be calculated, first for variables 1 and 2, then variables 3 and 4, etc. The data set KAP_N has one observation consisting of the number of categories for each pair of variables in KAP_DAT. Thus if \( m \) sets of statistics are to be computed, there will be \( 2m \) variables in KAP_DAT and \( m \) in KAP_N. For an \( n \times n \) table, the corresponding pair of variables in KAP_DAT must be coded from 1 to \( n \).

Labelling for up to 10 sets of computations may be supplied by including the following statements prior to use of KAPPAAL:

MACRO NOTE 1 label 1 %
MACRO NOTE 1 label 2 %
etc
KAPPAAL

These labels will be printed prior to the table of observations for each pair of variables.

A listing of the Macro is given in the appendix, and the following example demonstrates its use.

Example

Helzer et al (1980) conducted a reliability study of a structured diagnostic interview on a total of 101 psychiatric inpatients. Three types of raters were used: physicians using an itemized interview (A raters), physicians using a verbatim version of the interview (B raters), and trained non-physicians using the verbatim interview (C raters). The following program evaluates agreement on the 98 patients who were interviewed by both A and B raters for the following two variables: (1) presence of a period of low mood existing at least 4 weeks, coded as present or absent, and (2) the diagnosis of depression, coded as absent, questionable, probable, or definite.
DATA KAP DAT;
MERGE INT_A(IN_A) INT_B(IN_B); BY ID;
IF A AND B;
KEEP X1-X4; IN_A X1 X2 X3 X4;
IF DX DEP='PROB' THEN X3=3;
IF DX DEP='DEF' THEN X3=4;
DATA KAP N;Y1=2;Y2=4; OUTPUT;
MACRO NOTE1 DEMONSTRATION OF "THE KAPPA MACRO%
MACRO NOTE2 THIS WOULD BE THE SECOND LABEL IF THERE WERE ONE%
KAPPLE%__

DEMONSTRATION OF THE KAPPA MACRO%
MATRIX OF OBSERVATIONS

<table>
<thead>
<tr>
<th>PROB</th>
<th>COL1</th>
<th>COL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW1</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>ROW2</td>
<td>4</td>
<td>65</td>
</tr>
</tbody>
</table>

OKAP KAPPA ST_ERR 95 PER CENT CLS CORR McNemar
.0.605091 0.091642 0.424997 0.785105 0.615597 3.266667

THIS WOULD BE THE SECOND LABEL IF THERE WERE ONE%
MATRIX OF OBSERVATIONS

<table>
<thead>
<tr>
<th>PROB</th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROW1</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>ROW2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ROW3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ROW4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>65</td>
</tr>
</tbody>
</table>

OKAP KAPPA ST_ERR 95 PER CENT CLS CORR McNemar
.0.665199 0.0804097 0.504216 0.819422 0.717842 1.14286

***************

Discussion
Kappa allows for the marginal distributions of the two observers to differ. Scott (1955) proposed the reliability statistic \( \kappa \) which is appropriate when the marginal distributions are identical. If, for example, observers are drawn at random from a pool of raters, then \( \kappa \) rather than \( \kappa_1 \) should be used as a measure of inter-rater reliability. For ordinal data, the intra-class correlation coefficient, computed by including each paired observation along with its transposition, should be used (Krippendorff, 1970). The general issue of the appropriateness of \( \kappa \) vs. \( \kappa_1 \) has not fully been settled (C. an exchange of letters between K. Krippendorff and J. Fleiss in Biometrics 34:pp. 142-144, 1978).

Reliability measures for a 2x2 table have been reviewed by Fleiss (1975) and Cloninger et al (1979). The product moment correlation reduces to the \( \phi \) coefficient, and we have \( \kappa \leq \phi \leq \kappa_1 \), with equality occurring when the two off-diagonal elements are equal.

A potential difficulty with the interpretation of Kappa can result from the model for "chance" agreement as defined in (1). If two psychiatrists were operating in a state of complete ignorance, and randomly assigned diagnoses at their respective base rates, then their agreement would be random. However, as Maxwell (1977) points out, in practice there are certain cases which are clear-cut and a psychiatrist will be confident of his diagnosis, and other cases which are problematic and provide a greater risk for error in agreement. Maxwell proposes an alternative to \( \kappa \), called the RE (random error) coefficient as a measure of agreement to reflect this alternative model for chance agreement. An alternative approach is to use a finer classification when diagnosing patients where symptomatology are close to the threshold of diagnosis, and may give insight into the sources of unreliability.

161
Finally, no single measure of agreement can be expected to serve all purposes. We have found examination of Kappa, along with the Pearson correlation, both for the diagnoses and the number of underlying positive symptoms, to be useful for evaluating agreement and for adding insight into the properties of the diagnostic instrument.

Acknowledgements

This research is supported in part by USPHS grants AA-03539, MH-31302, MH-25430, AA-03852, 278-79-0017DB. The authors wish to express their gratitude to Ms. Barbara Morris for her diligent typing.

References

APPENDIX

MACRO NOTE1 
MACRO NOTE2 
MACRO NOTE3 
MACRO NOTE4 
MACRO NOTE5 
MACRO NOTE6 
MACRO NOTE7 
MACRO NOTE8 
MACRO NOTE9 
MACRO NOTE10 

MACRO KAPPA1
PROC MATRIX; FETCH NCAT DATA=KAP N; FETCH Z DATA=KAP DAT; N=ROW(N);
NKAP=NCOL(NCAT); I=0; OKAP=J(1,5,-2); LOOP1: I=I+1; DIM=NCAT(I,1);
PROB=J(DIM,DIM,0); Z=0; LOOP2: I=I+1;
PROB(Z(12,2*I-1),Z(12,2*I))=PROB(Z(12,2*I-1),Z(12,2*I))+1;
IF I LE N-1 THEN GO TO LOOP2;
IF I=1 THEN NOTE NOTE1;
IF I=2 THEN NOTE NOTE2;
IF I=3 THEN NOTE NOTE3;
IF I=4 THEN NOTE NOTE4;
IF I=5 THEN NOTE NOTE5;
IF I=6 THEN NOTE NOTE6;
IF I=7 THEN NOTE NOTE7;
IF I=8 THEN NOTE NOTE8;
IF I=9 THEN NOTE NOTE9;
IF I=10 THEN NOTE NOTE10;
NOTE MATRIX OF OBSERVATIONS; PRINT PROB; PROB=PROB#/N;
T=TRACE(PROB); T2=(ROWSUM(PROB))*ROWSUM(PROB');
OKAP(1,1)=T1=T1-T2)/1-(T1-T2); T3=(ROWSUM(PROB)+ROWSUM(PROB'))*VECDIAG(PROB);
T5=(J(DIM,1)*ROWSUM(PROB))'; T6=J(DIM,1)*ROWSUM(PROB');
T6=ROWSUM((ROWSUM(PROB))'; T7=J(1,1)-T1; T7=(J(1,1)-T2)*T2=J(1,1)-T2; T8=2*T1-T3-T4-T5-T6-T9; T7=(J(DIM,1)*T7));
T12=(J(DIM,T7)); OKAP(1,2)=OKAP(1,2)-1; OKAP(1,1)=OKAP(1,1)1.96*OKAP(1,2);
Z=Z*(Z+1)/Z*(Z+1); SUMJ(1,2); XFX=XFX+Z; XFX=XFX-SUM#/N;
S=SQRT(VECDIAG(XFX)); T13=S*XFX; OKAP(1,5)=T13/(5,2);
M=J(DIM,-1); M=J(DIM,-1); M=J(DIM,-1); M=J(DIM,-1); M=J(DIM,-1); M=J(DIM,-1}; OKAP(1,6)=(C-OKAP(1,5))/OKAP(1,5); OKAPCOL='KAPPA' 'ST ERR' '95 PER' 'CEET CLS ' 'CORR' 'MONEMAR'; OKAPROW=''; PRINT OKAP ROWNAME=OKAPROW COLNAME=OKAPCOL;
IF I LE NKAP-1 THEN GO TO LOOP1;

163