Paper 280

Measurement of Interrater Agreement
A SAS/IML® Macro Kappa Procedure for Handling Incomplete Data

Honghu Liu, GIM/HSR, UCLA Department of Medicine, LA, CA
Ron D. Hays, GIM/HSR, UCLA Department of Medicine, LA, CA

ABSTRACT

Reliability refers to the extent to which the same score is obtained on multiple administrations. Interrater reliability is evaluated by comparing scores assigned to the same targets by two or more raters. Kappa is perhaps the most commonly used index of interrater agreement. Kappa is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement. Kappa can be obtained from SAS PROC FREQ, but this procedure can only handle complete data (i.e., each rater uses every possible choice on the response scale at least once). For incomplete data, one will either be unable to get kappa (if non-square table) or one will get a wrong kappa (if irregular square table) through SAS®. This paper introduces a SAS® macro procedure that calculates the correct kappa for either complete or incomplete data, and allows the user to specify any of three different weighting schemes. The SAS® products included are Base SAS®, SAS/STAT®, SAS/IML® and SAS® MACRO. This procedure can be run on any computer platform with a working SAS® system by even an entry-level analyst.

INTRODUCTION

Measuring interrater agreement is a common issue in business and research. Reliability refers to the extent to which the same number or score is obtained on multiple administrations or from independent methods of measurement. Interrater reliability is evaluated by comparing scores assigned to the same targets by two or more raters. Kappa is one of the most popular indicator of interrater agreement for nominal and ordinal data. The current kappa procedure in SAS® PROC FREQ works only with complete data (i.e., each rater uses every possible choice on the response scale at least once). For incomplete data, one will either be unable to get kappa (if it is a non-square table) or one will get a wrong kappa (if it is a irregular square table). Unfortunately, incomplete contingency tables are common especially when the number of possible choices on the response scale is large and the sample size is relative small. Therefore, a procedure which can handle both complete and incomplete data is needed. This paper describes a new user-friendly macro that computes kappa for both complete and incomplete data, and allows the users to specify any of three weighting schemes including user defined own weights.

AGREEMENT AND KAPPA STATISTICS

The kappa statistic was first proposed by Cohen (1960). For norminal data, kappa is mathematically equivalent to the intraclass correlation (the intraclass coefficient is a widely used measure of interrater reliability for the case of quantitative ratings). For ordinal and interval-level data, weighted kappa and the intraclass correlation are equivalent under certain conditions (Fleiss & Cohen 1973).

Suppose, as shown below, there are two raters, rater A and rater B who rate N subjects with k possible choices on the response scale:

	Rater B				
Rater A	1	2	…	k	Total
1	p_{11}	p_{12}	. . .	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	. . .	p_{2k}	$p_{2.}$
.					
.					
.					
k	p_{k1}	p_{k2}	. . .	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	. . .	$p_{.k}$	1

The overall proportion of observed agreement is defined as:

$$p_0 = \sum_{i=1}^{k} p_{ii} \ ,$$

the overall proportion of chance-expected agreement is defined as:

$$p_c = \sum_{i=1}^{k} p_{i.} p_{.i} \ ,$$

and unweighted kappa is defined as:

$$\hat{k}_{uw} = (p_0 - p_c) / (1 - p_c) \ . \qquad (1)$$

The standard error of kappa (1) is:

$$s.e.(\hat{k}_{uw}) = (1 / (1 - p_c)\sqrt{N}) \\ * \sqrt{p_c + p_c^2 - \sum_{i=1}^{k} p_{i.} p_{.i} \ (p_{i.} + p_{.i})}$$

For weighted Kappa, suppose we have weights w_{ij} $(i = 1, \ldots, k, \ j = 1, \ldots, k)$ which are assigned on certain rational (for example clinical) grounds. These weights are restricted to

$$0 <= w_{ij} <= 1$$

with

$$w_{ii} = 1, \quad \text{for } i = 1, \ldots, k$$

and

$$0 <= w_{ij} < 1 \quad \text{for } i \neq j$$

and

$$w_{ij} = w_{ji} \ .$$

Taking account the weights, the observed weighted proportion of agreement is now

$$p_{w0} = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} w_{ij} \ ,$$

the chance expected weighted proportion of agreement is

$$p_{wc} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.i} \ ,$$

and the weighted kappa is

$$\hat{k}_w = (p_{w0} - p_{wc}) / (1 - p_{wc}) \qquad (2)$$

If plug $w_{ij} = 0$ for all $i \neq j$, then we will get unweighted kappa.

INCOMPLETE DATA

In the idea situation, if the maximum possible choices on the rating scale is m, we will expect to have an m by m square table to be used to calculate kappa. For example, if the maximum number of possible choices is 4, then for complete data, a regular 4x4 square table will look like (frequency table):

		Rater 1				
	Score	1	2	3	4	
	1	a	b	c	d	r1
Rater 2	2	e	f	g	h	r2
	3	i	j	k	l	r3
	4	m	n	o	p	r4
		c1	c2	c3	c4	

with r1, r2, r3, r4, c1, c2, c3 and c4 are all greater than zero. However in reality, this is not necessarily the case. Instead, very often, one or both raters will not use every possible choice on the response scale. Therefore, some of the margins are zero, producing an irregular table. The problematic irregular tables can be classified into two categories:

(1) Non-square tables:

If the numbers of zero marginal values are different for the rows and columns, one will end up with a non-square table. For example in the above case, if r2=0, but c1, c2, c3 and c4 are all greater than zero, then one will have a table like the following:

		Rater 1				
	Score	1	2	3	4	
Rater 2	1	a	b	c	d	r1
	3	i	j	k	l	r3
	4	m	n	o	p	r4
		c1	c2	c3	c4	

this is a 3x4 non-square table.

(2) Irregular square tables:

If the numbers of zero marginal values are the same for the rows and columns, but they are on different levels of the rating scale, then one will encounter an irregular square table. For example if r2=0 and c1=0, then the table becomes:

		Rater 1			
	Score	2	3	4	
Rater 2	1	b	c	d	r1
	3	j	k	l	r3
	4	n	o	p	r4
		c2	c3	c4	

This is a 3 by 3 square table, but it is irregular. The diagonals are composed of different scores for the rows and columns.

For situation (1), SAS® system will claim that it has encountered a non-square table and will not calculate kappa. For situation (2), SAS® will treat it as a regular square table and calculate kappa based on this incorrect assumption.

A SAS®/IML MACRO PROCEDURE

In this section, we introduce a new macro that calculates correct kappa statistics for both complete and incomplete data. This SAS® macro procedure is written in SAS® MACRO language, Base SAS®, SAS/STAT® and SAS/IML®**.**

The format of the input data for this procedure can be either case wide raw data or frequency table data. There are three weighting scheme options: unweighted kappa, weighted kappa defined by Cicchetti and Allison (1971) and user specified weighting. This procedure calculates the standard error of kappa statistics and provides a statistical test for the null hypothesis that kappa equals to zero. The procedure also provides four levels of choice of confidence intervals for the kappa statistic.

At the beginning of the procedure, the data set is read in, and a two way table is generated by PROC FREQ. Then, the two way table is converted into a one dimension working table. This one dimension table is formed by using the parameter LEVEL which is equal to the maximum number of possible choices of the rating scale. If the data is incomplete, the missing row(s) or/and column(s) in the two way table are properly inserted. Therefore, the one dimension table has LEVEL*LEVEL number of entries, although some of the entries can have zero frequencies. A weighting vector is constructed depending on what weighting scheme one has chosen. If it is unweighted, then the weight of 1 for all the diagonal elements and 0 for all the off diagonal elements are assigned; if it is weighted by the weights defined by Cicchetti and Allison, then the weights

$$w_{ij} = 1 - \left| S_i - S_j \right| / (S_k - S_1) \quad i, j = 1, \dots, k$$

are generated, where k is the maximum number of possible choices of the rating scale, S_i and S_j are the scores for row i and column j; if the user specifies weights, the weighting vector is formed by the string from parameter w. The kappa statistic, its standard error and statistical testing are calculated in SAS/IML®. Four levels of confidence intervals (85%, 90%, 95% and 99%) are calculated in base SAS®. The syntax of the procedure is:

%kappa(**x,y,level,f,w,ci,dataset**)

Where **x** is the variable for rater one and **y** for rater two. Since this procedure can handle incomplete data, for case wide data, the value of **x** and **y** are required to be integers which correspond to the level of ratings in the complete data situation. **f** represents the type of frequency data. For usual case wide data, enter value 1. For frequency table data, **f** is the variable containing the value of the frequency for each cell. **w** is the weights for kappa. For unweighted kappa, just simply enter *uwt*. For weighted kappa (defined by Cicchetti and Allison,1971) enter *wt*. For user specified weights, enter the weights on one line with a space in between. For example, a 3x3 tables could have the following:

1 0.5 0.2 0.5 1 0.5 0.2 0.5 1.

ci is confidence interval for the kappa, the choices are 80, 85, 90, 95 and 99. **dataset** is the name of the data to be used in the analysis, the default (without entering a SAS data set name) is the last active data set.

To use the procedure, first to invoke the macro code (e.g., by using %include statement or just copy the entire macro/IML code to your SAS® program), then issue

%kappa(**x,y,level,f,w,ci,dataset**)

with all the parameters properly substituted.

DATA EXAMPLE

In this section, we show two examples in which incomplete data were used for calculating kappa.

In a study involved serious ill hospitalized patients, the same question regarding patient's quality of life (QOL) was asked from both the patients and their surrogate at the study entry. The QOL scale was an ordinal measure with four levels: excellent, good, fair and poor. Table 1 summarizes the data:

(Table 1)

```
q_pt                  q_sg

Frequency |
Percent   |
Row Pct   |
Col Pct   |excellt |good    |poor    |  Total
          |        |        |        |
----------+--------+--------+--------+
excellt   |     10 |     33 |     23 |     66
          |   1.24 |   4.08 |   2.85 |   8.17
          |  15.15 |  50.00 |  34.85 |
          |  20.41 |  10.15 |   5.30 |
----------+--------+--------+--------+
good      |     31 |    162 |    100 |    293
          |   3.84 |  20.05 |  12.38 |  36.26
          |  10.58 |  55.29 |  34.13 |
          |  63.27 |  49.85 |  23.04 |
----------+--------+--------+--------+
fair      |      5 |     85 |    106 |    196
          |   0.62 |  10.52 |  13.12 |  24.26
          |   2.55 |  43.37 |  54.08 |
          |  10.20 |  26.15 |  24.42 |
----------+--------+--------+--------+
poor      |      3 |     45 |    205 |    253
          |   0.37 |   5.57 |  25.37 |  31.31
          |   1.19 |  17.79 |  81.03 |
          |   6.12 |  13.85 |  47.24 |
----------+--------+--------+--------+
Total           49      325      434      808
              6.06    40.22    53.71   100.00
```

We can see from table 1 that none of the surrogates has answered 'fair' for patient's quality of life, so we end up with a 4x3 non-square table. To calculate kappa statistic for this contingency table by SAS®, one will get an error message in SAS® log saying "Agree statistics are computed only for tables where the number of rows equals the number of columns." But using the new macro procedure, we can get kappa statistic for this non-square table. Since the data (named ps_agree) was a case wide data set with a four levels of rating scale, for an unweighted kappa with a 95% confidence interval, we can simply issue:

%kappa(q_pt, q_sg, 4, 1, 95, ps_agree);

in the SAS program and got the unweighted kappa results:

```
        Kappa Statistic with 95% Confidence Interval

 KAPPA        SE        P_VALUE     LOWER_95     UPPER_95

0.21672    0.021015        0         0.17553      0.25791
```

As the second example, this same cohort of patients were followed up for 6 months, the same question was then asked again from both patient and surrogate. Table 2 below shows the data (due to missing values, the sample size is now much smaller):

```
                (Table 2)
q_pt             q_sg

Frequency |
Percent   |
Row Pct   |
Col Pct   |excellt |good    |poor    |  Total
          |        |        |        |
----------+--------+--------+--------+
excellt   |     25 |     63 |      3 |     91
          |   7.18 |  18.10 |   0.86 |  26.15
          |  27.47 |  69.23 |   3.30 |
          |  75.76 |  30.58 |   2.75 |
----------+--------+--------+--------+
fair      |      7 |    122 |     40 |    169
          |   2.01 |  35.06 |  11.49 |  48.56
          |   4.14 |  72.19 |  23.67 |
          |  21.21 |  59.22 |  36.70 |
----------+--------+--------+--------+
poor      |      1 |     21 |     66 |     88
          |   0.29 |   6.03 |  18.97 |  25.29
          |   1.14 |  23.86 |  75.00 |
          |   3.03 |  10.19 |  60.55 |
----------+--------+--------+--------+
Total           33      206      109      348
              9.48    59.20    31.32   100.00
```

From table 2 we can see that none of the patients has answered ‘good’ and none of the surrogates has answered ‘fair’, therefore, we end up with an irregular 3x3 square table. To calculate kappa, SAS will treat the table as a regular 3x3 square table (e.g., treats the 122 observations with patient’s answer of ‘fair’ and surrogate’s answer of ‘good’ as a true diagonal value) and gets an unweighted kappa=0.363 (incorrect). However, if we use the new macro to calculate kappa for this data, the procedure actually calculates kappa statistic based on the following modified table:

(Table 3)

```
q_pt            q_sg

Frequency |
          |
          |
          |excellt |good    |fair    |poor    |  Total
          |        |        |        |        |
----------+--------+--------+--------+--------+
excellt   |     25 |     63 |   0    |      3 |    91
----------+--------+--------+--------+--------+
good      |      0 |      0 |   0    |      0 |    0
----------+--------+--------+--------+--------+
fair      |      7 |    122 |   0    |     40 |    169
----------+--------+--------+--------+--------+
poor      |      1 |     21 |   0    |     66 |    88
----------+--------+--------+--------+--------+
Total           33      206     0          109     348
```

We can see in table 3, a ‘fair’ column and a ‘good’ row are inserted in the table. Those cells that were in wrong positions in table 2 have now been relocated to right positions. The macro procedure gives the following correct results:

```
    Kappa Statistic with 95% Confidence Interval

 KAPPA         SE        P_VALUE     LOWER_95     UPPER_95

0.17577     0.014794        0         0.14678      0.20477
```

CONCLUSION

This SAS macro Kappa procedure has several strengths. First, the most important feature of this procedure is that it can handle both complete and incomplete data. Second, this procedure has three different weighting scheme options for users to chose from. The option of user defined weights is unique and useful and is also not available in the current SAS kappa procedure. Third, this procedure has four levels of confidence intervals options (85%, 90%, 95% and 99%). This procedure is also user-friendly and has a simple syntax.

This procedure calculates kappa for two raters only, and does not apply to designs with more than two raters. It is also important to note that kappa is strongly dependent on the marginal distribution. A procedure which can calculate several different reliability measures in one package may be a useful future project.

The SAS code of the procedure is available upon request from the authors.

REFERENCE

SAS/STAT® User’s Guide, SAS Institute Inc., SAS Campus Drive, Cary, NC.

SAS/IML® Software: Usage and Reference, Version 6, SAS Institute Inc., Cary, NC.

SAS® Guide to Macro Processing, Version 6, Second Edition, SAS Institute Inc., Cary, NC.

Fleiss, J. L. (1981), Statistical method for rates and proportions, John Wiley, New York.

Cicchetti, D. V. and Allison, T. (1971), A new procedure for assessing reliability of scoring EEG sleep recordings, American Journal of EEG technology, 11, 101-109.

Cohen, J(1960), A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20, 37-46.

Fleiss, J. L. & Cohen, J (1973), The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33, 613-619.

CONTACT AUTHORS

Your comments and questions are valued and encouraged. Contact the authors at:

Honghu Liu
General Internal Medicine & Health Service Research
UCLA Department of Medicine
LA, CA 90095-1736
Phone 310-794-7396
Fax 310-206-0719
email hhliu@ucla.edu

Ron Hays
General Internal Medicine & Health Service Research
UCLA Department of Medicine
LA, CA 90095-1736
Phone 310-794-7508
Fax 310-206-0719
email Ronald_Hays@rand.org