Paper 242-2009

More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters

Michael Cunningham, University of Pittsburgh, Pittsburgh, PA

ABSTRACT

The kappa coefficient is a widely used statistic for measuring the degree of reliability between raters. SAS® procedures and macros exist for calculating kappa with two or more raters, but none address situations when the kappa coefficient alone does not sufficiently describe the level of reliability. When the prevalence of a rating in the population is very high or low, the value of kappa may indicate poor reliability even with a high observed proportion of agreement. Researchers have recommended reporting several other values in addition to the kappa to address this and another paradox of the kappa statistic. This program, developed in SAS® 9.1, calculates kappa, but also outputs the observed and expected proportions of agreement, the prevalence and bias indices, and the prevalence adjusted bias adjusted kappa (PABAK) for two raters. Designed for input of the rater responses in the familiar 2x2 table format using the SAS %WINDOW statement, users with minimal SAS experience will be able to report these statistics to more fully characterize the extent of inter-observer agreement between two raters.

INTRODUCTION

The kappa coefficient is a widely used statistic for measuring the degree of reliability between raters. Highmark, Inc., one of the leading health insurers in Pennsylvania, uses the kappa statistic as an important component of its quality improvement and management programs. A typical assessment of a process involves the sampling of medical records by a rater, in most cases a nurse, who assigns a rating based on information within each record. A second nurse, independently from the first, examines the same records and also assigns a rating. These ratings are usually categorized as “yes” or “no”, based on how each nurse would answer a specific question about the record. The results of these responses are placed into a 2x2 table and sent to the University of Pittsburgh for analysis and interpretation.

Standard statistical packages, spreadsheets, or a calculator can be used to calculate the kappa coefficient from data in a 2x2 table. Based on the value of kappa, a general level of agreement can be assigned to indicate the amount of reliability between the ratings of the two observers. Therefore, in this setting, a high level agreement (as indicated by a high value of kappa) would show that for a given process, the two nurses generally abstract the same information from the sampled records.

The kappa statistic alone is appropriate if the marginal totals for the 2x2 table are relatively balanced, but if the prevalence of a given response is very high or low, the value of kappa may indicate a low level of reliability even when the observed proportion of agreement is quite high. Researchers have recommended reporting several other values in addition to the kappa to address this paradox of the kappa statistic. While there is no consensus as to all of the values that should be reported, there is agreement that the kappa alone is insufficient. This program, developed in SAS® version 9.1, calculates kappa, but also outputs the observed and expected proportions of agreement, the prevalence and bias indices, and the prevalence-adjusted bias-adjusted kappa (PABAK) for two raters to more fully characterize the extent of the inter-rater reliability (IRR) between two raters.

KAPPA STATISTIC

The results of a two rater analysis are often entered into a 2x2 table (Figure 1).

		Rater B		
		yes	no	Totals
Rater A	yes	a	b	a+b
	no	c	d	c+d
	Totals	a+c	b+d	N

Figure 1. Table of N Ratings for Rater A and Rater B

The kappa coefficient, κ, is calculated as $\kappa = (p_o - p_e)/(1 - p_e)$, where the observed proportion of agreement $p_o = (a+d)/N$ and the expected proportion of agreement $p_e = ((a+c)(a+b) + (b+d)(c+d))/N^2$.

The values of kappa range from -1 to +1, with -1 indicating perfect disagreement and +1 indicating perfect agreement between the raters. Several authors have categorized the values of kappa to indicate the strength of this agreement.

Landis and Koch's (1977) categorization is widely referenced and has been used in this program:

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

An interesting phenomenon known as the kappa paradox (Feinstein and Cicchetti, 1990) occurs when the observed proportion of agreement is high but the value of the kappa statistic is low. If this paradox is present and a standard statistical package is used to calculate kappa, an interpretation based solely on the value of kappa may give a misleading result.

As an example, consider the data in Figure 2:

		Rater B		
		yes	no	Totals
Rater A	yes	95	4	99
	no	1	0	1
	Totals	96	4	100

Figure 2. Sample Data Illustrating Kappa Paradox

The proportion of observed agreement in Figure 2 is 0.95 (95/100), but the kappa statistic is -0.0163, "poor" agreement (Landis and Koch, 1977). In this situation, the very high prevalence of "yes" responses in this population has resulted in a low value of kappa, even though raters A and B agreed on 95% of the ratings.

A number of researchers have proposed reporting additional values and statistics in addition to the kappa to provide a clearer picture of the IRR (Cicchetti and Feinstein, 1990; Lantz and Nebenzahl, 1996; Sim and Wright, 2005; Viera and Garrett, 2005). There is some disagreement as to which additional values should be reported, but there is a general consensus that the observed proportion of agreement, the expected proportion of agreement, the proportion of positive agreement, the proportion of negative agreement, the prevalence index, and the bias index are among those that should be reported with the kappa.

Consider again the data from Figure 2. It has been shown that kappa = -0.0163 and $p_o =$0.95. The values for these additional measures are shown below:

- expected proportion of agreement, $p_e = ((a+c)(a+b)+(b+d)(c+d))/N^2$ = 0.9508
- proportion of positive agreement, $p_{pos} = (2a)/(N+a-d)$ = 0.9744
- proportion of negative agreement, $p_{neg} = (2d)/(N-a+d)$ = 0.000
- prevalence index, $p_{index} = (a-d)/N$ = 0.95
- bias index, $b_{index} = (b-c)/N$ = 0.03

The values of 0.9744 and 0.95 for the proportion of positive agreement and the prevalence index, respectively, help to paint a clearer picture of the data in the 2x2 table. If the kappa alone was interpreted, you would conclude that there is poor reliability between raters. But if you had also calculated and reported the proportion of positive agreement and the prevalence index, you would realize that the low value of kappa was a result of the high prevalence of the "yes" responses in the population.

Another statistic, the prevalence adjusted bias adjusted kappa (PABAK), adjusts the kappa for imbalances caused by differences in the prevalence and bias (Byrt et al., 1993). Calculation of the PABAK depends solely on the observed proportion of agreement between raters, PABAK $= 2p_o - 1$. Some researchers have been critical of using the PABAK (Hoehler, 2000). The criticism seems to focus on the clinical setting, where it would be desirable to have a perfect 50/50 distribution of "yes" and "no" ratings in a population. But in our experience, certain processes (fortunately) have very high completion rates, as demonstrated in Figure 2, so low values of kappa with high prevalence rates are regularly observed. One other benefit of reporting the PABAK is that it may be more useful to show a single summary measure rather than reporting the kappa and several other prevalence and bias indices. Several other recent reliability studies have reported the PABAK as a part of their analyses (Girianelli et al., 2007; Cibere et al., 2008); therefore it is included in the program. In addition to the PABAK, this program includes all of the other measures discussed previously. It is left to the individual to decide which of these measures are appropriate and how they should be interpreted.

KAPPA PROGRAM

The data from Figure 2 will now be used to demonstrate how the kappa statistic and the other bias and prevalence measures described earlier can be obtained in SAS.

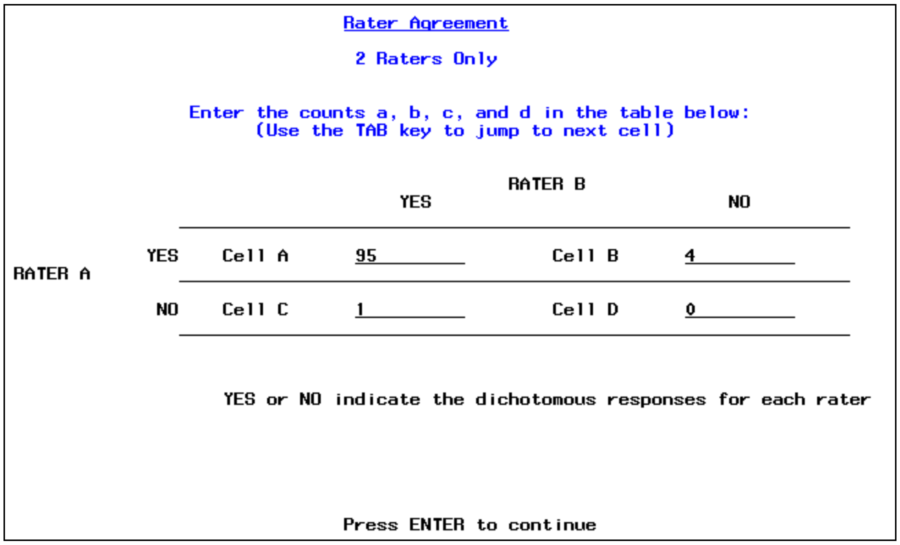
INPUT RATING RESPONSES USING %WINDOW

To allow for easier input of the rater responses, the %WINDOW macro is used to enter the values for cells a, b, c, and d in the 2x2 table format shown in Figure 1. For this example, a=95, b=4, c=1, and d=0. Running the code below will open up a window named KAP, where the cell values are entered.

```
%window kap color=white
#2 @33 'Rater Agreement' attr=(highlight,underline) color=blue
#4 @34 '2 Raters Only' attr=highlight color=blue
#7 @19 'Enter the counts a, b, c, and d in the table below:' attr=highlight
color=blue
#8 @25 '(Use the TAB key to jump to next cell)' color=blue
#11 @16 '                                RATER B                        '
#12 @38 'YES' attr = highlight @68 'NO' attr = highlight
#13 @18 '_____________________________________________________________'
#15 @15 'YES' attr = highlight @22 'Cell A'   @34 a 10 attr = underline required=yes
#15 @52 'Cell B'   @64 b 10 attr = underline required=yes
#16 @3 'RATER A' @18 '_____________________________________________________________'
#18 @16 'NO' attr = highlight @22 'Cell C'   @34 c 10 attr = underline required=yes
#18 @52 'Cell D'   @64 d 10 attr = underline required=yes
#19 @18 '_____________________________________________________________'
#23 @22 'YES or NO indicate the dichotomous responses for each rater'
#30 @33 'Press ENTER to continue' attr=highlight ;

%macro def ;
      %let a =  ;
      %let b =  ;
      %let c =  ;
      %let d =  ;
      %display kap ;
%mend def ;
%def ;data one;
set two;
if mix(var1, var2) > 0 then do;
```

The DEF macro following the %WINDOW statement macro sets the initial value of cells a, b, c, and d to blank spaces. Running the code above and entering the values from Figure 2 will produce the KAP window shown in Figure 3.



Rater Agreement

2 Raters Only

Enter the counts a, b, c, and d in the table below:
(Use the TAB key to jump to next cell)

			RATER B		
			YES		NO
RATER A	YES	Cell A	95	Cell B	4
	NO	Cell C	1	Cell D	0

YES or NO indicate the dichotomous responses for each rater

Press ENTER to continue

Figure 3. KAP Input Window

DEFINE FORMATS

The macro variables a, b, c, and d have now been assigned the values of 95, 4, 1, and 0, respectively. The next portion of the program defines the formats used in the program output:

```
proc format ;
  value rating
    0 = "poor"
    1 = "slight"
    2 = "fair"
    3 = "moderate"
    4 = "substantial"
    5 = "almost perfect"
    6 = "cannot calculate kappa"
    ;
  value rs
    1 = "yes"
    2 = "no"
    ;
run ;
```

The RATING formats are defined to include the Landis and Koch (1977) interpretation of the value of kappa in the output. If there are too many zero cells in the 2x2 table, the kappa statistic cannot be calculated, and the output will indicate this (RATING value = 6).
The RS formats for 1 = “yes” and 2 = “no” are used in the “2x2 Table” section discussed later.

CALCULATE KAPPA AND OTHER MEASURES OF AGREEMENT

The relatively lengthy CALCS DATA set is used for all of the calculations. The values from the 2x2 table are now used in a DATA step to calculate kappa:

```
data calcs ;
 a = &a ;
 b = &b ;
 c = &c ;
 d = &d ;
 N = a+b+c+d ;
```

The values entered in the 2x2 table are defined as variables a, b, c, and d. The total number of ratings, N, is the sum of these four cells. The proportion of observed agreement (p_o), the expected proportion of agreement (p_e), the proportion of positive agreement (p_{pos}) and the proportion of negative agreement (p_{neg}) are then calculated:

```
 po = (a+d)/N ;
 pe = ((a+c)*(a+b)+(b+d)*(c+d))/N**2 ;
 ppos = (2*a)/(N+a-d) ;
 pneg = (2*d)/(N-a+d) ;
```

Calculation of the prevalence index, (p_{index}), the bias index, (b_{index}), and the PABAK follows:

```
 pi = (a-d)/N ;
 bi = (b-c)/N ;
 pabak = 2*po-1 ;
```

The kappa statistic and its asymptotic standard error (Fleiss, 2003) are determined:

```
 kappa = (po-pe)/(1-pe) ;
 q = ((a/N)*(1-(((a+b)/N)+((a+c)/N))*(1-kappa))**2)+((d/N)*
     (1-(((c+d)/N)+((b+d)/N))*(1-kappa))**2);
 r = ((1-kappa)**2)*((b/N)*(((a+c)/N)+((c+d)/N))**2+(c/N)*
     (((b+d)/N)+((a+b)/N))**2) ;
 s = (kappa - pe*(1-kappa))**2 ;
*Asymptotic standard error ;
 se_kappa = sqrt((q+r-s)/(N*(1-pe)**2)) ;
 LL_95_CI = kappa-1.96*se_kappa ;
 if LL_95_CI < -1.00 then LL_95_CI = -1.00 ;
 UL_95_CI = kappa+1.96*se_kappa ;
```

```
  if UL_95_CI > 1 then UL_95_CI = 1.00 ;
```

The 95% lower and upper confidence limits around kappa are calculated and set to -1 or +1 if the confidence limit is less than -1 or greater than +1. Had you specified the KAPPA option in the TABLE statement of a typical PROC FREQ, you would obtain these same values of kappa, the asymptotic standard error, and the 95% confidence limits. This program also provides a test of the null hypothesis kappa = 0 versus kappa > 0. The standard error under the null, the z-statistic, and the p-value are calculated (Rosner, 2000):

```
   se_kappa_null =sqrt((1/(N*(1-pe)**2))*(pe+(pe**2)-
      ((((a+b)/N)*((a+c)/N)*(((a+b)/N)+((a+c)/N))+((c+d)/N)*((b+d)/N)*(((c+d)/
      N)+((b+d)/N)))))) ;
   z = kappa/se_kappa_null ;
   p = 1 - cdf('Normal',z,0,1) ;
```

The final two sections of the CALCS DATA step define the labels for the variables that were created:

```
   label se_kappa = "Kappa Std. Error" ;
   label LL_95_CI = "95% CI Lower Limit" ;
   label UL_95_CI = "95% CI Upper Limit" ;
   label se_kappa_null = "Kappa Std. Error (Under Ho)" ;
   label z = "Z (Under Ho:Kap=0)" ;
   label p = "One sided p-value (Under Ho:Kap=0)" ;
   label po = "Observed Agreement (Po)" ;
   label pe = "Expected Agreement (Pe)" ;
   label ppos = "Positive Agreement (Ppos)" ;
   label pneg = "Negative Agreement (Pneg)" ;
   label pi = "Prevalence Index" ;
   label bi = "Bias Index" ;
   label kappa = "Kappa" ;
   label pabak = "PABAK" ;
```

Next, the STRENGTH variable that is used to assign the RATINGS format to the specific value of kappa is defined. The ranges of kappa that define the strength of association were discussed previously.

```
   strength = 0 ;
   if kappa gt 0.00 and kappa le 0.20 then strength = 1 ;
   if kappa gt 0.20 and kappa le 0.40 then strength = 2 ;
   if kappa gt 0.40 and kappa le 0.60 then strength = 3 ;
   if kappa gt 0.60 and kappa le 0.80 then strength = 4 ;
   if kappa gt 0.80 and kappa le 1.00 then strength = 5 ;
   if kappa = . then strength = 6 ;
   format strength rating. ;
   label strength = "Strength of Agreement" ;
  run ;
```

To print the results with the appropriate formats, the following statement can be used:

```
  proc print data = calcs label noobs ;
      var kappa strength se_kappa LL_95_CI UL_95_CI se_kappa_null z p po pe ppos pneg
             pi bi pabak ;
      format po pe ppos pneg pi bi kappa se_kappa LL_95_CI UL_95_CI se_kappa_null p
             pabak 6.4 z 4.2 ;
  run ;
```

For the Figure 2 data, the following output is obtained from the PROC PRINT:

```
                                   The SAS System

                                                                              One sided
                   Strength      Kappa    95% CI    95% CI    Kappa Std.    Z (Under    p-value
                      of         Std.     Lower     Upper       Error       Ho:Kap=      (Under
         Kappa     Agreement     Error    Limit     Limit     (Under Ho)       0)       Ho:Kap=0)
         -.0163       poor      0.0132    -.0422    0.0097      0.0793        -.21        0.5813


        Observed     Expected     Positive     Negative
        Agreement    Agreement    Agreement    Agreement    Prevalence     Bias
          (Po)         (Pe)        (Ppos)       (Pneg)         Index       Index     PABAK
          0.9500      0.9508      0.9744      0.0000       0.9500      0.0300    0.9000
```

As shown earlier, kappa = -0.0163, indicating poor reliability. But the prevalence index is 0.95, indicating that the raters A and B were in agreement for 95 of the 100 observations. In this situation, the kappa statistic alone is not sufficient in characterizing the reliability of these ratings. The PABAK, which is calculated from the observed proportion of agreement, is 0.90.

2X2 TABLE

When the KAPPA option is used in a PROC FREQ to calculate a kappa statistic, the output 2x2 table will not print if the kappa coefficient cannot be calculated (when there are too many zero cells). This drawback can be overcome by using the following two DATA steps before running the PROC FREQ.
The first DATA step assigns the possible outcomes of ratings from the 2 raters:

```
data rater_outcomes ;
    input rater_a rater_b ;
    datalines ;
1 1
1 2
2 1
2 2
;
run ;
```

The next DATA step creates the WT variable and assigns it the values based on the a, b, c, and d macro variable values:

```
data rater_response ;
    set rater_outcomes ;
    wt = &a ;
    if rater_a = 1 & rater_b = 2 then wt = &b ;
    if rater_a = 2 & rater_b = 1 then wt = &c ;
    if rater_a = 2 & rater_b = 2 then wt = &d ;
run ;
```

The PROC FREQ below will generate a 2x2 table that matches the data entered in the KAP %WINDOW. The options in the TABLE and WEIGHT statements are used to only show the counts in each cell a, b, c, and d, and to allow zero cells to be included in the table. The FORMAT statement will label the cells “yes” and “no” to match the %WINDOW.

```
proc freq data = rater_response  ;
    table rater_a*rater_b / norow nocol nopercent ;
    weight wt/ zeros  ;
    format rater_a rater_b rs. ;
run ;
```

Running this procedure with the example data produces the following table:

```
     Table of rater_a by rater_b

rater_a     rater_b

Frequency|yes     |no      |  Total
---------+--------+--------+
yes      |     95 |      4 |     99
---------+--------+--------+
no       |      1 |      0 |      1
---------+--------+--------+
Total          96        4      100
```

CONCLUSION

The kappa statistic is widely used as a measure of agreement between raters. In some situations when the prevalence of a given response is very high or very low, interpretation of the kappa statistic alone may not be meaningful. In the situation of two raters, the SAS code in this paper has been used to calculate additional measures, such as the prevalence and bias indices and the PABAK, that can be used to characterize inter-rater reliability when this paradox of the kappa statistic is observed.

REFERENCES

Byrt T, Bishop J, Carlin JB. (1993): Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46:423-429.

Cibere J, Thorne A, Bellamy N, et al. (2008): Reliability of the hip examination in osteoarthritis: effect of standardization. *Arthritis & Rheumatism* 59:373-381.

Cicchetti DV and Feinstein AR. (1990): High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 6:551-558.

Feinstein AR and Cicchetti DV. (1990): High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 6:543-549.

Fleiss JL, Levin B, Paik MC. (2003): "Statistical Methods for Rates and Proportions, 3rd Edition". John Wiley & Sons, Inc., Hoboken, New Jersey.

Girianelli VR and Santos Thuler LC. (2007): Evaluation of agreement between conventional and liquid-based cytology in cervical cancer early detection based on analysis of 2,091 smears: experiences at the Brazilian National Cancer Institute*.* *Diagnostic Cytopathology* 35:545-549.

Hoehler FK. (2000): Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology* 53:499-503.

Landis JR, Koch GG. (1977): The measurement of observer agreement for categorical data*.* *Biometrics* 33:159-174.

Lantz CA and Nebenzahl E. (1996): Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology* 49:431-434.

Rosner BA (2000): "Fundamentals of Biostatistics, 5th Edition" Pacific Grove, CA: Duxbury.

Sim J, Wright CC. (2005): The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy* 85:257-268.

Viera AJ, Garrett JM. (2005): Understanding interobserver agreement: the kappa statistic. *Family Medicine* 37:360-363.

ACKNOWLEDGMENTS

Development of this program was supported through an agreement between Highmark, Inc. and the University of Pittsburgh, Department of Biostatistics. I would like to thank Linda Ruhl of Highmark, Inc. and Dr. Gary Marsh of the University of Pittsburgh for guidance and assistance with this project.

CONTACT INFORMATION

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

Michael Cunningham
University of Pittsburgh, Department of Biostatistics, Graduate School of Public Health
A440 Crabtree Hall, 130 DeSoto Street
Pittsburgh, PA 15261
Phone: (412) 624-3039
Fax: (412) 624-9969
E-mail: mac20@pitt.edu
Web: http://www.biostat.pitt.edu/