Paper 180-30

Calculation of the Kappa Statistic for Inter-rater Reliability: The Case Where Raters Can Select Multiple Responses from a Large Number of Categories

Catherine R. Stein, US Army Center for AMEDD Strategic Studies, San Antonio, TX
Raymond B. Devore, Jr, US Army Center for AMEDD Strategic Studies, San Antonio, TX
Barbara E. Wojcik, US Army Center for AMEDD Strategic Studies, San Antonio, TX

ABSTRACT

A common measure of rater agreement where outcomes are nominal is the kappa statistic (a chance-corrected measure of agreement). You can use PROC FREQ to calculate the kappa statistic, but only if the given frequency table is square (that is, raters used the same categories). In most rater analyses involving nominal outcomes, a rater assigns a single response based on a small number of categories (usually fewer than 10). What if for each observation raters were able to select multiple responses (with, for example, one having 5 responses and another having 8) from a very large number of categories (approximately one thousand)? This paper shows you how to obtain the kappa statistic using PROC FREQ in BASE® SAS when raters select multiple responses from a large number of categories. The discussion is intended for users of all skill levels.

INTRODUCTION

We were recently tasked to determine if reliable mapping is possible between two different classification systems related to medical conditions—Department of Defense patient condition (PC) codes, used to refer to groups of related conditions when planning for medical readiness during combat operations, and the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) diagnosis codes, used in medical records to define specific patient diseases and injuries. Currently, there are 389 PC codes and each has a treatment brief associated with it that describes for each echelon of care or treatment level within the theater: the condition of the patient, the treatment to be provided, and the probabilities on disposition of the patient (return to duty, evacuation to the next treatment level, or death). All of this information is general in nature and based on the opinions of subject matter experts. We analyzed the mapping efforts of three registered health information specialists who had been given the treatment briefs and asked to assign all pertinent ICD-9-CM diagnoses codes to each PC code.

In order to have a meaningful and useful conversion between two coding systems, there has to be agreement, or consistency, in the judgments of the coders or raters (i.e., inter-rater reliability). Two methods are commonly used to measure rater agreement where outcomes are nominal: percent agreement and Cohen's chance-corrected kappa statistic (Cohen, 1960). In general, percent agreement is the ratio of the number of times two raters agree divided by the total number of ratings performed. The kappa statistic estimates the proportion of agreement among raters after removing the proportion of agreement which would occur by chance. The upper limit of kappa is +1.00 and occurs when there is total agreement. A kappa of 0.00 indicates agreement at a chance level. Negative kappa values represent agreement which is less than chance (lower limit is -1.00).

Most rater analyses with nominal outcomes involve situations where, for an observation, a rater assigns a single response based on several (usually fewer than 10) categories. However, the mapping between the two classification systems differed in two major ways: a) coders could assign multiple ICD9 codes to each PC code and b) coders had approximately 1,000 categories from which to choose (ICD-9-CM diagnoses reduced to 3-character codes (ICD9 codes)). We conducted literature searches, but were unable to find rater analyses involving either of these two conditions.

The FREQ procedure computes the kappa statistic based on frequency tables with one rater's responses being the rows and a second rater's responses being the columns. To perform kappa calculations in SAS, the frequency tables have to be square, which results only if both raters used the same categories for their responses. Because of the large number of ICD9 codes the coders could choose from, it was probable that some of the tables obtained in PROC FREQ to compute the kappa statistics for our study would not be square. This paper discusses the solution we found for obtaining the kappa statistic with SAS when raters are able to select multiple responses from a large number of categories.

SAMPLE DATA

A major part of the solution to working with multiple responses from a large number of categories is to have your data structured so that the desired frequency tables will result. First, you want to produce frequency tables for a pair of raters so that their response categories form the rows and columns of the tables. Second, you want both raters to have used all of the categories in the table. To obtain both these goals, we found it useful to get our data into a preliminary form where, in our example, there is one record for each PC code-ICD9 code combination used by any of the coders. On each record there is a flag variable for each coder, with "1" indicating the coder assigned that ICD9 code to that PC code, or the flag is set to missing. For illustration purposes, we restrict the analysis to two PC codes and start with the following data set:

```
* READ IN DATA SET CONSISTING OF 1 RECORD PER PC-ICD9
  COMBINATION AND INDICATE WHICH CODERS ASSIGNED THAT MAPPING ;
data indata ;
input PC $ ICD $ coder1 coder2 coder3 ;
datalines ;
0001  800  . 1 .
0001  801  . 1 .
0001  802  . 1 .
0001  803  1 1 1
0001  804  . 1 .
0001  850  1 1 1
0172  863  1 1 1
0172  867  1 1 1
;
```

In this example, you can see that the three coders did not assign the same ICD9 codes to PC code 0001, and hence, any frequency table comparing two of these coders would not be square and SAS would not compute the kappa statistic. We found a solution for producing square tables in a paper providing hints on calculating the kappa statistic using SAS software (John Uebersax, 2002). The author used pseudo-observations to make the tables square. Following his methodology, we created pseudo-records for each PC-ICD9 combination that existed in the data. We assigned a weight of one (1) to the real records and a very small weight (0.0000000001) to the pseudo-records. The pseudo-records ensured that each coder had responses for every ICD9 assigned by any other coder, and the small weights meant the pseudo-records had no effect on the calculated value of kappa.

Using the INDATA data set, we create the final data set for use by PROC FREQ. First, we transform the actual data: we replace the flag values of "1" with the given ICD9 code value, eliminate the ICD9 code variable, and add the weight variable assigning it a value of 1. Also, in order to get our square tables, we will need to replace the missing flag values with token values of "x." Following is the code for converting the data.

```
* REPLACE '1' CODER VALUES WITH ICD9 CODE, MISSING VALUE
  WITH 'X', DROP ICD VARIABLE, AND ADD WEIGHT OF 1 ;
data mappings (keep=PC CdrA CdrB CdrC wgt) ;
  set indata ;
  by PC ;
  length CdrA CdrB CdrC $ 3 ;
  if coder1=1 then CdrA=ICD ;
              else CdrA=' x ' ;
  if coder2=1 then CdrB=ICD ;
              else CdrB=' x ' ;
  if coder3=1 then CdrC=ICD ;
              else CdrC=' x ' ;
 * ASSIGN WEIGHT OF '1' TO REAL
   RECORDS ;
  wgt = 1 ;
run ;
```

Table 1 shows the data for PC code 0001 before and after the conversion.

Table 1. Conversion of initial data set records to weighted records for kappa calculation.

Initial PC-ICD9 Records				
PC	ICD9	Coder1	Coder2	Coder3
0001	800	.	1	.
0001	801	.	1	.
0001	802	.	1	.
0001	803	1	1	1
0001	804	.	1	.
0001	850	1	1	1

⇒

Converted PC-Coder Records				
PC	CoderA	CoderB	CoderC	Wgt
0001	x	800	x	1
0001	x	801	x	1
0001	x	802	x	1
0001	803	803	803	1
0001	x	804	x	1
0001	850	850	850	1

Next, we create one pseudo-observation for each PC-ICD9 combination. We assign the ICD9 code value to each coder indicating they all used the given PC-ICD9 combination. We also give the pseudo-observation the weight of 0.0000000001 (1E-10) so that its contribution to kappa will be negligible. Also, we need to create one pseudo-observation where all coders had the missing "x" value—this record accounts for missing values and keeps the frequency table square.

We then combine the real and pseudo-observations and sort by PC code. The following code creates the pseudo-observations and the desired data set:

```
* CREATE DUMMY DATA RECORDS TO ENSURE SQUARE TABLE FOR OBTAINING
  KAPPAS. ASSIGN A TINY WEIGHT SO DUMMY OBSERVATIONS DO NOT EFFECT
  KAPPA VALUES. INCLUDE ROW OF MISSING ('x') VALUES FOR EACH PC CODE ;
data dummy
        (keep=PC CdrA CdrB CdrC wgt) ;
  set indata ;
     by PC ;
  length CdrA CdrB CdrC $ 3 ;
  CdrA=ICD ;
  CdrB=ICD ;
  CdrC=ICD ;
  wgt = .0000000001 ;
  output ;
  if last.PC then
    do ;
      CdrA=' x ' ;
      CdrB=' x ' ;
      CdrC=' x ' ;
      output ;
    end ;
run ;

* CONCATENATE REAL & DUMMY DATA & SORT BY PC CODE ;
data wt_PcIcd ;
  set mappings dummy ;
run ;

proc sort data=wt_PcIcd ;
  by PC ;
run ;
```

Table 2 shows the pseudo-observations created for PC code 0001.

Table 2. Pseudo-observations (on right) created from actual observations (on left).

Initial PC-ICD9 Records				
PC	ICD9	Coder1	Coder2	Coder3
0001	800	.	1	.
0001	801	.	1	.
0001	802	.	1	.
0001	803	1	1	1
0001	804	.	1	.
0001	850	1	1	1

⇒

Pseudo PC-Coder Records				
PC	CoderA	CoderB	CoderC	Wgt
0001	800	800	800	1E-10
0001	801	801	801	1E_10
0001	802	802	802	1E_10
0001	803	803	803	1E_10
0001	804	804	804	1E_10
0001	850	850	850	1E_10
0001	x	x	x	1E-10

INTER-RATER ANALYSIS

Suppose we want to assess the reliability between coders in mapping individual PC codes. Also, suppose we have chosen to evaluate the inter-rater reliability using pairwise measurements among the three coders. Using the WT_PCICD data set consisting of CoderA-C records (both actual and pseudo), we create a subset for each pair of coders.

LESS THAN 100% AGREEMENT EXAMPLE

The following code creates a data set for assessing the reliability between coder 1 (CdrA) and coder 2 (CdrB) and then obtains the kappa statistic using PROC FREQ.

```
* CREATE DATA SET FOR COMPARING CODERS 1 & 2 ;
data testAB ;
  set wt_PcIcd ;
  * DELETE REAL RECORDS WHERE ICD CODE WAS NOT USED BY EITHER CODER ;
  if CdrA=' x ' and CdrB=' x ' and wgt=1
    then delete ;
run ;

* PRODUCE FREQ TABLE COMPARING CODERS 1 & 2. USE WEIGHT STATEMENT TO
  OBTAIN SQUARE TABLES, BUT USE NON-WEIGHTED KAPPA RESULTS          ;
proc freq data=testAB ;
  by PC ;
  weight wgt ;
  tables CdrA*CdrB / norow nocol nopercent
                     agree ;
run ;
```

SAS PROC FREQ produced the following frequency table and kappa statistic for CoderA vs. CoderB (Figure 1). CoderA's mappings form the rows and CoderB's mappings form the columns. Note that the value of "1" in a cell on the table's diagonal (top left to bottom right) represents agreement—only ICD9 codes 803 and 850 were mapped to PC code 0001 by both CoderA and CoderB.

The SAS System

PC=0001

The FREQ Procedure

Table of CdrA by CdrB

CdrA / Frequency, CdrB	x	800	801	802	803	804	850	Total
x	1E-10	1	1	1	0	1	0	4
800	0	1E-10	0	0	0	0	0	1E-10
801	0	0	1E-10	0	0	0	0	1E-10
802	0	0	0	1E-10	0	0	0	1E-10
803	0	0	0	0	1	0	0	1
804	0	0	0	0	0	1E-10	0	1E-10
850	0	0	0	0	0	0	1	1
Total	1E-10	1	1	1	1	1	1	6

Statistics for Table of CdrA by CdrB

Test of Symmetry

Statistic (S)	4.0000
DF	21
Pr > S	1.0000

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2941	0.1558	-0.0113	0.5995
Weighted Kappa	0.3889	0.2066	-0.0161	0.7939

Sample Size = 6.0000000007

Figure 1. Output from SAS PROC FREQ comparing the mapping of PC code 0001 by two coders.

Note that SAS produced statistics for both a simple kappa and a weighted kappa. Although we used weights, the actual data records had weights of "1" and the kappa of interest to us is the simple kappa of 0.29. For comparison, the actual percent agreement, without the adjustment for agreement that would have occurred by chance is calculated as follows using only the actual observations:

% agreement = (number of agreements/number of ICD9 codes compared)*100%
= (2/6)*100%
= 33.3%

100% AGREEMENT EXAMPLE

If there is 100% agreement between two coders, is kappa, which should equal 1.0, affected by the pseudo-observations? Following are the results from PROC FREQ of assessing the agreement between coders 1 and 2 in mapping PC code 0172 (Figure 2). The results show that although pseudo-observations are not needed when there is complete agreement, their inclusion has no effect on the computation of kappa—kappa still equals 1.0.

```
                                The SAS System

------------------------------------ PC=0172 -------------------------------------

                                The FREQ Procedure

                                Table of CdrA by CdrB

                CdrA      CdrB
                Frequency| x      |863     |867     |  Total
                ---------+--------+--------+--------+
                 x       |  1E-10 |      0 |      0 |  1E-10
                ---------+--------+--------+--------+
                863      |      0 |      1 |      0 |      1
                ---------+--------+--------+--------+
                867      |      0 |      0 |      1 |      1
                ---------+--------+--------+--------+
                Total       1E-10        1        1        2


                    Statistics for Table of CdrA by CdrB


                                   Test of Symmetry

                                  ----------------------
                                  Statistic (S)        .
                                  DF                   3
                                  Pr > S               .


                                   Kappa Statistics
          Statistic          Value       ASE     95% Confidence Limits
          ------------------------------------------------------------
          Simple Kappa      1.0000    0.0000       1.0000       1.0000
          Weighted Kappa    1.0000    0.0000       1.0000       1.0000


                                Sample Size = 2.0000000003
```

Figure 2. Output from SAS PROC FREQ comparing the mapping of PC code 0001 by two coders.

For comparison, the actual percent agreement, without the adjustment for agreement that would have occurred by chance is calculated as follows using only the actual observations:

% agreement = (number of agreements/number of ICD9 codes compared)*100%
= (2/2)*100%
= 100%

FURTHER CONSIDERATIONS

In our problem of assessing rater reliability in mapping between the two coding systems, we debated on how we would measure reliability given the multiple responses and many-possible-category situation. We decided to obtain reliability measures from three different viewpoints: 1) mappings of each individual PC code as illustrated in our example (i.e., what was the agreement between two coders in mapping a particular PC code), 2) identical mappings of individual PC codes (i.e., what was the agreement between two coders across all PC codes, if to be in agreement, two coders had to code a PC code to exactly the same set of ICD9 codes), and 3) mappings without regard to individual PC codes (i.e., what proportion of the time did two coders map to the same ICD9 code summed across all

PC codes). Some additional steps are necessary, but the same general techniques we illustrated, apply for the other two assessment views.

Results of applying the three reliability assessments to the example data are summarized in the following set of tables. Tables 3 and 4 present the results of comparing coders on the mapping of individual PC codes using the SAS programming provided in this paper. Table 3 presents the pairwise comparison by PC code and Table 4 summarizes the results by coder pair.

Table 3. Pairwise assessment of inter-rater reliability based on mappings of each PC code.

PC Code	Coders	# of Agreements	# of ICD9s Compared	% Agreement	Kappa Statistic	
					Kappa	95% Confidence Interval
0001	AB	2.0	6.0	33.3	0.2941	-0.0113, 0.5995
0001	AC	2.0	2.0	100.0	1.0000	1.0000, 1.0000
0001	BC	2.0	6.0	33.3	0.2941	-0.0113, 0.5995
0172	AB	2.0	2.0	100.0	1.0000	1.0000, 1.0000
0172	AC	2.0	2.0	100.0	1.0000	1.0000, 1.0000
0172	BC	2.0	2.0	100.0	1.0000	1.0000, 1.0000

Table 4. Summary of interrater agreements of three coders (A, B, C) summarized by coder pair.

Coders	Mean # Agreements	Mean # ICD9 Comparisons	Mean % Agreement	Mean Kappa Statistic
AB	2.0	4.0	66.7	0.6471
AC	2.0	2.0	100.0	1.0000
BC	2.0	4.0	66.7	0.6471
Overall	2.0	3.3	77.8	0.7647

Table 5 summarizes the results when only identical mappings of individual PC codes is considered to represent agreement, and Table 6 shows the agreement statistics when mappings were assessed without regard to individual PC codes.

Table 5. Interrater agreements based on identical mappings of PC codes.

Coders	# of Perfect Agreements	# of PCs Compared	% Agreement	Kappa Statistic	
				Kappa	95% Confidence Interval
AB	1.0	2.0	50.0	0.3333	0.0254, 0.6413
AC	2.0	2.0	100.0	1.0000	1.0000, 1.0000
BC	1.0	2.0	50.0	0.3333	0.0254, 0.6413
Mean	1.3	2.0	66.7	0.5556	

Table 6. Overall inter-rater agreements without regard to individual PC codes.

Coders	# of Agreements	# of ICD9 Comparisons	% Agreement	Kappa Statistic	
				Kappa	95% Confidence Interval
AB	4.0	8	50.0	0.4667	0.1464, 0.7870
AC	4.0	4	100.0	1.0000	1.0000, 1.0000
BC	4.0	8	50.0	0.4667	0.1464, 0.7870
Mean	4.0	6.7	66.7	0.6444	

CONCLUSION

The methods we presented in this paper, involving only simple SAS data steps, enabled us to perform an assessment of mapping between two complex coding systems. We believe these techniques will work in other situations where determining inter-rater reliability is based on raters choosing multiple responses from hundreds or thousands of categories.

REFERENCES

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20, 37-46.

Uebersax, John. "Calculating Kappa with SAS" revised 20 July 2002. <http://ourworld.compuserve.com/homepages/jsuebersax/saskappa.htm> (July 2, 2003).

CONTACT INFORMATION

If you have any questions or comments, contact the author at:

Catherine R. Stein
Center for AMEDD Strategic Studies (CASS)
1608 Stanley Rd Ste 47
Fort Sam Houston, TX 78234-5047
Work Phone: 210-221-9135
Email: catherine.stein@amedd.army.mil
Web:

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.